




# Human-Like Modulation Sensitivity Emerging through Optimization to Natural Sound Recognition

 Takuya Koumura,  Hiroki Terashima, and  Shigeto Furukawa

NTT Communication Science Laboratories, Atsugi, Kanagawa 243-0198, Japan

Natural sounds contain rich patterns of amplitude modulation (AM), which is one of the essential sound dimensions for auditory perception. The sensitivity of human hearing to AM measured by psychophysics takes diverse forms depending on the experimental conditions. Here, we address with a single framework the questions of why such patterns of AM sensitivity have emerged in the human auditory system and how they are realized by our neural mechanisms. Assuming that optimization for natural sound recognition has taken place during human evolution and development, we examined its effect on the formation of AM sensitivity by optimizing a computational model, specifically, a multilayer neural network, for natural sound (namely, everyday sounds and speech sounds) recognition and simulating psychophysical experiments in which the AM sensitivity of the model was assessed. Relatively higher layers in the model optimized to sounds with natural AM statistics exhibited AM sensitivity similar to that of humans, although the model was not designed to reproduce human-like AM sensitivity. Moreover, simulated neurophysiological experiments on the model revealed a correspondence between the model layers and the auditory brain regions. The layers in which human-like psychophysical AM sensitivity emerged exhibited substantial neurophysiological similarity with the auditory midbrain and higher regions. These results suggest that human behavioral AM sensitivity has emerged as a result of optimization for natural sound recognition in the course of our evolution and/or development and that it is based on a stimulus representation encoded in the neural firing rates in the auditory midbrain and higher regions.

**Key words:** auditory; modulation; neural network; neurophysiology; psychophysics; sound recognition

## Significance Statement

This study provides a computational paradigm to bridge the gap between the behavioral properties of human sensory systems as measured in psychophysics and neural representations as measured in nonhuman neurophysiology. This was accomplished by combining the knowledge and techniques in psychophysics, neurophysiology, and machine learning. As a specific target modality, we focused on the auditory sensitivity to sound AM. We built an artificial neural network model that performs natural sound recognition and simulated psychophysical and neurophysiological experiments in the model. Quantitative comparison of a machine learning model with human and nonhuman data made it possible to integrate the knowledge of behavioral AM sensitivity and neural AM tunings from the perspective of optimization to natural sound recognition.

## Introduction

Amplitude modulation (AM) is a critical sound feature for hearing (Fig. 1). Not only is AM associated with basic hearing sensations such as loudness fluctuation, pitch, and roughness (Joris et al., 2004) but it is also an essential clue for recognizing natural

sounds, including everyday sounds and speech (Dudley, 1939; Shannon et al., 1995; Gygi et al., 2004). The significance of AM sensitivity to our hearing functions is supported by its correlations with speech recognition performance, as revealed by experiments mostly conducted on hearing-aid and cochlear-implant users (Cazals et al., 1994; Fu, 2002; Luo et al., 2008; Won et al., 2011; De Ruiter et al., 2015; Bernstein et al., 2016).

The properties of AM sensitivity have been investigated mainly through two separate approaches, psychophysics and neurophysiology. On the one hand, psychophysical studies have identified a wide variety of sensitivity curves in the form of the temporal modulation transfer function (TMTF; Viemeister, 1979; Dau et al., 1997a; Lorenzi et al., 2001a, b). The TMTF is defined as the AM-detection threshold (i.e., the minimum AM depth required for detection) as a function of the AM rate (Fig. 2). Typically, it is measured with a sinusoidal AM (Fig. 1c). It

Received Oct. 25, 2022; revised Mar. 20, 2023; accepted Mar. 28, 2023.

Author contributions: T.K., H.T., and S.F. designed research; T.K. performed research; T.K. contributed unpublished reagents/analytic tools; T.K. analyzed data; and T.K., H.T., and S.F. wrote the paper.

This work was supported by Japan Society for the Promotion of Science Grant JP20H05957.

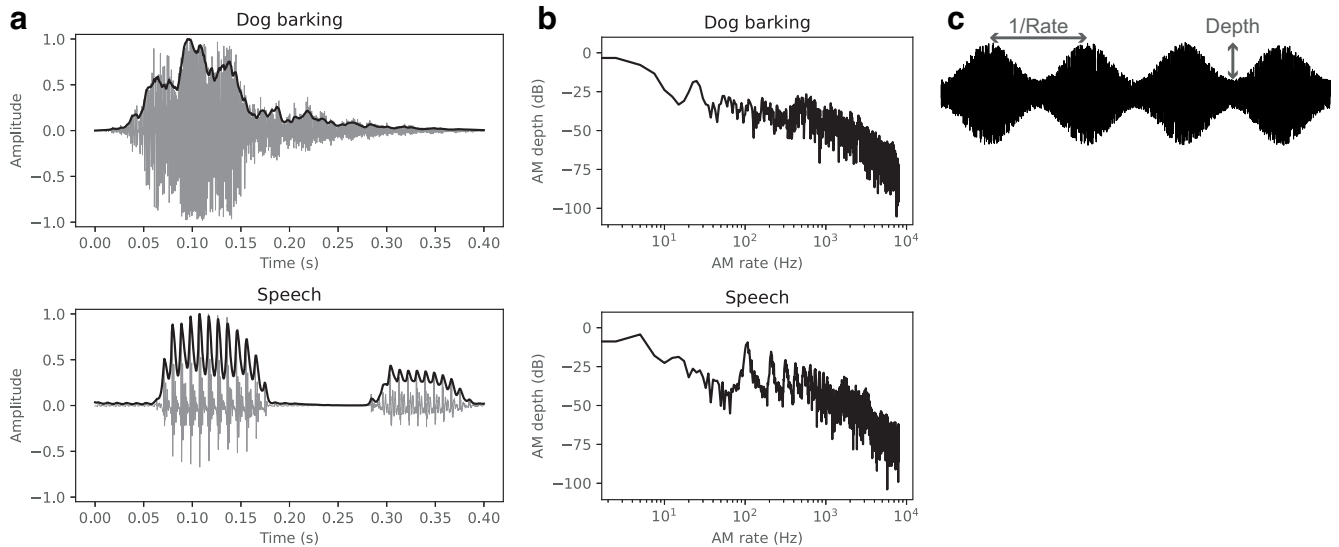
The authors declare no competing financial interests.

Correspondence should be addressed to Takuya Koumura at koumura@cycentum.com.

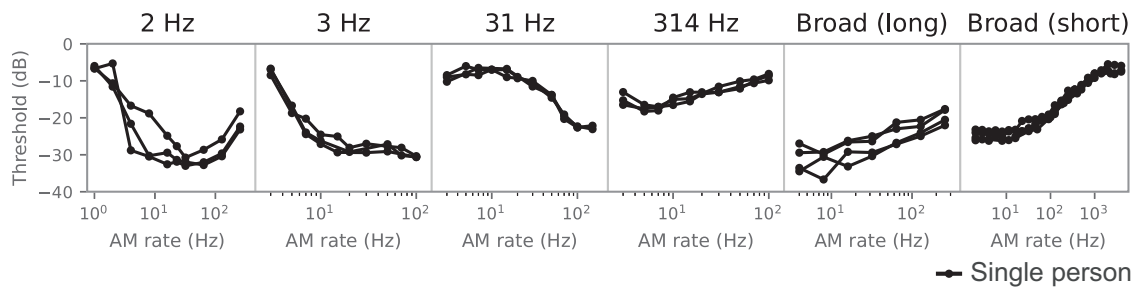
<https://doi.org/10.1523/JNEUROSCI.2002-22.2023>

Copyright © 2023 Koumura et al.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.



**Figure 1.** *a*, Examples of AM in natural sounds. Excerpts of a dog barking (top) and speech (bottom) are shown. Sound waveforms and their amplitude envelopes are shown by gray and black lines, respectively. *b*, Modulation spectra of the sounds in *a*. Each sound has a distinct modulation pattern. *c*, Illustration of the AM depth and rate (actually, the inverse of the rate) of sinusoidally amplitude-modulated white noise. Generally, the shallower the AM depth is the more difficult AM becomes to detect.



**Figure 2.** TMTFs of humans, sorted by the carrier bandwidth of the stimulus. The TMTF is defined as the AM detection threshold as a function of the AM rate. Amplitude modulation of broadband carriers yields low-pass-shaped TMTFs with lower thresholds at low AM rates and higher thresholds at high AM rates, whereas it yields high-pass-shaped TMTFs for narrowband carriers. Other stimulus parameters also appear to affect TMTFs. The depicted TMTFs were taken from psychophysics papers (Viemeister, 1979; Dau et al., 1997a; Lorenzi et al., 2001a,b). Each line shows a TMTF in a single person.

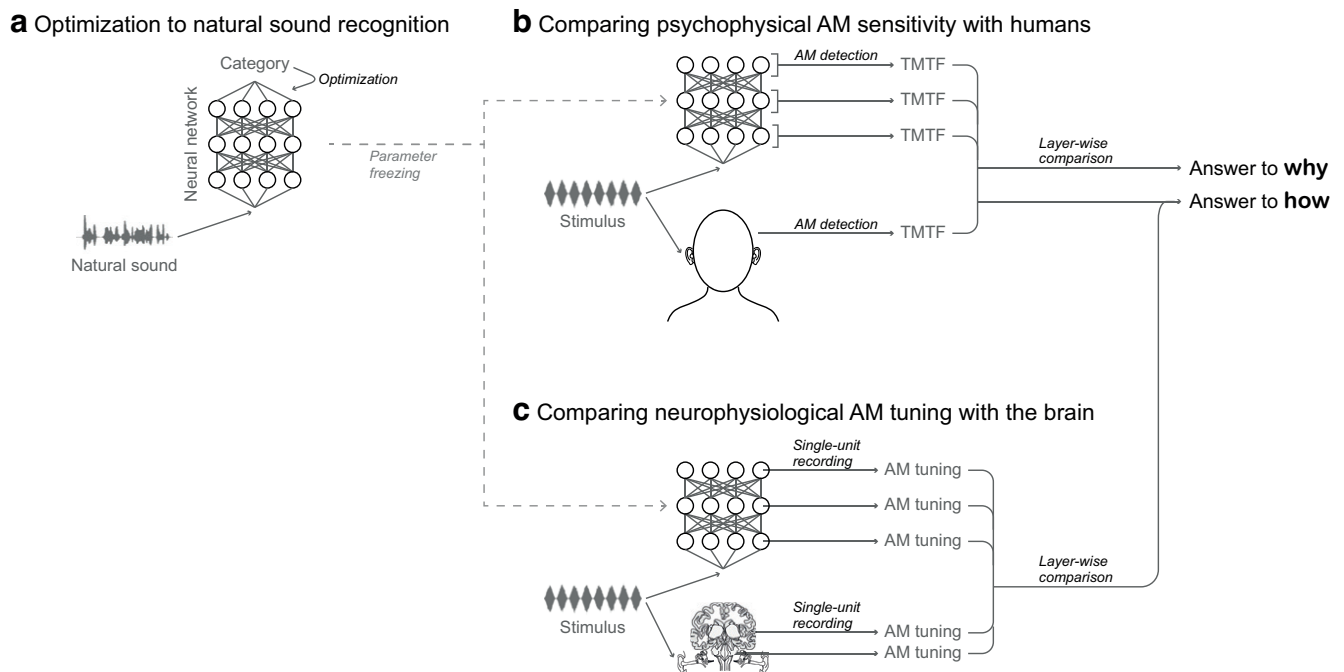
shows apparent interactions with the carrier bandwidth. The detection thresholds are higher (less sensitive) at an AM rate equal to the carrier bandwidth. These patterns have been interpreted in terms of frequency masking in the modulation domain. Stimulus parameters other than the carrier bandwidth (e.g., stimulus duration) may be additional factors determining the TMTF form.

On the other hand, neurophysiological studies have found that many neurons throughout the mammalian auditory nervous system (ANS) show tuning to AM (Joris et al., 2004). Their spike rate and/or spike timing depends on the stimulus AM rate. Their preferred AM rate varies widely over the range of behaviorally detectable values. Although these findings suggest that AM-tuned neurons are somehow involved in behavioral AM sensitivity, the lack of single-unit neural data in humans has made it difficult to establish a direct link with human behavior.

Inspired by the psychophysical and neurophysiological findings, Dau et al. (1997a,b) have proposed that a bandpass filter bank in the modulation domain, called a modulation filter bank (MFB), is involved in auditory signal processing. They built a computational model that includes an MFB with which they reproduced a variety of psychoacoustic properties including stimulus-parameter-dependent TMTFs (Dau et al., 1997a,b). To reproduce a wider range of psychoacoustic phenomena, they

have gradually incremented and refined the model components that each performs a specific signal processing computation (Derleth et al., 2001; Jepsen et al., 2008). Building a model in such a bottom-up fashion is advantageous for theorizing on what kinds of signal processing are implemented in the human auditory system. However, to fully understand the properties of AM sensitivity, we should also answer two critical questions, Why has it emerged during our evolution and development? and How is it realized by our neural mechanisms? The neural mechanisms of AM sensitivity have been studied mostly by animal neurophysiology, but explaining why and how in a single computational framework would help us understand those neurophysiologically elucidated mechanisms from the perspectives of human behavior and the process of their emergence.

To provide answers to these questions, we built a computational model that performs natural sound recognition and compared its psychophysical and neurophysiological properties with those of the auditory system (Fig. 3). First, to investigate why AM sensitivity has emerged, we optimized an artificial neural network (NN) for natural sound recognition (Fig. 3a) as a way of simulating the optimization that is presumably happening in the auditory system during its evolution and development. We assumed that better recognition of natural sounds yields better evolutionary fitness and hypothesized that natural sound recognition



**Figure 3.** *a–c*, Schematic illustration of the framework of the present study, consisting of three stages. Humans have evolved and developed the ability to precisely recognize natural sounds (*a*). We realized a computational simulation of this process by optimizing a model for natural sound recognition. Specifically, we used a deep NN that takes a sound waveform as input and estimates its category. We froze the learned parameters and measured the AM sensitivity in the NN by using the same procedure as in human psychophysical experiments (*b*). A TMTF was computed for each layer. It was compared with previously reported human AM-sensitivity data in an attempt to answer why AM sensitivity has emerged in humans in its current form. We measured neurophysiological AM tuning in the units in the NN by using the same procedure as in animal neurophysiological experiments (*c*). On the basis of the similarity of the AM tuning with the auditory brain regions and the results of the psychophysical experiments, we could infer possible neural mechanisms underlying behavioral AM sensitivity.

plays a major role in shaping human AM sensitivity. We used everyday sounds (Piczak, 2015) and speech sounds (<https://doi.org/10.35111/17gk-bn40>) as examples of natural sounds. Then, we simulated psychophysical experiments on the NN to see whether human-like AM sensitivity emerges in some of its layers (Fig. 3*b*). This kind of two-step optimization and analysis procedure has explained a number of auditory properties (Lewicki, 2002; Terashima and Okada, 2012; Khatami and Escabi, 2020; Ashihara et al., 2021; Saddler et al., 2021; Francl and McDermott, 2022), as well as properties in other sensory modalities (Kriegeskorte and Douglas, 2018; Kanwisher et al., 2023).

