Artificial Neural Network Trained for Sound Recognition Exhibiting Human-like Sensitivity to Sound Amplitude Modulation

Takuya Koumura

Abstract

Amplitude modulation (AM) is one of the most important sound physical dimensions for auditory perception. Human listeners can detect subtle AM in such a way that their sensitivity to AM depends on the stimulus parameters. Why does the auditory system exhibit such a form of AM sensitivity? How does the brain conduct AM detection? To answer these questions, my research colleagues and I conducted a computational study. We trained an artificial neural network (NN) model for sound recognition and simulated AM-detection experiments in the model. We found the emergence of human-like AM sensitivity in the model. The AM structure in the sounds for model training was essential for this emergence. The layers exhibiting human-like AM detection had AM-related properties similar to the neurons in the auditory midbrain and higher brain regions. These results suggest that AM sensitivity in humans might also be a result of the adaptation of the auditory system to sound recognition and that the auditory system might detect AM using neural activities in the auditory midbrain and higher brain regions.

Keywords: auditory perception, amplitude modulation, artificial neural network

1. Background

Sounds in our everyday life (e.g. speech, music, and environmental sounds) exhibit rich patterns of amplitude modulation (AM) (**Fig. 1**). AM is a slow change in sound amplitude. It is one of the most important sound features for auditory perception. Different patterns of AM evoke different hearing sensations such as pitch and roughness [1]. Humans can recognize a sound only with its AM cue [2, 3]. AM is often characterized by its rate (or speed) and depths (or magnitude