Finally, to investigate how AM sensitivity is realized, we performed neurophysiological experiments on the same model and made a hierarchical correspondence with the ANS (Fig. 3*c*). By taking advantage of a method established in our previous study that maps the AM representation between an NN and the ANS based on single-unit activity (Koumura et al., 2019), we roughly mapped the layerwise AM sensitivity measured in the psychophysical simulation onto the hierarchical processing stages in the ANS. In this way, we could infer which brain regions are most likely to be responsible for human AM sensitivity.

Parts of this article have been previously presented in Koumura et al. (2020).

## Materials and Methods

**Model construction and evaluation.** We used a multilayer feedforward NN as a model of the auditory system. Each layer consisted of a dilated convolution (van den Oord et al., 2016) followed by an exponential linear unit (ELU; Clevert et al., 2016). Convolution was along the time axis. Above the topmost layer was a classification layer consisting of a convolution with a filter size of one. In this way, the model worked as a fully convolutional NN. The input time window was 0.2 s. In other words, the model estimated the sound category of every 0.2 s of the input

sound. During optimization, softmax cross entropy for sound categories was computed at a single time step of the model output (corresponds to the input sampling rate), and the parameters (namely, convolutional weights and biases) were updated to minimize the error. During the evaluation, the output of the classification layer was averaged over time to estimate a single category per sound clip for the everyday sounds or per phoneme interval for the speech sounds.

The trainable parameters of the model were the connection weights and biases in the convolutional layers. Initially, the connection weights were random, and the biases were zero. These parameters were optimized for sound recognition with a standard backpropagation method using the Adam optimizer with a learning rate of  $10^{-4}$ . We refer to a model with initial parameters (random weights and zero biases) as a “nonoptimized model” and a model after optimization as an “optimized model.” The sound data were divided into training and validation sets. We used the early stopping strategy. This means that the parameter update was conducted with part of the training set until recognition accuracy stopped improving for the other part of the training set.

The model was very similar to the one in our previous study in that it consisted of a stack of a dilated temporal convolution followed by an ELU activation function and that it was optimized to categorize everyday sounds or speech sounds based on the softmax cross entropy (Koumura et al., 2019). On the other hand, there are some nonessential differences. In the present study we newly sampled the architectural parameters (namely, the number of layers, number of units per layer, convolutional filter width, and convolutional dilation width), the connection weights and biases are newly optimized, the input duration of the previous model was 0.19 or 0.26 s, the previous study used the Eve optimizer for optimization, and the previous study used a subset of the environmental sound classification (ESC)-50 dataset without human-originated sounds because it focused on comparison with the nonhuman ANS, whereas the present study used the entire dataset as described below.

**Sound data for optimization.** We used two datasets for optimization, ESC-50 (Piczak, 2015) and TIMIT (Linguistic Data Consortium; <https://doi.org/10.35111/17gk-bn40>). Both datasets are commonly used for sound recognition and are relatively small (Fonseca et al., 2022). We did

not use larger datasets because our purpose was not to achieve state-of-the-art sound-recognition performance. The optimization to the two datasets was conducted independently using different NNs.

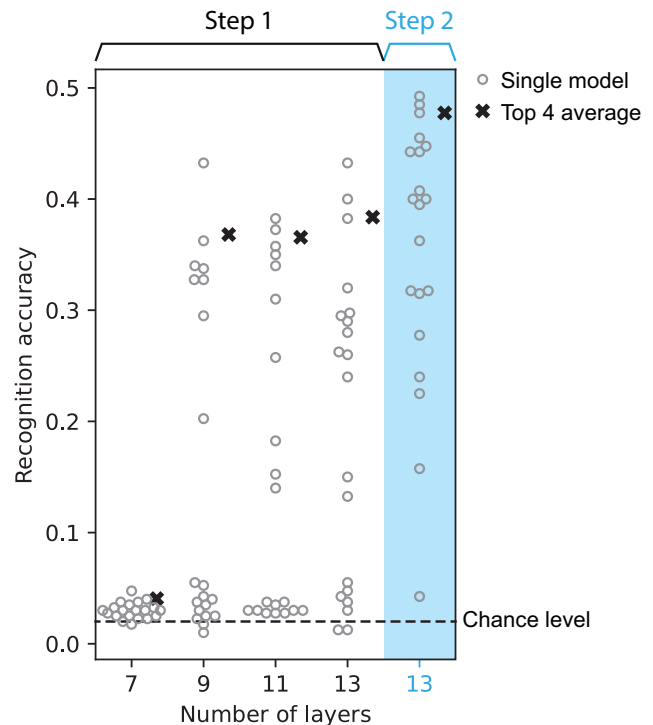
ESC-50 defines five folds. We used folds 1–4 for training and fold 5 for validation. In the training, folds 1–3 were used for the parameter update and fold 4 was used for early stopping. Some sound clips end with absolute zero amplitude values, probably to make the clip duration 5 s in otherwise shorter sounds. We excluded such zero tailings. The dataset contains 50 categories of everyday sounds, which are roughly grouped into the following five category groups: animals, natural soundscapes and water sounds, human nonspeech sounds, interior/domestic sounds, and exterior/urban noises. The optimization objective is a 50-way classification of a sound input. Although the name of the dataset ESC stands for environmental sound classification, in this study we call it “everyday sound” because the dataset contains not only environmental sounds (e.g., rain, sea waves, crackling fire) but also sounds from a single event (e.g., sneezing, door knock, mouse click). Such sounds are often called everyday sounds (Van Grootel et al., 2009; Norman-Haignere et al., 2015).

TIMIT defines training and test sets. The test set includes sentences spoken by the core-test speakers and non-core-test speakers. For validation, we used sentences spoken by the core-test speakers. For training, we used the training set for the parameter update and the sentences spoken by the non-core-test speakers for early stopping. We excluded sentences included in both the training and test sets. This process ensured that there was no duplication of sentences or speakers in the training and validation sets. We merged 61 categories contained in the dataset into 39 categories as proposed by the previous study (Lee and Hon, 1989). Because a single sound clip consists of a sequence of phonemes, a 0.2 s input can contain multiple phonemes. During training, the optimization objective was to estimate the phoneme category at the center of the 0.2 s input. During evaluation, the output of the classification layer was averaged over the interval of a single phoneme to produce a single output for each phoneme interval.

Before being fed to the model, the sound signals were high-pass filtered at 20 Hz, and the 10 ms raised-cosine ramps were applied to the onset and the offset. During training, the sound amplitude was slightly varied clip by clip. During the evaluation, it was fixed to the mean value of that for training.

**Architecture search.** For the architecture search, we tested architectures that varied in the number of layers, number of units per layer, and convolutional filter size and dilation width. The number of layers was 7, 9, 11, or 13. For each number of layers, we sampled 20 models by varying the number of units per layer, convolutional filter size, and convolutional dilation width. The convolutional filter width and the dilation width were randomly sampled for each layer with the constraint on the input time window being 0.2 s. The number of units per layer was either 32, 64, 128, 256, or 512. To avoid an expensive computation of training all models over numerous iterations (Zhou et al., 2020), we conducted a two-step architecture search as follows. In the first step, we sought the number of layers that would potentially achieve the highest recognition accuracy. All models were trained until the recognition accuracy for a subset of the training set stopped improving for 32 epochs. Average recognition accuracy at this point of the four best models among those with the same number of layers was the highest for the 13-layer models (Fig. 4). Thus, we selected the 13-layer architectures and discarded the others. In the second step, we further trained those 20 models until the recognition accuracy on a subset of the training set stopped improving for 96 epochs. We selected the four models with the highest recognition accuracy for the subsequent psychophysical and neurophysiological analyses.

**Experimental design and statistical analyses.** To measure the AM detection threshold in the model, we simulated AM detection experiments in human psychophysics. To compare our results fairly with those produced by humans, the simulations duplicated the procedure of the human experiments as precisely as possible. One exception was that in human studies a detection threshold is estimated with a staircase method, whereas we computed the AM detection accuracy for each modulation depth independently.



**Figure 4.** Sound recognition accuracy of the models with different architectures. Left, In the first step of the search process, four models with 13 layers had the highest average accuracy (area with the white background). Right, In the second step, the accuracy of the models with 13 layers improved after further optimization (area with the blue background).

We simulated a two- or three-interval forced-choice (IFC; 2IFC or 3IFC) task. In each trial, two or three stimuli were presented to the model, one of them being modulated, the others not. The task was to correctly identify the modulated interval. We conducted this task by assuming an AM detection process based on the model activities. We conducted 128 trials for each AM depth, from which the proportion of correct trials was calculated.

The proportion of correct trials plotted against the AM depth yields a psychometric curve. It was fitted with an asymmetric sigmoid function (Richards, 1959; Fekedulegn et al., 1999). The detection threshold was defined as the AM depth at which detection accuracy was 70.7% on the fitted curve. In some conditions, the threshold could not be estimated because the proportion of correct responses was either too high or too low at all tested AM depths. Excluding such a condition would result in an overestimation of the similarity to human TMTFs. To avoid the overestimation, instead of excluding such a condition, the threshold was clipped to the maximum or minimum values of the tested range of the AM depth. The range is described below.

**AM detection based on time-averaged unit activities.** An  $x$ IFC ( $x = 2$  or 3) task was conducted by estimating the modulated interval from model activities. Specifically, we assumed AM detection based on time-averaged unit activities. For each stimulus interval, unit activities in the model were averaged over time. From the time-averaged activities in a single layer, a logistic regression was trained to estimate whether the stimulus was modulated. The proportion of correct trials was computed in a 4-fold cross validation of a total of 128 trials. In each of the 32 held-out trials, the probability of the stimulus being modulated was calculated for each stimulus interval, and the interval with the maximum probability was considered as the response in that trial. If that interval was actually the modulated interval, the trial was considered correct. L2 regularization was applied to logistic regression. The regularization coefficient was optimized in another 4-fold cross validation within the training set.

**Stimulus.** We tested six stimulus parameters from three independent human studies. All of them were sinusoidally amplitude modulated narrowband or broadband white noise. The stimulus parameters and



generation procedure were as close as possible to those in the human studies, except for amplitude scaling; in the human studies, it was based on sound pressure levels, whereas in this study, it was based on the root mean square (RMS). The stimulus RMS was adjusted to the average RMS of the training set. In the short broadband condition, the stimulus amplitude was scaled before applying modulation (Viemeister, 1979). In the other conditions, it was scaled after modulation (Dau et al., 1997a; Lorenzi et al., 2001a, b).

In the narrowband carrier conditions, Gaussian noise was bandpass filtered with a digital Fourier transform. In the 314 Hz carrier bandwidth condition, bandpass filtering was applied after modulation (Dau et al., 1997a). In the other conditions, bandpass filtering was applied before modulation (Dau et al., 1997a; Lorenzi et al., 2001b). The Gaussian noise carrier was sampled independently in each stimulus.

AM depth is expressed in dB relative to the sound amplitude. In the case of sinusoidal AM, a sound with an AM depth  $m$  in dB and rate  $f$  is defined as follows:

$$(1 + 10^{\frac{m}{20}})\sin(2\pi ft + \varphi)C(t),$$

where  $\varphi$  is the AM starting phase,  $t$  is time, and  $C(t)$  is a carrier signal. The AM starting phase was fixed to zero in the 3 Hz, 31 Hz, and 314 Hz bandwidth conditions and in the long broadband condition (Viemeister, 1979; Dau et al., 1997a). In the other conditions, it was randomly sampled independently in each stimulus (Lorenzi et al., 2001a, b).

The range and steps of the AM rate differ among the human experiments. For each condition, we chose eight AM rates evenly spaced on a log scale within the range in the particular human experiment. The AM depths ranged from  $-60$  to  $0$  dB in the 2 Hz carrier bandwidth condition and from  $-40$  to  $0$  dB in the other conditions. They were spaced every 4 dB.

*Quantitative comparison of model and human TMTFs.* Previous human studies reported TMTFs in multiple human subjects. We took TMTF values from those studies and compared them with those in our model. Before calculating the quantitative similarities, we averaged the TMTFs across subjects. When the AM rates did not match among subjects, linear interpolation along the log-scaled AM rate was conducted.

Likewise, we averaged TMTFs of the four selected models. Then, we compared the averaged human TMTFs and averaged model TMTFs in terms of their relative patterns and absolute values. Similarity of their relative patterns was quantified by the pattern similarity index, that is, the correlation coefficient of the human and model TMTFs. Similarity of the absolute values was quantified by the discrepancy index, that is, the RMS deviation as follows:

*Discrepancy index*

$$= \sqrt{\text{mean}_{\text{condition}, f_m} ((y_{\text{model}}(\text{condition}, f_m) - y_{\text{human}}(\text{condition}, f_m))^2),$$

where *condition* and  $f_m$  are the experimental conditions (size = 6) and the AM rates in each condition (size = 8), and  $y$  is a detection threshold.

The net difference between the human and model TMTFs was defined as the average signed difference between them as follows:

$$\text{Net difference} = \text{mean}_{\text{condition}, f_m} (y_{\text{model}}(\text{condition}, f_m) - y_{\text{human}}(\text{condition}, f_m)).$$

Positive/negative values of net difference mean larger/smaller thresholds in the model than in humans on average. To take all stimulus conditions into account, TMTFs in all conditions were pooled when calculating those indices.

*Statistical analysis of correlations between the recognition accuracy and the (dis)similarity.* Correlations between the recognition accuracy and the pattern similarity and the discrepancy were assessed with Pearson correlation coefficients. The  $p$  values were Bonferroni corrected for the number of layers.

*Manipulation of the training data for exploring critical features.* To evaluate the importance of amplitude envelope (Env) and temporal fine structure (TFS), we made degraded versions of the training data by disrupting either the Env or TFS components of the sound.

Single-band Env signals were made by combining the Env component of a sound and a TFS component of white noise as in the following:

$$\text{Single band Env signal} = \text{real}(\text{Env}_x(t)\exp(i\text{TFS}_{wn}(t))),$$

where  $\text{Env}_s$  and  $\text{TFS}_s$  are the Env and TFS components of a signal  $s$ ,  $x$  and  $wn$  are the original sound and a white noise with the same RMS as  $x$ ,  $t$  is time,  $i$  is the imaginary unit, and  $\text{real}$  converts a complex signal to its real part. The Env and TFS components are defined as the magnitude and phase of the Hilbert-transformed complex analytic signal.

Single-band TFS signals were made by flattening the Env component of a sound as follows:

$$\text{Single band TFS signal} = \text{real}(\text{FlatEnv}_x\exp(i\text{TFS}_x(t)))$$

$$\text{FlatEnv}_x = \text{RMS}(\text{Env}_x(t)),$$

where  $\text{FlatEnv}_s$  is a flattened Env of a signal  $s$ , which takes a constant RMS value of the Env component.

When making multiband Env and TFS signals, we first decomposed the sound into sub-bands with a linear bandpass filter bank. The filter center frequencies ranged from 20 Hz to the Nyquist frequency and were spaced every one equivalent rectangular bandwidth (Moore, 2013). Because the Nyquist frequency of the everyday sound dataset is 22.05 kHz and that of the speech sound dataset is 8 kHz, the number of bands was 42 and 33, respectively. Then we computed the Env and TFS components for each sub-band. The Env or TFS components were disrupted in the same way as in the single-band signals, and the multiband signals were added to form the final output as follows:

$$\text{Multi band Env signal} = \sum_f \text{real}(\text{Env}_{x_f}(t)\exp(i\text{TFS}_{w_{mf}}(t)))$$

$$\text{Multi band TFS signal} = \sum_f \text{real}(\text{FlatEnv}_{x_f}\exp(i\text{TFS}_{x_f}(t))),$$

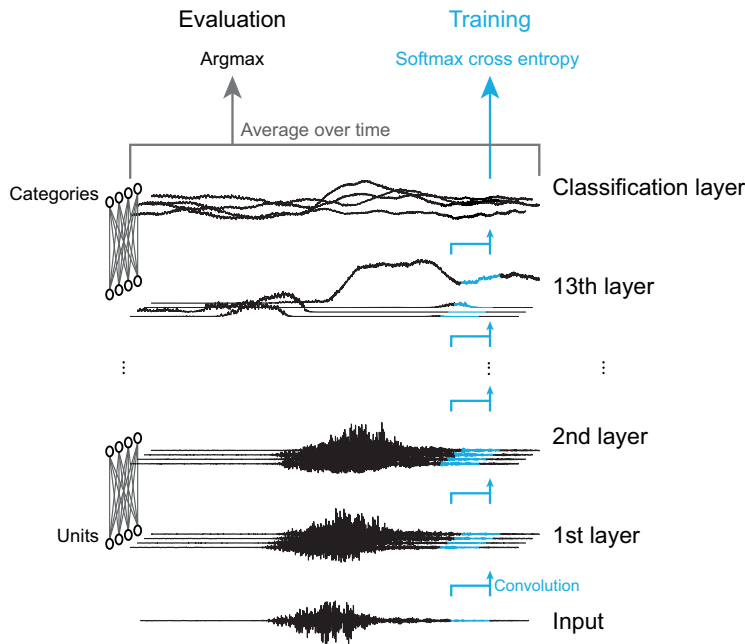
where  $s_f$  is the  $f$ th sub-band of the frequency-decomposed signal  $s$ . Other than the difference in the training data, the procedures of the optimization and analysis were completely the same as the models trained on the original sounds.

*AM detection based on template correlation.* For each sound interval in the xIFC task, the correlation was calculated between the unit activities in a layer in response to the stimulus and the template. The interval with the largest correlation was taken to be the response to the trial. The correlation was defined by the sum of products as in the previous study (Dau et al., 1997a). It was calculated for all units in each layer as follows:

$$\text{Correlation}_{\text{layer}} = \sum_{\text{unit}, t} x_{\text{layer}, \text{unit}}(t)\text{Template}_{\text{layer}, \text{unit}}(t),$$

where  $x_{\text{layer}, \text{unit}}$  is the activity in a specific unit in the target layer, and  $t$  is time.

To make a template, first, we averaged the unit activities across 128 independent fully modulated and nonmodulated stimuli. The template was defined as average unit activities for fully modulated stimuli minus the average unit activities for nonmodulated stimuli (Dau et al., 1997a) as in the following:



**Figure 5.** Schematic illustration of the NN architecture. Units in the first layer took a waveform as input and applied a nonlinear temporal convolution to it. Subsequent layers took the activations in the layer below as input. Above the topmost convolution layer (13th layer in the figure) was a classification layer. The number of units in the classification layer equals the number of sound categories. During training, softmax cross entropy was calculated for a single time frame at a time (corresponding to the input sampling rate). During the evaluation, values in the classification layer were averaged over time, and the category with the maximum average value was chosen as the estimated output category. The classification layer was not included in the psychophysical or neurophysiological analysis. This figure is a simplified illustration. The length of the convolutional filters and the number of units are not the same as those in the actual architectures used in this study.

$$\begin{aligned}
 \text{Template}_{\text{layer,unit}}(t) &= \text{mean}_{i=1 \text{ to } 128} \\
 (x_{\text{layer,unit,modulated}_i}(t)) &- \text{mean}_{i=1 \text{ to } 128} \\
 (x_{\text{layer,unit,nonmodulated}_i}(t)),
 \end{aligned}$$

where  $x_{\text{layer,unit,modulated}_i}$  and  $x_{\text{layer,unit,nonmodulated}_i}$  are the unit activities in response to the  $i$ th modulated and nonmodulated stimulus, respectively.

**Neurophysiological similarity between NN layers and brain regions.** The neurophysiological similarity between NN layers and brain regions was computed in the same way as in our previous study, except that the resolution of the AM rates at which AM tuning was computed was decreased in this study for reducing the computational cost. A detailed description of the method is provided in our previous paper (Koumura et al., 2019).

Unit activities in the model were recorded while presenting it with sinusoidally amplitude-modulated broadband white noise. The AM tuning was defined in terms of the time-averaged unit activities and the synchrony of the activities to the stimulus modulation. It was characterized by the best AM rate and the upper cutoff rate. The best AM rate was defined as the AM rate at which the tuning curve reached a maximum. The upper cutoff rate was defined as the AM rate at which the tuning started to decrease. Distributions of best and upper cutoff rates were compared between NN layers and brain regions. Similarity between an NN layer and a brain region was defined as one minus the Kolmogorov–Smirnov distance of the distributions. The AM tuning in the ANS was taken from previous neurophysiological studies (Müller-Preuss, 1986; Langner and Schreiner, 1988; Schreiner and Urbas, 1988; Batra et al., 1989; Frisina et al., 1990; Preuss and Müller-Preuss, 1990; Joris and Yin, 1992; Rhode and Greenberg, 1994; Zhao and Liang, 1995; Bieser and Müller-Preuss, 1996; Condon et al., 1996; Schulze and Langner, 1997; Eggermont, 1998; Huffman et al., 1998; Joris and Smith, 1998; Joris and Yin, 1998; Kuwada and Batra, 1999; Krishna and Semple, 2000; Lu and Wang, 2000; Lu et al., 2001; Liang et al., 2002; Batra, 2006; Zhang and Kelly, 2006; Bartlett and Wang, 2007; Scott et al., 2011; Yin et al., 2011).

**Data availability.** All data and code are available at <https://github.com/cycentum/Human-like-Modulation-Sensitivity-through-Natural-Sound-Recognition>.

## Results

### Optimizing a neural network for natural sound recognition

Our NN consists of multiple layers, which in turn consist of multiple units (Fig. 5). An input sound waveform was fed to the first layer, which performed temporal convolution and a static nonlinear operation. The outputs of the first layer were fed to the second layer, and this process continued to the topmost layer. There was no feedback or recurrent connections. Above the topmost layer was a classification layer that computed the categories of the input sound. The classification layer was not included in the psychophysical or neurophysiological analysis. To reduce the number of hard-coded assumptions and clarify the relationship between the optimization procedure and the emergent properties, we applied an NN directly to a raw sound waveform without any preprocessing (Hoshen et al., 2015; Tokozume and Harada, 2017). This is in contrast with typical auditory models that attempt to implement a hard-coded frequency-decomposition stage in the cochlea (Bruce et al., 2018; Verhulst et al., 2018).

The model was optimized to correctly classify natural sounds. We used two types of sounds, everyday sounds (Piczak, 2015) and speech sounds (<https://doi.org/10.35111/17gk-bn40>). The optimization objective was to correctly estimate the category of an everyday sound or the phoneme categories in a speech sound. We built and analyzed a model for each sound type. Because the results were generally consistent across different sound types, below we report the results for everyday sounds before those for speech sounds.

The recognition performance of an NN generally depends on its architecture (Bergstra and Bengio, 2012; Bergstra et al., 2013; Klein et al., 2017). In this study, we trained multiple NNs with different architectures and performed psychophysical and neurophysiological analyses on the NNs that achieved the highest recognition accuracy. To reduce possible biases by a specific architecture, unless otherwise stated, the reported recognition accuracy, TMTFs, and neurophysiological similarities are averages of the results of the four models with the highest recognition accuracies. This could be considered as a modeled version of reporting average quantities in multiple participants in human studies (Francl and McDermott, 2022). Four architectures with 13 layers were selected by performing an architecture search (see above, Materials and Methods for the detailed procedure). Their parameters are in Table 1.

After optimization, we evaluated the recognition performance for sounds not used in the model construction. The recognition accuracy was 0.477. This value is well above the chance level (0.02) but lower than that of state-of-the-art machine learning studies (Gong et al., 2021). Although tuning the hyperparameters or increasing the amount of training data may lead to an improvement in accuracy, we used the model as is in the subsequent analysis because our goal was to understand the properties of the human hearing system, not to pursue accuracy improvements.

**Table 1. Architectural parameters of the models with the highest recognition accuracy**

Architecture	Number of units per layer	Convolutional filter size (from lower to higher layers)	Dilation width (from lower to higher layers)
Architecture 1	256	5, 3, 6, 5, 2, 7, 8, 8, 7, 4, 4, 6, 5	231, 603, 18, 138, 14, 97, 105, 7, 137, 381, 193, 208, 266
Architecture 2	512	3, 4, 4, 3, 6, 6, 4, 6, 8, 8, 7, 7, 3	449, 12, 161, 374, 175, 193, 120, 209, 161, 47, 151, 16, 465
Architecture 3	128	7, 4, 4, 2, 7, 3, 7, 5, 2, 4, 3, 7, 2	42, 125, 341, 603, 96, 410, 44, 269, 747, 152, 528, 122, 823
Architecture 4	512	3, 7, 2, 7, 7, 4, 5, 5, 3, 8, 8, 8, 2	676, 105, 581, 54, 16, 192, 214, 2, 173, 173, 184, 92, 887

The layers all had the same number of units for simplicity. The size and dilation width of the convolutional filter were randomly sampled for each layer. The input time window of the filter was calculated as (dilation width)  $\times$  (filter size  $- 1$ )  $+ 1$ .

### Simulating psychophysical experiments as a way of measuring the AM sensitivity of the models

To investigate the relationship between sound recognition and AM sensitivity, we measured the TMTFs in each of the four best models by simulating psychophysical AM detection experiments. To fairly compare the TMTF of the model with those of humans, we replicated the procedures of human psychophysical experiments as precisely as possible. We simulated six psychophysical experiments from three independent human studies (Viemeister, 1979; Dau et al., 1997a; Lorenzi et al., 2001a, b). In all of them, human subjects conducted a 2IFC or 3IFC task. In each trial of the task, two or three stimuli were sequentially presented, and only one among them was modulated. The task of a subject was to identify the modulated stimulus. At a given AM rate, an AM detection threshold was estimated with an adaptive method to find the AM depth that gave a 70.7% correct rate (Levitt, 1971).

The stimuli were sinusoidally amplitude-modulated broadband or narrowband Gaussian noise. They differ in their stimulus parameters (Table 2). Because the most notable difference is in the carrier bandwidth (2 Hz, 3 Hz, 31 Hz, 314 Hz, or broadband), hereafter, we specify the conditions with their carrier bandwidths, except for two conditions with the broadband carrier. We call the broadband condition with the 0.5 s stimulus duration “broadband, short,” and the broadband condition with the 2 s duration “broadband, long.”

In the present study, to conduct an  $x$ IFC task, we presented a stimulus to the model and averaged the activity of each unit over the stimulus duration (Fig. 6). Then, from the vector representing the time-averaged activity of the units in a single layer, we estimated the probability of the stimulus being modulated with logistic regression. The stimulus with the maximum probability was considered to be the response of the model to that  $x$ IFC trial. If the interval actually contained the modulated stimulus, the trial was considered correct. For simplicity, the threshold was estimated with a constant stimulus method. That is, the proportion of correct trials was computed independently for each AM depth. An asymmetric sigmoid function was fitted to the plot of the proportion of correct responses versus AM depth (Fig. 6c). The threshold was defined as the AM depth at which the proportion of correct trials was 70.7% on the fitted curve.

### Emergence of human-like TMTFs in the model

The forms of the TMTFs of the model (detection thresholds as a function of AM rate) varied depending on the stimulus condition and model layer (Fig. 7, orange lines). The forms of the human TMTFs (black dotted lines) also depend largely on the stimulus condition. In all conditions, the model and human TMTFs tended to overlap in the middle to higher layers.

Quantitative analyses supported the above observations. The similarity of the TMTFs of the models to that of humans was evaluated in terms of relative patterns (reflecting mainly similarity in TMTF shape) and absolute values (reflecting similarity in both shape and sensitivity in decibels; Fig. 8). An index of

**Table 2. Stimulus parameters in the AM detection experiments**

Carrier bandwidth	Duration	AM starting		Reference
		phase	Amplitude equalization	
2 Hz	2 s	Random	After applying AM	Lorenzi et al., 2001b
3 Hz	1 s	Constant	After applying AM	Dau et al., 1997a
31 Hz	1 s	Constant	After applying AM	Dau et al., 1997a
314 Hz	1 s	Constant	After applying AM	Dau et al., 1997a
Broadband	0.5 s	Constant	Before applying AM	Viemeister, 1979
Broadband	2 s	Random	After applying AM	Lorenzi et al., 2001a

Other parameters such as the fade duration vary among the studies, but not all of them are shown here.

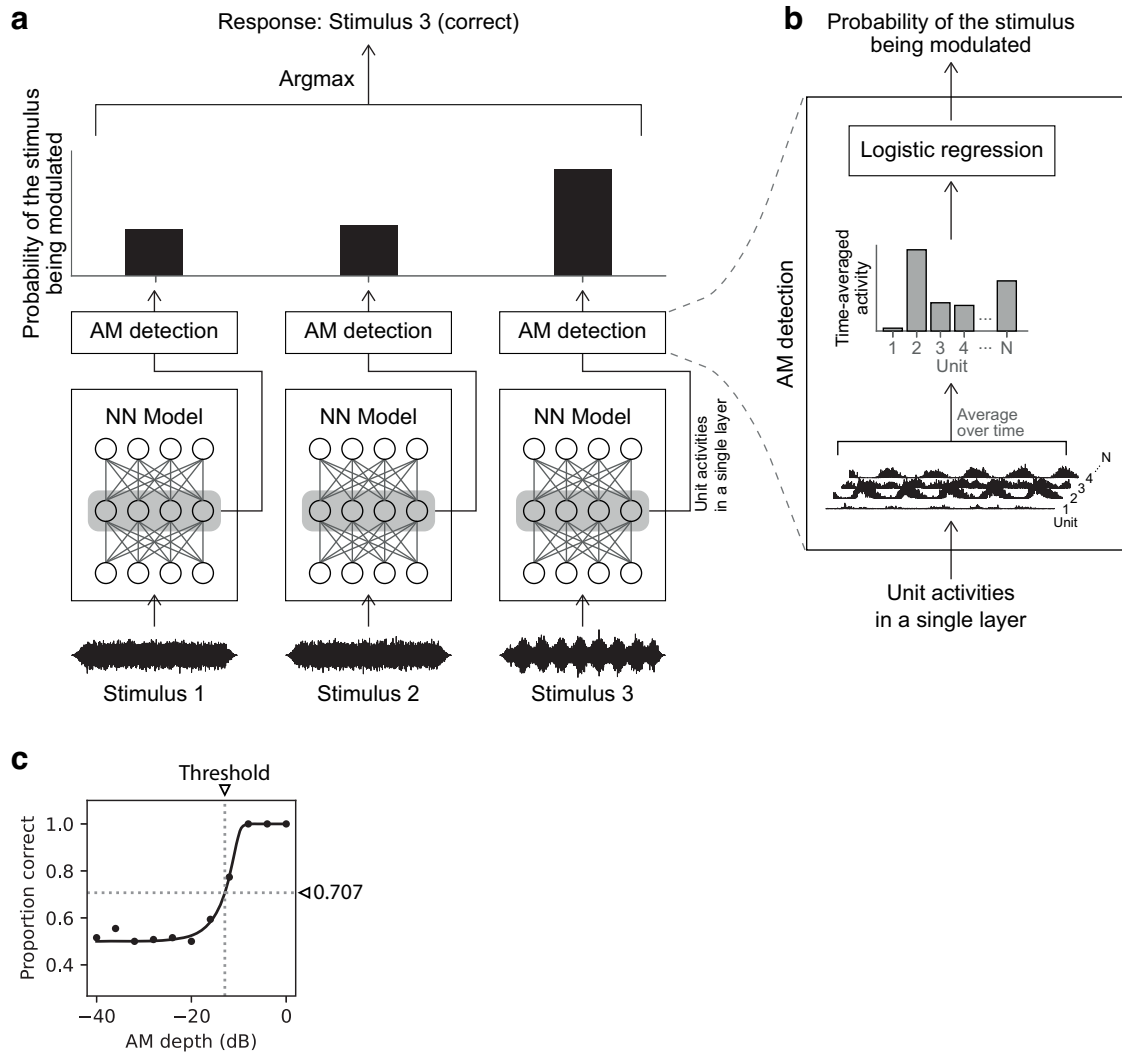
similarity of relative patterns, the correlation coefficient, was calculated from pairs of model and human TMTFs. Hereafter, we call it the pattern similarity index (Fig. 8, top). As an index of absolute measure of dissimilarity, we calculated the root mean square (RMS) deviation and called it the discrepancy index (Fig. 8, bottom). To take all stimulus conditions into account, TMTFs in all stimulus conditions were pooled when calculating the indices. The two measures consistently indicated that layers around the 10th layer exhibited TMTFs most similar to those of humans (highest pattern similarity and lowest discrepancy). This result indicates the emergence of human-like AM sensitivity in the model optimized for natural sound recognition.

Human-like TMTFs did not emerge in the nonoptimized model with random initial parameters. The sound recognition accuracy in the nonoptimized model was 0.013, which was as low as the chance level, 0.02. Generally, the TMTFs in the nonoptimized model were relatively invariant across the layers and showed marked discrepancies from those of humans and the optimized models. These discrepancies were particularly apparent in the higher layers (Fig. 7). These observations are supported by the quantitative analyses, showing a low pattern similarity index and high discrepancy index throughout the layers (Fig. 8). These results suggest that optimization to sound recognition is an essential factor for the emergence of human-like TMTFs and that the NN architecture only could not explain human AM sensitivity.

### Models with better recognition performance were more human-like

An additional analysis revealed a close link between the sound recognition performance of the model and the TMTF similarities. During the NN architecture search, we trained 20 models with 13 layers with different architectures. (Remember that the above analyses targeted the best 4 models of the 20.) The 20 models exhibited recognition accuracies ranging from 0.043 to 0.492 and produced TMTFs with a varying degree of similarity to those of humans (Fig. 9a,b). We examined the relationship between the recognition accuracy and TMTF similarity indices (i.e., pattern similarity and discrepancy indices).

The TMTF similarity indices correlated with the recognition accuracy in the higher layers (Fig. 9c). The positive and negative correlations, respectively for pattern similarity and discrepancy



**Figure 6.** *a*, Schematic illustration of the AM detection method in a 3IFC trial. Three stimuli were presented to the model, and the probabilities of the stimuli being modulated were estimated for each layer from its unit activities. The probability was estimated independently for each stimulus. The interval with the maximum probability was taken to be the response of the model to the task. It was calculated for each layer. In this example, it is the third interval, which is correct because the third stimulus was modulated. *b*, The boxes labeled AM detection in *a* are expanded for a detailed illustration of the probability estimation method. Logistic regression was applied to the time-averaged unit activities in a single layer.  $N$  denotes the number of units in the layer. *c*, An example of a psychometric curve obtained from a single layer. The proportion of correct trials (filled circles) was fitted with an asymmetric sigmoid curve (solid line). The detection threshold (vertical dotted line) was defined as the AM depth at a 0.707 correct proportion (horizontal dotted line).

indices, mean that the models that performed sound recognition better exhibited AM sensitivity more similar to that of humans. This result further supports the idea that there is a strong relationship between optimization for natural sound recognition and emergent AM sensitivity.

### Training signals must have natural AM patterns for the emergence of human-like AM sensitivity

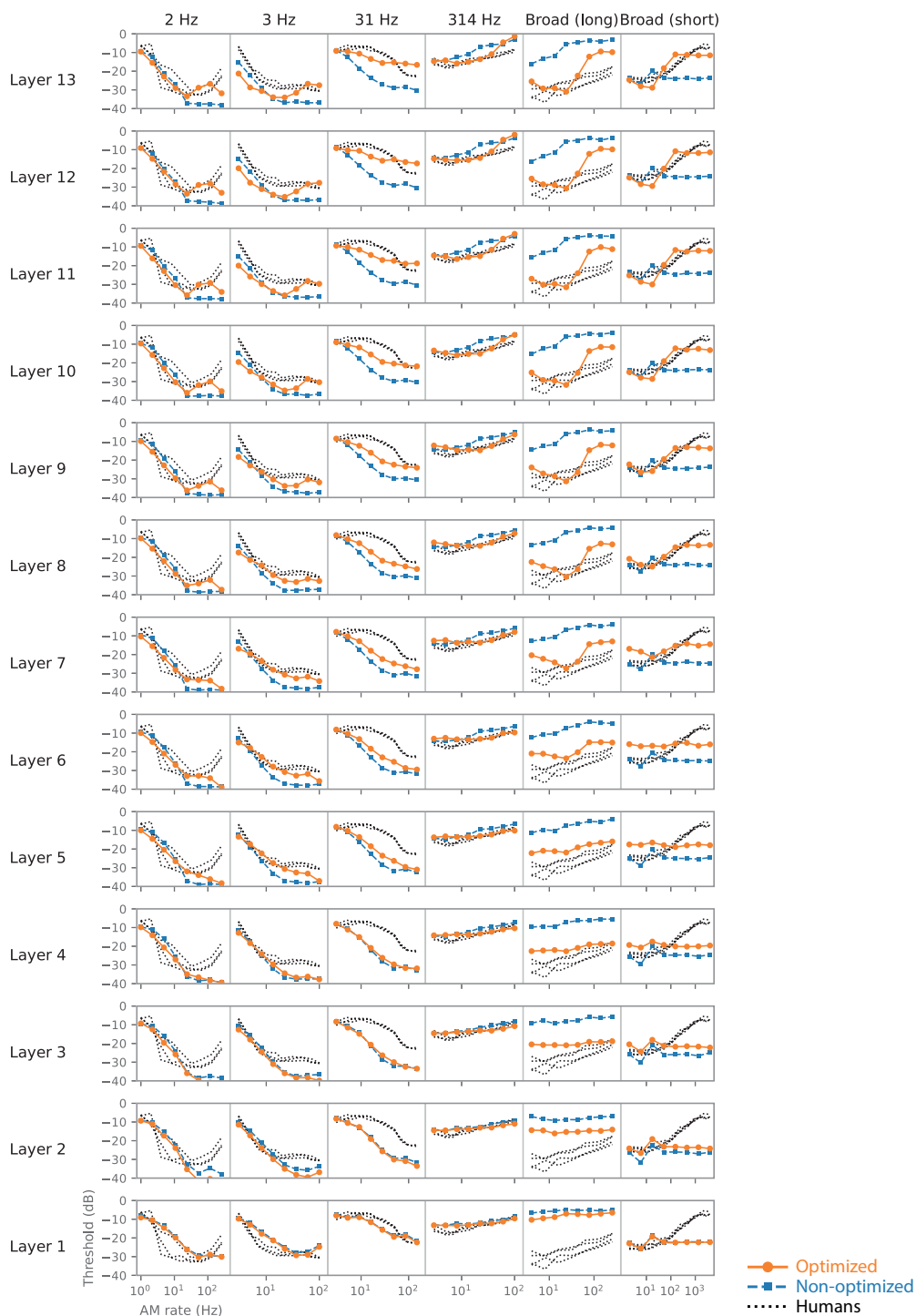
What components of the optimization for natural sound recognition are essential for acquiring human-like AM sensitivity? We hypothesized that natural AM patterns in the sound are the critical feature. To test this hypothesis, we conducted control experiments in the models optimized for the manipulated sound signals. Only the training signals were manipulated. No modifications were made to the stimuli for measuring AM sensitivity.

The manipulation involved dividing a sound signal into an amplitude envelope and TFS and disrupting either of them while preserving the other for the entire signal or each sub-band. This is a common strategy for manipulating the AM structure of

sound in auditory science (Smith et al., 2002; Lorenzi et al., 2006). Specifically, we tested the following four types of signals (see above, Materials and Methods for details): single-band Env signals, which preserved Env while TFS was replaced with that of broadband noise in the entire signal; multiband Env signals in which the original signal was divided into multiple frequency bands, and Env for each band was preserved while TFS was replaced with a random narrowband noise corresponding to that band, then the multiband signals were added together; single-band TFS signals, which, for the entire signal preserved TFS while Env was flattened; and multiband TFS signals in which the original signal was divided into multiple frequency bands, and TFS for each band was preserved while Env was flattened, then the multiband signals were added together.

We optimized the NN models to recognize the manipulated sounds. Hereafter, we call the optimized models for the above signals the single-band Env model, multiband Env model, single-band TFS model, and multiband TFS model. We refer to the model trained for intact sounds (i.e., the one described in the earlier sections) as the original model.

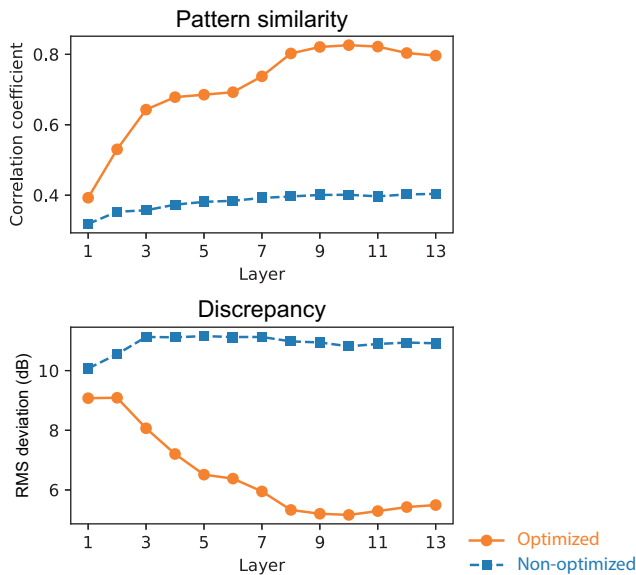




**Figure 7.** TMTFs in the model optimized to everyday sounds (orange circles), those in the nonoptimized model (blue squares), and those in humans (black dotted lines). The columns correspond to different experimental conditions, and the rows correspond to the different layers. The TMTFs in the higher layers of the optimized model appear to be more similar to those of humans than those of the lower layers or the nonoptimized model.

Differences in the TMTFs across the models were more apparent for higher layers. TMTFs of the single-band ENV model appeared to be closest in general shape to the human TMTFs (Fig. 10a, blue lines). The pattern similarity index for the single-band Env model was at a comparable level to the original model throughout its layers (Fig. 10b, top). However, the discrepancy index of the single-band Env model deviated from the original one in the layers above the eighth, exhibiting higher values (Fig. 10b, middle). These results indicate that the model had TMTFs

whose patterns were similar to those of humans, while its sensitivity to AM was higher than that of humans (i.e., lower thresholds; the blue lines were generally lower than the black dotted lines; Fig. 10a). This difference in AM sensitivity was quantified as the net difference, the average signed difference between the model TMTFs and the human TMTFs (Fig. 10b, bottom). The net difference was largely negative in the single-band Env model, indicating that its thresholds were on average lower than that of humans. These results suggest that optimizing to sounds that



**Figure 8.** Quantitative comparison of the TMTFs in the model and humans. Pattern similarity index (top) and discrepancy index (bottom) in the models optimized to everyday sounds (orange circles) and the nonoptimized models (blue squares) are shown. The relatively higher layers of the optimized models show large pattern similarity and small discrepancy. The lower layers and the nonoptimized models show low similarity.

only retain natural single-band AM patterns made the model more sensitive than humans to AM. This is probably because the model was biased toward exploiting the AM that was the only available feature for recognition.

The TMTFs of the multiband Env model were somewhat similar to those of humans. Both pattern similarity and discrepancy indices gradually approached the original model with increasing layer numbers, reaching a comparable level at the 12th and 13th layers. In contrast, the TMTFs of the single-band and multiband TFS models were consistently different from those of humans. This result, together with the results of the two Env models, suggests that natural AM patterns are essential for the emergence of human-like AM sensitivity.

It is important to note that all models achieved sufficiently high accuracy in the sound recognition task, well above chance level, and that the accuracies of the single-band Env and TFS models were comparable (Fig. 11). This indicates that all the signals contained a sufficient amount of information for sound recognition and that all the models were capable of using the information. Thus, the results of TFS models exhibiting TMTFs that were not similar to humans could not be attributed to failures in their optimization.

### Comparison with temporal-template-based AM detection strategy

In this study, we assumed AM detection is based on time-averaged activities in the model. On the other hand, previous computational studies used a method based on the temporal template for simulating the psychophysical AM detecting process (Dau et al., 1997a). There might be a possibility that different detection strategies yield different forms of TMTF. In Dau et al.'s (1997a) study, template-based AM detection was performed on the outputs of the MFB. A template for a given AM rate was generated by averaging the MFB outputs over multiple independent carrier instances. In an *x*IFC trial, a correlation coefficient was

calculated between the template and the MFB output for a test signal. The stimulus interval with the highest correlation was chosen as the response of the model. Here, to test whether the same detection method works well for our NN model, we applied it to the unit activities in a single NN layer (Fig. 12a). A template was defined as the average unit activities in response to fully modulated stimuli minus the average activities in response to nonmodulated stimuli. The average was taken over multiple carrier instances. Then, in each trial of the *x*IFC task, the response of the model was defined as the stimulus interval with the largest correlation between the activities of the model and the template. In accordance with the previous study (Dau et al., 1997a), the correlation was non-normalized, that is, the dot product.

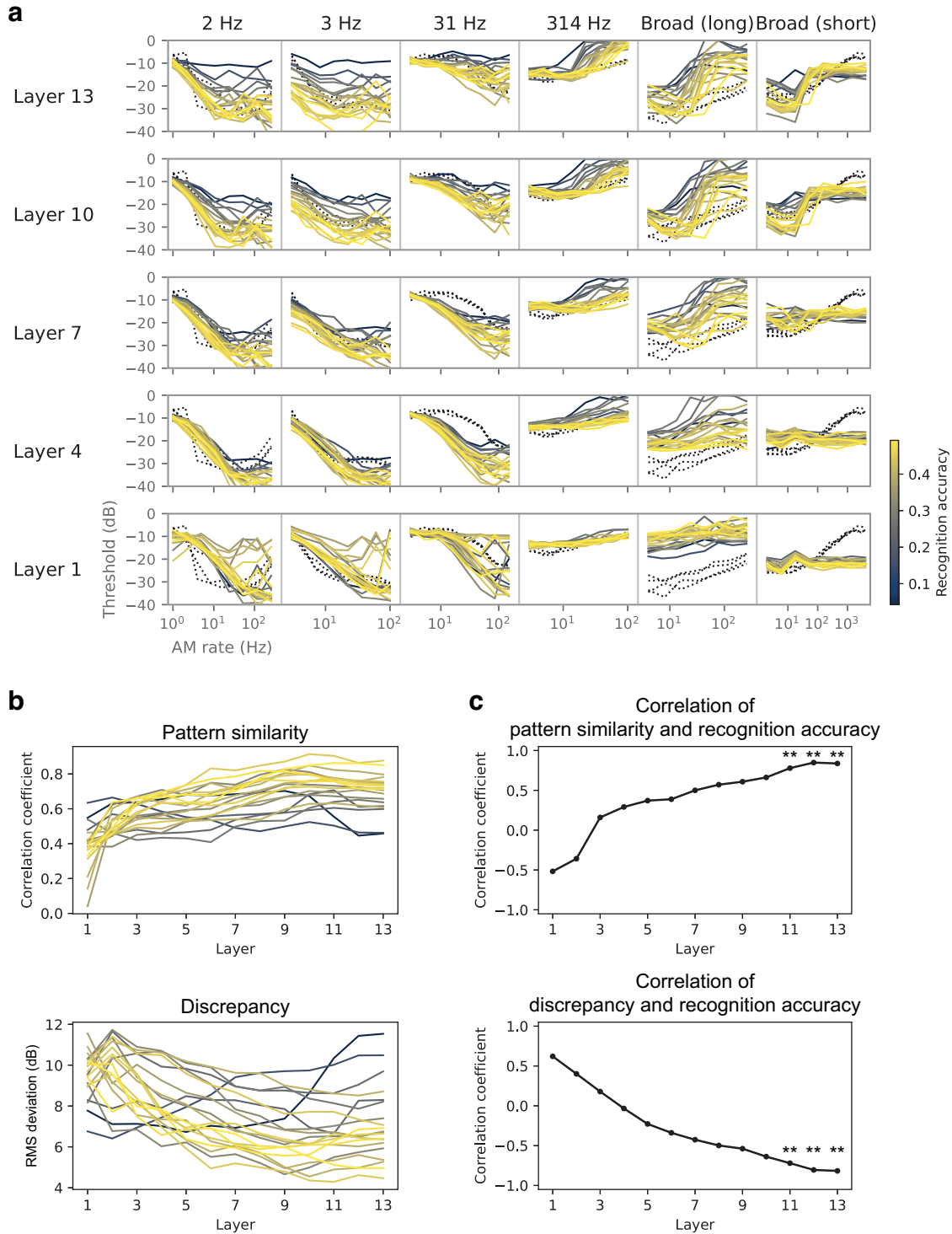
The resulting TMTFs differed from human TMTFs (Fig. 12b). The similarity to human TMTFs was the highest at the top-most layers (Fig. 12c, open circles), but it was still lower than the similarity of the TMTFs of the time-average-based detector (Fig. 12c, gray circles; same data as in Fig. 8). This result indicates that human-like AM sensitivity was not observed in our model when it used temporal correlation with templates for AM detection. Template-based detection might work well for MFB outputs but not for NN activities. Our results alone could not elucidate the reason for this difference. Perhaps different AM detection strategies should be applied to different sound representations (in an NN or in an MFB output).

### Neurophysiology of the model suggested involvement of the auditory midbrain and higher regions

In an earlier section, we indicated that human-like AM sensitivity was observed when the stimulus representation in layers around the 10th layer was used for AM detection (Fig. 8). Does this finding provide a significant insight into neural processes in the ANS? More specifically, to which brain regions do these layers correspond?

We addressed this question by mapping the NN layers onto the auditory brain regions based on the similarity of their neurophysiological AM tuning, as in our previous study (Koumura et al., 2019). Neurophysiological AM tuning in the model was measured by simulating neurophysiological experiments. AM tuning in a unit was calculated from its response to modulated white noise. It was characterized by its best rate (the AM rate with the maximum tuning value) and upper cutoff rate (the AM rate at which the tuning curve starts to decrease). The distribution of the best and upper cutoff rates in each layer was compared with that in the ANS to yield the similarity between the NN layers and the brain regions.

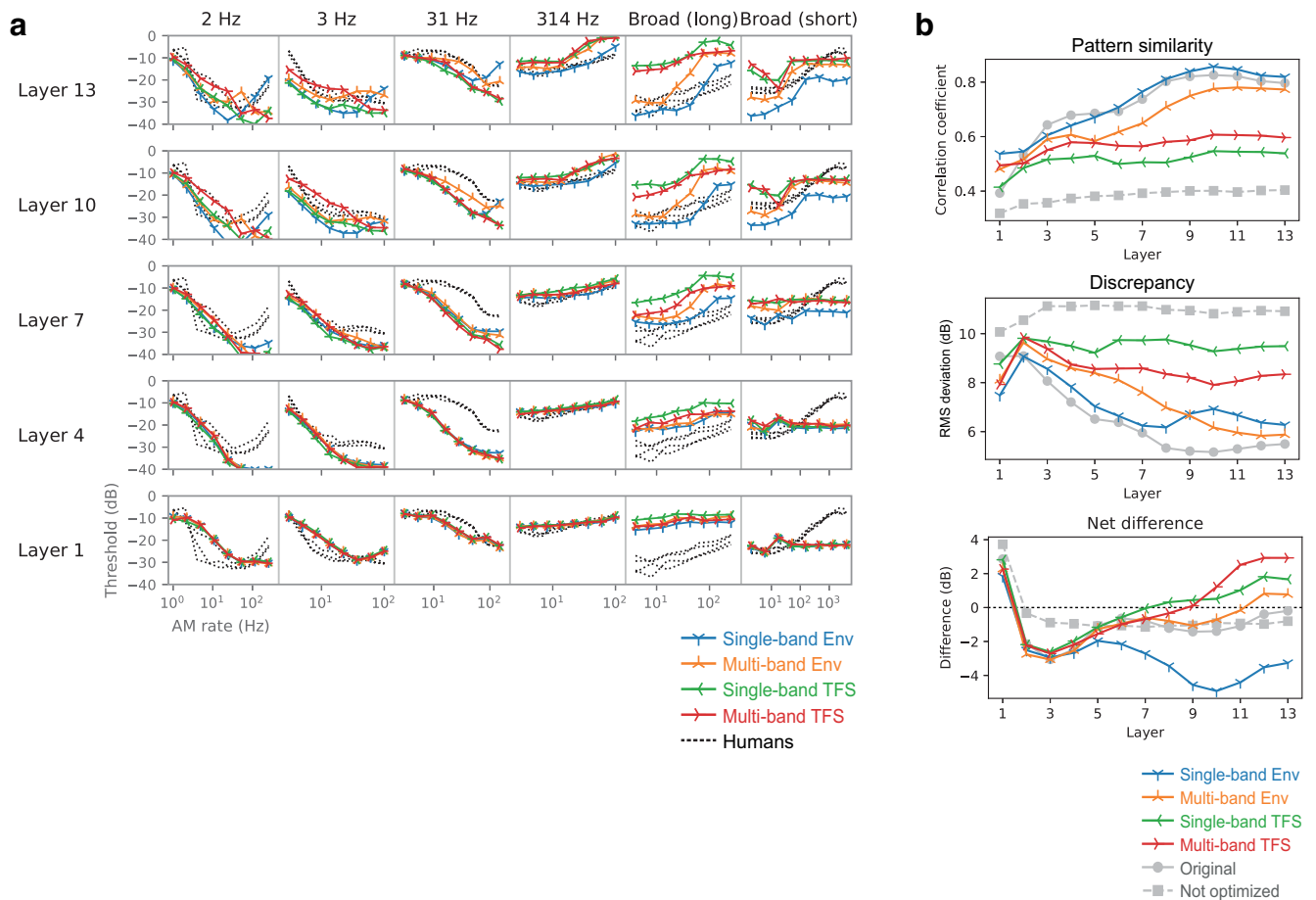
The neuronal tuning properties in the ANS were taken from the neurophysiological literature (Müller-Preuss, 1986; Langner and Schreiner, 1988; Schreiner and Urbas, 1988; Batra et al., 1989; Frisina et al., 1990; Preuss and Müller-Preuss, 1990; Rhode and Greenberg, 1994; Zhao and Liang, 1995; Bieser and Müller-Preuss, 1996; Condon et al., 1996; Schulze and Langner, 1997; Eggermont, 1998; Huffman et al., 1998; Joris and Smith, 1998; Joris and Yin, 1998, 1992; Kuwada and Batra, 1999; Krishna and Semple, 2000; Lu and Wang, 2000; Lu et al., 2001; Liang et al., 2002; Batra, 2006; Zhang and Kelly, 2006; Bartlett and Wang, 2007; Scott et al., 2011; Yin et al., 2011). The target brain regions from peripheral to central were auditory nerves (AN), the cochlear nucleus (CN), superior olivary complex (SOC), nuclei of the lateral lemniscus (NLL), inferior colliculus (IC), medial geniculate body (MGB), and auditory cortex (AC). AM tuning in the ANS is often described in terms of the spike synchrony to the



**Figure 9.** AM sensitivity in different architectures and its relationship to recognition performance. **a**, TMTFs in the models with different architectures. Each colored line shows results for a single model with a specific choice of NN architecture. The color indicates the recognition accuracy (legend at right) of the corresponding architecture. Black dotted lines show human TMTFs. **b**, Pattern similarity and discrepancy indices. **c**, Correlation coefficients between the (dis)similarity indices and the recognition accuracy. Statistically significant positive and negative correlations were found in the highest layers;  $**p < 0.01$  with a Bonferroni correction for the number of layers.

stimulus envelope and the average spike rate during stimulus presentation. The properties of the AM tuning transform along the peripheral to central pathway (Joris et al., 2004; Sharpee et al., 2011). The spike synchrony is tuned to a higher AM rate in the peripheral regions and to a lower AM rate in the central regions. The average firing rate is not tuned to the AM in the peripheral regions but is tuned in the central regions.

The neurophysiological similarity of the NN layers and brain regions shows that lower and higher layers were relatively similar to the peripheral and central brain regions, respectively (Fig. 13). Thus, the results of our previous study were replicated with the newly constructed model. Layers around the 10th layer roughly corresponded to the IC, MGB, and AC. This result indicates that the NN layers that exhibited human-like AM sensitivity had



**Figure 10.** *a*, TMTFs of the models optimized to degraded sounds. *b*, Their pattern similarity index (top), discrepancy (middle), and net difference from humans (bottom) are shown. The indices of the original optimized and nonoptimized models are shown as gray lines. Overall, in the higher layers, the TMTFs of the Env models were more similar to those of humans than were the TMTFs of the TFS models. Single-band Env models exhibited high pattern similarity but also showed a high discrepancy, indicating that the patterns of the TMTFs, but not their absolute values, were similar to those of humans. Their thresholds appeared to be lower than those of humans, as shown by the negative net difference.

similar neural representations to those of the IC and higher brain regions.

#### Emergent TMTFs through optimization to speech sounds

We conducted the same analysis on the models optimized for phoneme classification of speech sounds (<https://doi.org/10.35111/17gk-bn40>). The results were generally consistent with those of the models optimized for everyday sounds reported above. This indicates that human-like TMTFs robustly emerged in the models that were independently optimized to two different types of sound.

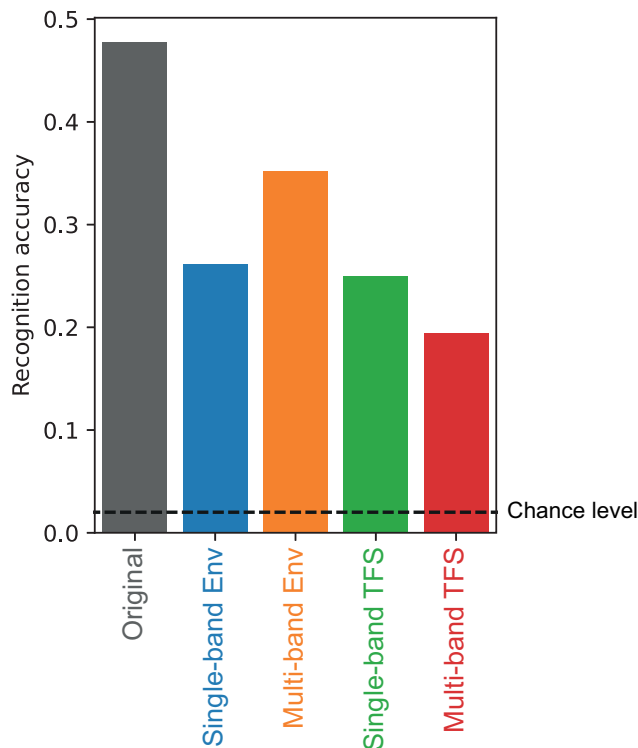
The phoneme classification accuracy was 0.747, which is high above the chance level, 0.026. The TMTFs of the layers around the eighth and ninth layers in the optimized model were similar to those of humans (Fig. 14*a,b*, orange lines), whereas neither the TMTFs of the nonoptimized model nor those calculated with the template correlation were similar. According to the neurophysiological analysis, the brain region most similar to the eighth and ninth layers was the IC (Fig. 14*c*). The MGB and AC also exhibited high neurophysiological similarity.

On the other hand, we did not observe high correlations between similarity to human TMTFs and recognition accuracy. The pattern similarity indices and the discrepancy indices did not exhibit appreciable variations across the 20 models with different architectures (Fig. 15*a*), and there was no significant correlation with recognition accuracy (Fig. 15*b*). This lack of

correlation may be explained by the dynamic range of the recognition accuracy in the models for speech sounds: Their recognition accuracies ranged from 0.499 to 0.771, whereas, for the models optimized to everyday sounds, the range was from 0.043 to 0.492. All models trained on speech sounds showed recognition accuracies well above chance level, whereas some models trained on everyday sounds exhibited very low recognition accuracy almost as low as chance level. Probably, the correlations between the recognition accuracy and the similarity to human TMTFs are nonlinear. Models that failed to perform sound recognition exhibited AM sensitivity that was not similar to humans, but models with good sound recognition performance exhibited more or less human-like AM sensitivity regardless of the small variation in their performance. Probably, once the performance of sound recognition surpasses a certain level, the similarity of AM sensitivity to humans does not change largely on any further increase in recognition accuracy.

When the Env or TFS was disrupted in the speech sounds, the recognition accuracies of the optimized models were well above the chance level, except for the multiband TFS model (Fig. 16*a*). The multiband Env model had the highest recognition accuracy, followed in order by the single-band TFS model, single-band Env model, and multiband TFS model. This order is consistent with human performance as shown in a previous study (Smith et al., 2002) and thus supports the conclusion that our models behave similarly to humans when recognizing those





**Figure 11.** Recognition accuracy of models optimized to degraded sounds. The result of a model optimized to the original sounds is also shown on the left. Generally, the recognition accuracy of the model dropped when it was optimized to degraded sounds, but the drop was not catastrophic.

degraded speech sounds. The multiband TFS model trained on everyday sounds showed relatively low recognition accuracy, although it was above the chance level (Fig. 11).

The TMTFs of the models optimized to degraded speech sounds were qualitatively consistent with those of the models optimized to degraded everyday sounds (Fig. 16*b*). The TMTFs of the single-band Env model exhibited high pattern similarity to those of humans, and their net difference shows their thresholds were lower than those of humans. The TMTFs of the multiband Env model were similar to those of humans in terms of both their pattern and absolute value. The TMTFs in the TFS models were not similar to those of humans.

## Discussion

### AM sensitivity emerging through sound recognition

The present study demonstrated that a model optimized to recognition of sounds with natural AM statistics exhibited human-like AM sensitivity. In building the model, we did not make any attempt to design model architectures or adjust parameters for achieving similarity to humans, nor did we use any knowledge about human AM sensitivity. The results therefore suggest that the nature of AM sensitivity in humans might also be a consequence of optimizing to natural sound recognition in the course of evolution and/or development in the natural environment and that it would not have emerged if the input sounds would have had different AM characteristics.

By simulating previous human experiments as precisely as possible, we could quantitatively compare the TMTFs of our model and of humans. We explained TMTFs in six experiments conducted in three independent human studies with a single unified framework. Our findings strengthen the existing knowledge

that general AM sensitivity is closely linked to sound recognition ability (Cazals et al., 1994; Fu, 2002; Luo et al., 2008; Won et al., 2011; De Ruiter et al., 2015; Bernstein et al., 2016).

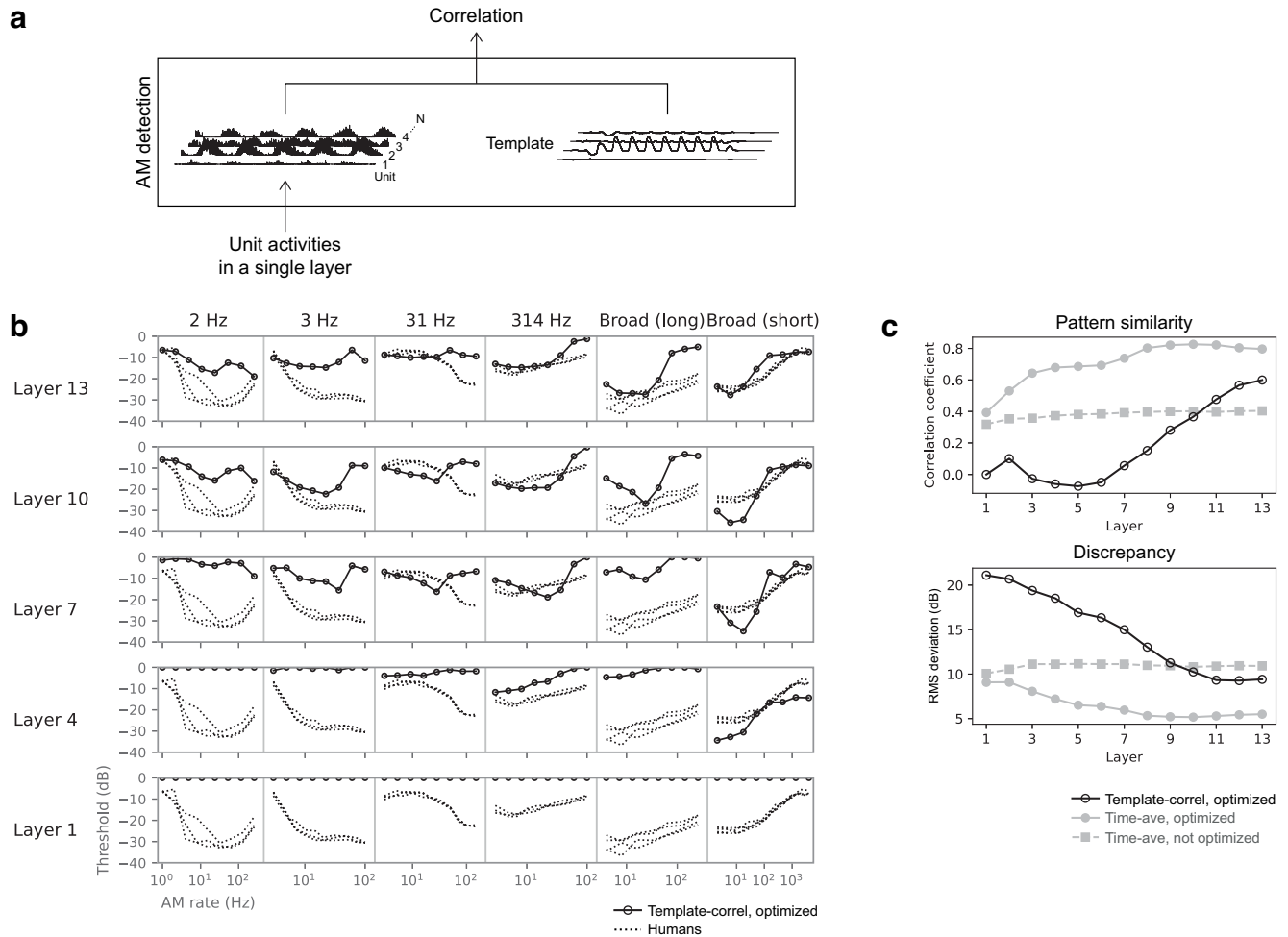
We built two types of models using either everyday sounds or speech sounds and analyzed each one independently. The results on the two datasets were qualitatively similar, although some relationships were stronger for everyday sounds. They suggested that human-like AM sensitivity is related to both sound types. This conclusion is consistent with the previous studies on cochlear frequency tuning and neural AM tuning, where qualitatively similar tunings were obtained from optimization to everyday sounds and speech sounds (Smith and Lewicki, 2006; Koumura et al., 2019). There might be a common representation of these sound types in the auditory system perhaps because the human auditory system has taken advantage of already evolved mechanisms to represent everyday sounds and built speech recognition functions on top of it.

### Relation to modulation filter bank theory

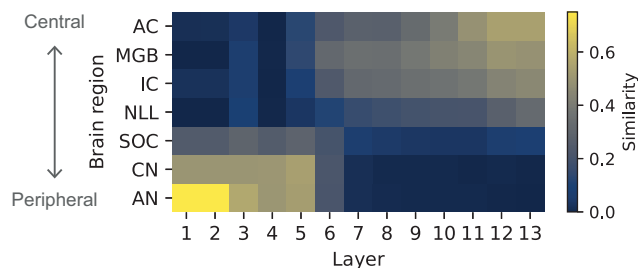
The MFB is a conceptual realization of midbrain neurons that are tuned to the modulation rate. It was formalized to explain various psychoacoustic phenomena, for example, frequency selectivity in the modulation domain (Dau et al., 1997*a,b*). Our NN model, in contrast, does not include any explicit implementation of auditory mechanisms (e.g., a cochlear filter bank or an MFB), nor does it attempt to reproduce any psychophysical phenomena (e.g., modulation masking). This is because the purpose of our study is not to delve into the auditory signal processing mechanisms but to investigate emergent TMTFs in a sound-recognition model and the effect of optimization and sound features to be optimized on the emergent AM sensitivity. Thus, it is difficult to make a fair comparison between models based on MFB theory and those in the present study. Nevertheless, we consider it worth discussing the present findings in relation to MFB theory.

Compared with previous computational studies involving the MFB (Jepsen et al., 2008), the TMTFs of our model were less similar to human TMTFs. The TMTFs in Jepsen et al. (2008) showed a pattern similarity index of 0.95 and a discrepancy index of 2.3 dB (derived from Jepsen et al., 2008, their Fig. 8; but there is a caveat as they did not test a 2 Hz bandwidth or long broadband conditions). This, however, is not surprising and does not indicate the inferiority of our model. That is, the previous study designed the model explicitly for reproducing human psychophysical properties including AM sensitivity, whereas our model does not explicitly try to reproduce any psychophysical or neurophysiological properties.

Our previous study showed that middle layers in an optimized NN (as the nuclei in the ANS) exhibited units with tuning to various AM rates, some of which probably work as modulation-domain filters (Koumura et al., 2019). This alone, however, does not guarantee that the units in the NN (and even the auditory neurons) function as the MFB model that predicts human's carrier-dependent TMTFs. It should be repeated that our previous study evaluated similarities only with the nonhuman ANS, and that both the NN and the ANS contained not only band-pass-like units (as assumed in the MFB) but also multi-peaked and broadly tuned units. Thus, the present study is not an (indirect) replication of our previous study or the MFB study but places the AM tuning in a broader context under considerably different configurations from those in the MFB theory. The emergence of human-like TMTFs in our configurations suggested that



**Figure 12.** *a*, Schematic illustration of the AM detection process based on correlation with a template. For the purpose of explanation, this illustration replaces Figure 6*b*, where the correlation in this figure corresponds to the output probability in Figure 6*b*. *b*, TMTFs obtained from AM detection based on template correlation (open circles). TMTFs of humans are shown as dotted lines. *c*, Pattern similarity index and discrepancy index from the template-based detector (black open circles). The similarity indices for the time-average-based detector (Fig. 8) are shown as filled symbols. AM detection based on template correlation did not result in human-like TMTFs.



**Figure 13.** Similarity of the neurophysiological tuning between brain regions and NN layers. Layers that showed TMTFs similar to those of humans roughly correspond to higher regions like the IC, MGB, and AC.

the NN may have implicitly acquired through optimization a function equivalent to an MFB.

Although we did not implement hardwired cochlear filters, our previous study suggested that lower layers in the optimized NN conducted some kind of initial frequency analysis (Koumura et al., 2019, their Fig. 15). Their frequency responses were multi-peaked but not sharply tuned to a single frequency as in the auditory periphery. Measuring TMTFs in a model with explicit implementation of ANS-like peripheral filters would be an interesting future work, considering that pitch-related psychophysical behavior is more

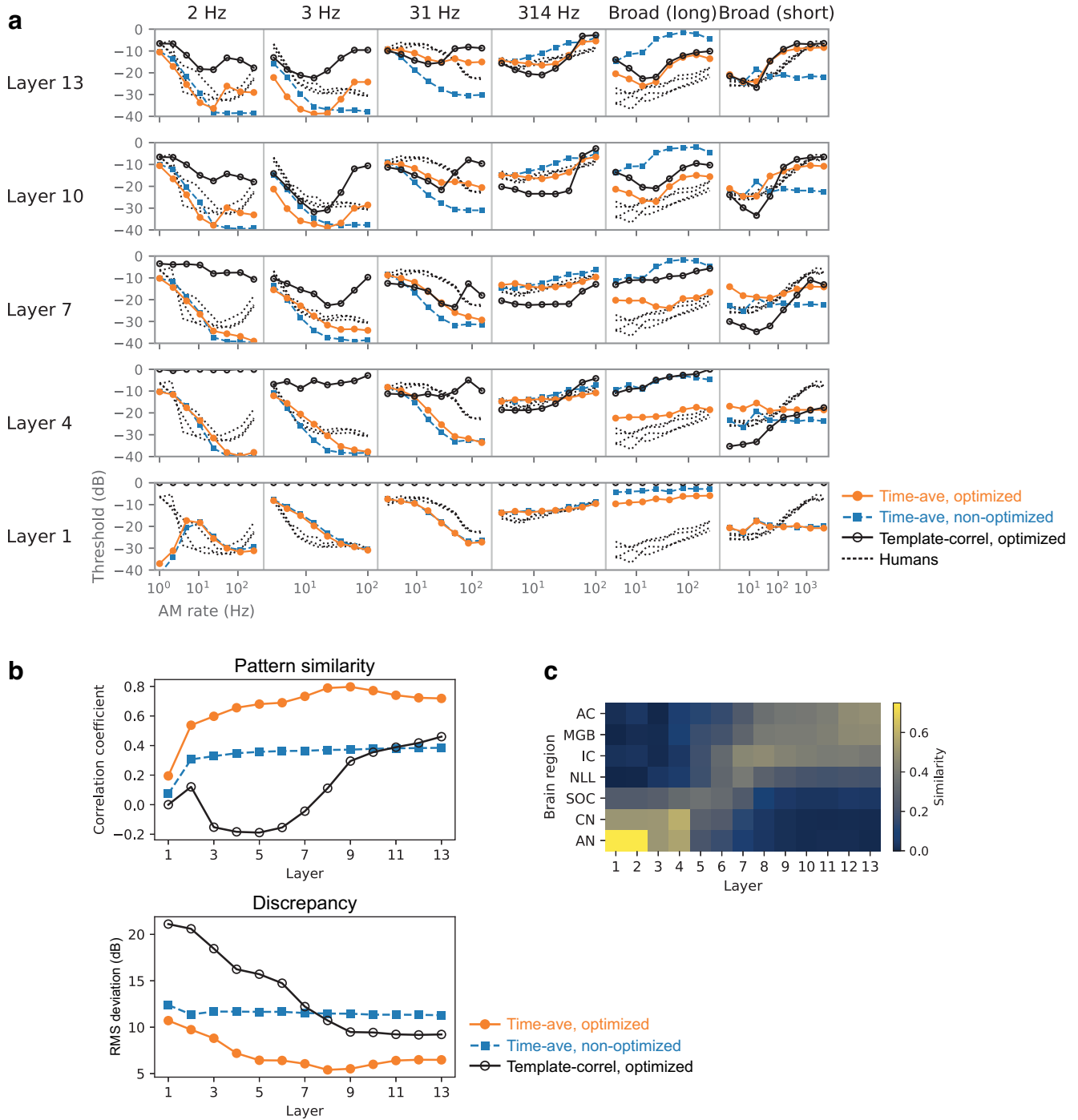
similar to humans in the model with hardwired ANS-like peripheral filters (Saddler et al., 2021).

In the MFB models, additive noise was applied to the internal representation of the model from which AM detection was simulated. In contrast, our model is deterministic, meaning that the unit responses are the same for the same stimulus, except for the nondeterministic behaviors of atomic operations in a graphics processing unit. In our AM detection experiments, the only source of stochasticity was the noise carrier and the AM starting phase in the conditions with a random starting phase, both of which were sampled independently stimulus by stimulus (Table 2). Psychophysically relevant internal noise (as proposed by Ewert and Dau, 2004) could increase the similarity of the TMTFs of the model to human ones.

### Neural mechanisms of behavioral AM sensitivity

#### Hierarchical brain regions and AM detection

In the optimized model, TMTFs in the middle to higher layers were most similar to humans' TMTFs. Here, we discuss the implications of this result in terms of signal processing and anatomic brain regions. Generally, an optimized NN behaves as an effective signal processor and feature extractor. While processing an input signal, each of its cascading layers computes its representation by integrating and nonlinearly transforming

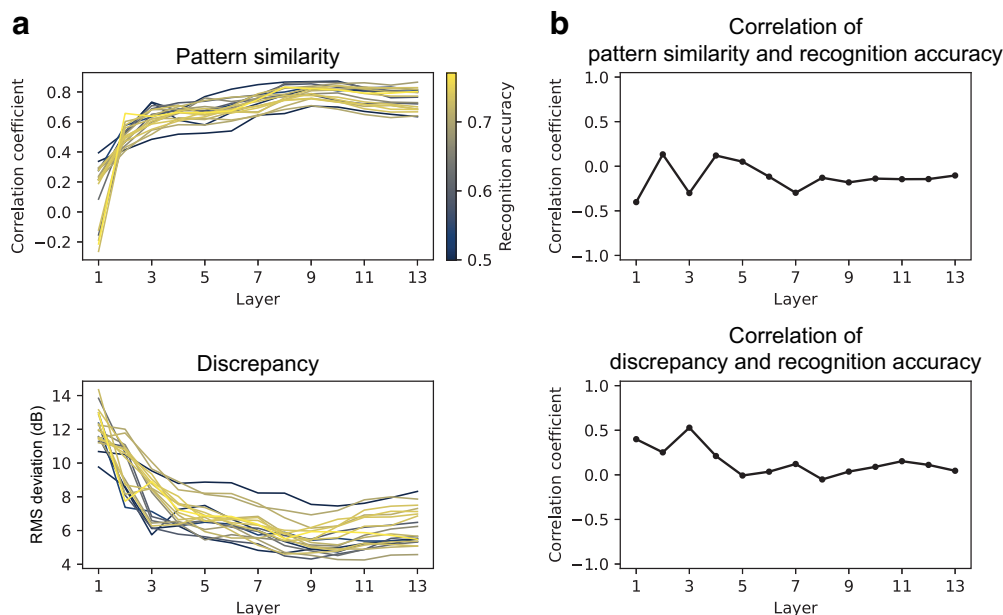


**Figure 14.** Results of the model optimized to speech sounds. **a**, TMTFs of the optimized model with the AM detection process based on time-averaged unit activities (orange circles), those of the nonoptimized model (blue squares), those from AM detection based on temporal correlation with templates (black open circles), and those of humans (black dotted lines). **b**, Pattern similarity index and discrepancy index between the model TMTFs and human TMTFs. **c**, Similarity of neurophysiological tuning between NN layers and auditory brain regions.

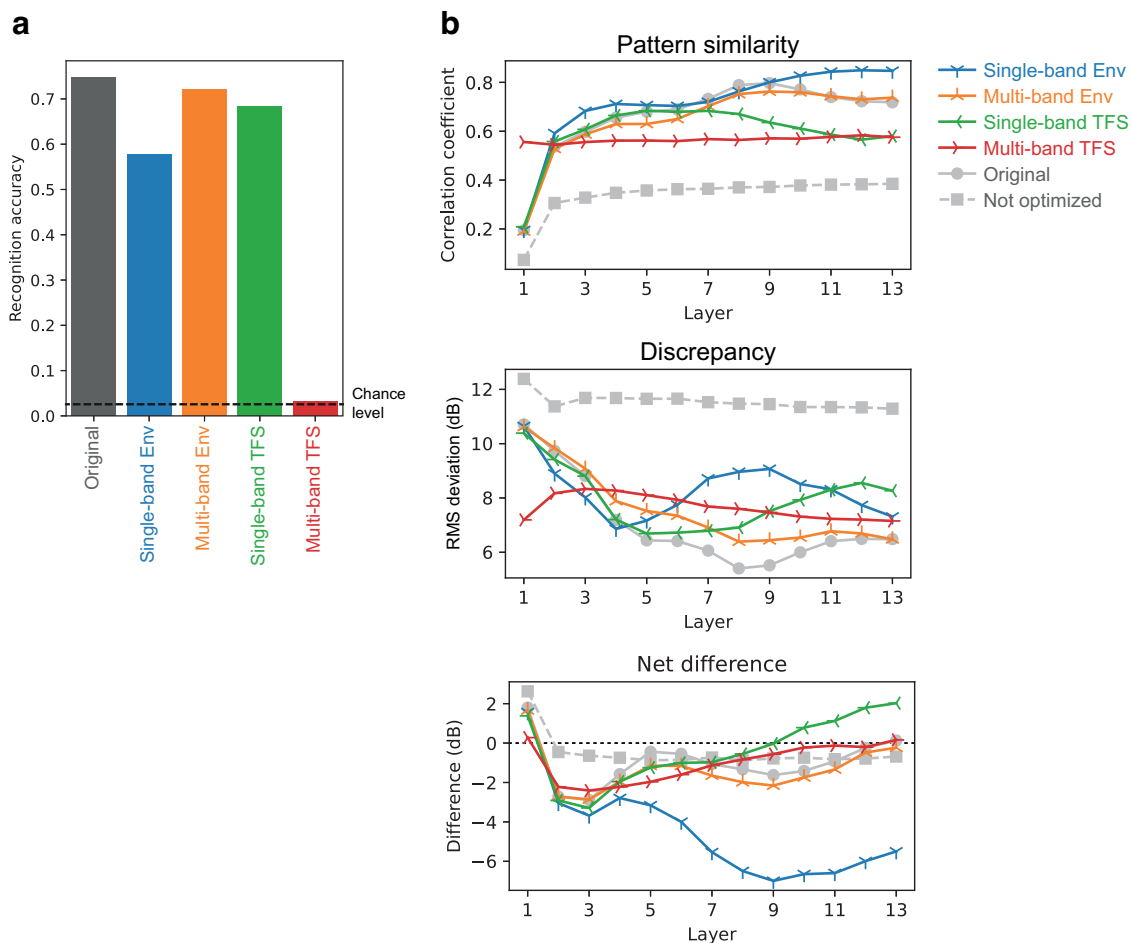
the one below. Its lower layers compute relatively simple and temporally and/or spatially local features. As the processing stage progresses, the extracted features gradually become more complex and global (Mahendran and Vedaldi, 2015; Yosinski et al., 2015). Therefore, our results suggest that the AM detection ability of humans might be based on relatively higher-order features of the stimulus.

Interestingly, a similar tendency can also be seen in the ANS. Peripheral regions in the auditory pathway are generally sensitive to fast temporal changes in a sound and relatively linear features, whereas central regions are sensitive to slower changes and more

nonlinear features (Joris et al., 2004; Sharpee et al., 2011). In the present study, we found that layers with human-like psychophysical TMTFs showed neurophysiological AM tuning similar to that in the IC, MGB, and AC (Figs. 13, 14c). This result suggests that human-like TMTFs could be observed when conducting an AM detection task from the stimulus representation in these brain regions. A human brain might also use stimulus representations in these regions when conducting AM detection tasks, but our results alone cannot distinguish whether such a computation is running within those brain regions or somewhere outside, possibly regions associated with higher-order cognitive



**Figure 15.** *a*, Pattern similarity and discrepancy indexes of the TMTFs in the models with different architectures optimized to speech sounds. The pattern of the similarity indexes appeared similar across different NN architectures. *b*, Correlation between similarity indexes and recognition accuracy. Significant correlation was not observed. These results are probably because of the small dynamic range of recognition accuracy.



**Figure 16.** *a*, Recognition accuracy of the models optimized to degraded speech sounds. Left, The result for the model optimized to original speech sounds is also shown. Recognition accuracy dropped when the models were optimized to degraded sounds, but the drop was not catastrophic except in the multiband TFS model. *b*, Pattern similarity indices (top), discrepancy indices (middle), and net difference (bottom) between model TMTFs and human TMTFs. The results are consistent with those of the models optimized to everyday sounds.



functions. Another unanswered question is which (possibly all?) of these regions is actually the source of the neural representation used by the AM detector (if it exists) implemented in the human brain. Nevertheless, the present results, at least, suggest that the stimulus representation necessary for AM detection emerges as early as in the IC and is kept until the signal reaches the AC.

### Neural AM representation relevant to AM detection

We obtained human-like TMTFs by assuming an AM detection process based on time-averaged unit activities. Time-averaged unit activities can be interpreted as the average neuronal firing rate (Koumura et al., 2019). Together with the above discussion, it can be suggested that the behavioral AM sensitivity of humans might be based on the average firing rate in the IC, MGB, and AC. This is consistent with the previous neurophysiological findings that relatively central auditory brain regions perform rate coding of AM (Joris et al., 2004).

The other AM coding strategy in the ANS is temporal coding (Joris et al., 2004). We also tested temporal-coding-based AM detection by performing the temporal template correlation method, but findings were elusive. Our results alone could not distinguish whether the brain relies on temporal coding when performing AM detection. Our template-based detection strategy might have been too simple to simulate a temporal-coding-based AM detection process in the human brain. Using a more sophisticated detection process (Bashivan et al., 2019) might result in more human-like TMTFs.

### Differences among experimental conditions

The TMTFs in the lower layers were less similar to those in humans, and as the layer number went higher, the similarity also became higher (Figs. 8, 14*b*). Detailed inspections of the individual stimulus parameters, however, indicate that changes in the form of TMTFs along the layers appeared to vary with the stimulus parameters. The change seems most prominent in the broadband carrier conditions, and least in the condition with the 314 Hz carrier bandwidth, showing almost constant TMTF forms across layers (Figs. 7, 14*a*). In this study, we could not see a consistent relationship between the TMTF difference across layers and the stimulus parameters. There is a possibility that humans perform AM detection with different strategies in different experimental conditions. These results highlight the importance of testing multiple stimulus parameters when investigating AM sensitivity.

### Analyzing machine learning models with a combination of psychophysics and neurophysiology

From a machine learning point of view, this study can be viewed as an attempt to understand an NN with a combination of psychophysical and neurophysiological methods. A number of methods have been proposed for analyzing the behavior and stimulus representations of an NN (Montavon et al., 2018; Cammarata et al., 2020). It would be important to analyze an NN from a variety of perspectives. We can learn from a tradition of psychophysical and neurophysiological studies that have established various methods to investigate complicated biological systems (Eijkman, 1992; Leibo et al., 2018; Barrett et al., 2019; RichardWebster et al., 2019). The present study demonstrated the utility of multidisciplinary analysis on a single platform.

## References

- Ashihara T, Moriya T, Kashino M (2021) Investigating the impact of spectral and temporal degradation on end-to-end automatic speech recognition performance. *Proc Interspeech* 2021:1757–1761.
- Barrett DG, Morcos AS, Macke JH (2019) Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Curr Opin Neurobiol* 55:55–64.
- Bartlett EL, Wang X (2007) Neural representations of temporally modulated signals in the auditory thalamus of awake primates. *J Neurophysiol* 97:1005–1017.
- Bashivan P, Kar K, DiCarlo JJ (2019) Neural population control via deep image synthesis. *Science* 64:eaav9436.
- Batra R (2006) Responses of neurons in the ventral nucleus of the lateral lemniscus to sinusoidally amplitude modulated tones. *J Neurophysiol* 96:2388–2398.
- Batra R, Kuwada S, Stanford TR (1989) Temporal coding of envelopes and their interaural delays in the inferior colliculus of the unanesthetized rabbit. *J Neurophysiol* 61:257–268.
- Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *J Mach Learn Res* 13:281–305.
- Bergstra J, Boulevarde EHL, Yamins DLK, Cox DD, Boulevarde EHL (2013) Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. Paper presented at the 30th International Conference on Machine Learning, Atlanta, June.
- Bernstein JGW, Danielsson H, Hällgren M, Stenfelt S, Rönnberg J, Lunner T (2016) Spectrotemporal modulation sensitivity as a predictor of speech-reception performance in noise with hearing aids. *Trends Hear* 20:233121651667038.
- Bieser A, Müller-Preuss P (1996) Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. *Exp Brain Res* 108:273–284.
- Bruce IC, Erfani Y, Zilany MSA (2018) A phenomenological model of the synapse between the inner hair cell and auditory nerve: implications of limited neurotransmitter release sites. *Hear Res* 360:40–54.
- Cammarata N, Carter S, Goh G, Olah C, Petrov M, Schubert L, Voss C, Egan B, Lim SK (2020) Thread: Circuits. *Distill*. Available at: <https://doi.org/10.23915/distill.00024>.
- Cazals Y, Pelizzone M, Saudan O, Boex C (1994) Low-pass filtering in amplitude modulation detection associated with vowel and consonant identification in subjects with cochlear implants. *J Acoust Soc Am* 96:2048–2054.
- Clevert D-A, Unterthiner T, Hochreiter S (2016) Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv:1511.07289*. <https://doi.org/10.48550/arXiv.1511.07289>.
- Condon CJ, White KR, Feng AS (1996) Neurons with different temporal firing patterns in the inferior colliculus of the little brown bat differentially process sinusoidal amplitude-modulated signals. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 178:147–157.
- Dau T, Kollmeier B, Kohlrausch A (1997a) Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J Acoust Soc Am* 102:2892–2905.
- Dau T, Kollmeier B, Kohlrausch A (1997b) Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J Acoust Soc Am* 102:2906–2919.
- De Ruiter AM, Debryne JA, Chenault MN, Francart T, Broxk JPL (2015) Amplitude modulation detection and speech recognition in late-implanted prelingually and postlingually deafened cochlear implant users. *Ear Hear* 36:557–566.
- Derleth RP, Dau T, Kollmeier B (2001) Modeling temporal and compressive properties of the normal and impaired auditory system. *Hear Res* 159:132–149.
- Dudley H (1939) Remaking speech. *J Acoust Soc Am* 11:169–177.
- Eggermont JJ (1998) Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. *J Neurophysiol* 80:2743–2764.
- Eijkman EGJ (1992) Neural nets tested by psychophysical methods. *Neural Networks* 5:153–162.
- Ewert SD, Dau T (2004) External and internal limitations in amplitude-modulation processing. *J Acoust Soc Am* 116:478–490.

- Fekedulegn D, Mac Siurtain MP, Colbert JJ (1999) Parameter estimation of nonlinear growth models in forestry. *Silva Fenn* 33:327–336.
- Fonseca E, Favory X, Pons J, Font F, Serra X (2022) FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Trans Audio Speech Lang Process* 30:829–852.
- Francl A, McDermott JH (2022) Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nat Hum Behav* 6:111–133.
- Frisina RD, Smith RL, Chamberlain SC (1990) Encoding of amplitude modulation in the gerbil cochlear nucleus: I. A hierarchy of enhancement. *Hear Res* 44:99–122.
- Fu Q-J (2002) Temporal processing and speech recognition in cochlear implant users. *Neuroreport* 13:1635–1639.
- Gong Y, Chung Y-A, Glass J (2021) AST: audio spectrogram transformer. [arXiv:2104.01778](https://doi.org/10.48550/arXiv.2104.01778). <https://doi.org/10.48550/arXiv.2104.01778>.
- Gygi B, Kidd GR, Watson CS (2004) Spectral-temporal factors in the identification of environmental sounds. *J Acoust Soc Am* 115:1252–1265.
- Hoshen Y, Weiss RJ, Wilson KW (2015) Speech acoustic modeling from raw multichannel waveforms. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, April.
- Huffman RF, Argeles PC, Covey E (1998) Processing of sinusoidally amplitude modulated signals in the nuclei of the lateral lemniscus of the big brown bat, *Eptesicus fuscus*. *Hear Res* 126:181–200.
- Jepsen ML, Ewert SD, Dau T (2008) A computational model of human auditory signal processing and perception. *J Acoust Soc Am* 124:422–438.
- Joris PX, Yin TCT (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. *J Acoust Soc Am* 91:215–232.
- Joris PX, Smith PH (1998) Temporal and binaural properties in dorsal cochlear nucleus and its output tract. *J Neurosci* 18:10157–10170.
- Joris PX, Yin TCT (1998) Envelope coding in the lateral superior olive. III. Comparison with afferent pathways. *J Neurophysiol* 79:253–269.
- Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. *Physiol Rev* 84:541–577.
- Kanwisher N, Khosla M, Dobs K (2023) Using artificial neural networks to ask ‘why’ questions of minds and brains. *Trends Neurosci* 46:240–254.
- Khatami F, Escabi MA (2020) Spiking network optimized for word recognition in noise predicts auditory system hierarchy. *PLOS Comput Biol* 16:e1007558.
- Klein A, Falkner S, Springenberg JT, Hutter F (2017) Learning curve prediction with Bayesian neural networks. Paper presented at the Fifth International Conference on Learning Representations, Toulon, France, April.
- Koumura T, Terashima H, Furukawa S (2019) Cascaded tuning to amplitude modulation for natural sound recognition. *J Neurosci* 39:5517–5533.
- Koumura T, Terashima H, Furukawa S (2020) “Psychophysical” modulation transfer functions in a deep neural network trained for natural sound recognition. *Proceedings of the International Symposium on Auditory and Audiological Research* 7:157–164.
- Kriegeskorte N, Douglas PK (2018) Cognitive computational neuroscience. *Nat Neurosci* 21:1148–1160.
- Krishna BS, Semple MN (2000) Auditory temporal processing: responses to sinusoidally amplitude-modulated tones in the inferior colliculus. *J Neurophysiol* 84:255–273.
- Kuwada S, Batra R (1999) Coding of sound envelopes by inhibitory rebound in neurons of the superior olivary complex in the unanesthetized rabbit. *J Neurosci* 19:2273–2287.
- Langner G, Schreiner CE (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J Neurophysiol* 60:1799–1822.
- Lee K-F, Hon H-W (1989) Speaker-independent phone recognition using hidden Markov models. *IEEE Trans Acoust, Speech, Signal Processing* 37:1641–1648.
- Leibo JZ, Masson D’automne CDM, Zoran D, Amos D, Beattie C, Anderson K, Castañeda AG, Sanchez M, Green S, Gruslys A, Legg S, Hassabis D, Botvinick MM (2018) Psychlab: a psychology laboratory for deep reinforcement learning agents. [arXiv:1801.08116](https://doi.org/10.48550/arXiv.1801.08116). <https://doi.org/10.48550/arXiv.1801.08116>.
- Levitt H (1971) Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49:467–477.
- Lewicki MS (2002) Efficient coding of natural sounds. *Nat Neurosci* 5:356–363.
- Liang L, Lu T, Wang X (2002) Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *J Neurophysiol* 87:2237–2261.
- Lorenzi C, Soares C, Vonner T (2001a) Second-order temporal modulation transfer functions. *J Acoust Soc Am* 110:1030–1038.
- Lorenzi C, Simpson MI, Millman RE, Griffiths TD, Woods WP, Rees A, Green GGR (2001b) Second-order modulation detection thresholds for pure-tone and narrow-band noise carriers. *J Acoust Soc Am* 110:2470–2478.
- Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci U S A* 103:18866–18869.
- Lu T, Liang L, Wang X (2001) Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat Neurosci* 4:1131–1138.
- Lu T, Wang X (2000) Temporal discharge patterns evoked by rapid sequences of wide- and narrowband clicks in the primary auditory cortex of cat. *J Neurophysiol* 84:236–246.
- Luo X, Fu QJ, Wei CG, Cao KL (2008) Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users. *Ear Hear* 29:957–970.
- Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. [arXiv:1412.0035](https://doi.org/10.48550/arXiv.1412.0035). <https://doi.org/10.48550/arXiv.1412.0035>.
- Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. *Digit Signal Process* 73:1–15.
- Moore BCJ (2013) An introduction to the psychology of hearing. Brill, Leiden, Netherlands.
- Müller-Preuss P (1986) On the mechanisms of call coding through auditory neurons in the squirrel monkey. *Eur Arch Psychiatry Neurol Sci* 236:50–55.
- Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* 88:1281–1296.
- Piczak KJ (2015) ESC: dataset for environmental sound classification. Paper presented at the 23rd ACM International Conference on Multimedia, October, Brisbane, Australia.
- Preuss A, Müller-Preuss P (1990) Processing of amplitude modulated sounds in the medial geniculate body of squirrel monkeys. *Exp Brain Res* 79:207–211.
- Rhode WS, Greenberg S (1994) Encoding of amplitude modulation in the cochlear nucleus of the cat. *J Neurophysiol* 71:1797–1825.
- Richards FJ (1959) A flexible growth function for empirical use. *J Exp Bot* 10:290–301.
- RichardWebster B, Anthony SE, Scheirer WJ (2019) PsyPhy: a psychophysics driven evaluation framework for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 41: 2280–2286.
- Saddler MR, Gonzalez R, McDermott JH (2021) Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nat Commun* 12:1–25.
- Schreiner CE, Urbas JV (1988) Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hear Res* 32:49–63.
- Schulze H, Langner G (1997) Periodicity coding in the primary auditory cortex of the Mongolian gerbil (*Meriones unguiculatus*): two different coding strategies for pitch and rhythm? *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 181:651–663.
- Scott BH, Malone BJ, Semple MN (2011) Transformation of temporal processing across auditory cortex of awake macaques. *J Neurophysiol* 105:712–730.
- Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304.

- Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. *Curr Opin Neurobiol* 21:761–767.
- Smith EC, Lewicki MS (2006) Efficient auditory coding. *Nature* 439:978–982.
- Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416:87–90.
- Terashima H, Okada M (2012) The topographic unsupervised learning of natural sounds in the auditory cortex. *Advanc Neural Inf Process Sys* 2:2312–2320.
- Tokozume Y, Harada T (2017) Learning environmental sounds with end-to-end convolutional neural network. *IEEE International Conference on Acous, Speech, and Signal Processing*, 2017:2721–2725.
- van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) WaveNet: a generative model for raw audio. *arXiv:1609.03499*. <https://doi.org/10.48550/arXiv.1609.03499>.
- Van Grootel MWW, Andringa TC, Krijnders JD (2009) DARES-G1: Database of annotated real-world everyday sounds. In: *Proceedings of the NAG/DAGA International Conference on Acoustics*, pp 43.
- Verhulst S, Altoè A, Vasilkov V (2018) Computational modeling of the human auditory periphery: auditory-nerve responses, evoked potentials and hearing loss. *Hear Res* 360:55–75.
- Viemeister NF (1979) Temporal modulation transfer functions based upon modulation thresholds. *J Acoust Soc Am* 66:1364–1380.
- Won JH, Drennan WR, Nie K, Jameyson EM, Rubinstein JT (2011) Acoustic temporal modulation detection and speech perception in cochlear implant listeners. *J Acoust Soc Am* 130:376–388.
- Yin P, Johnson JS, O'Connor KN, Sutter ML (2011) Coding of amplitude modulation in primary auditory cortex. *J Neurophysiol* 105:582–600.
- Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. *arxiv:1506.06579*. <https://doi.org/10.48550/arXiv.1506.06579>.
- Zhang H, Kelly JB (2006) Responses of neurons in the rat's ventral nucleus of the lateral lemniscus to amplitude-modulated tones. *J Neurophysiol* 96:2905–2914.
- Zhao H-B, Liang Z-A (1995) Processing of modulation frequency in the dorsal cochlear nucleus of the guinea pig: amplitude modulated tones. *Hear Res* 82:244–256.
- Zhou D, Zhou X, Zhang W, Loy CC, Yi S, Zhang X, Ouyang W (2020) EcoNAS: finding proxies for economical neural architecture search. *arXiv:2001.01233*. <https://doi.org/10.48550/arXiv.2001.01233>.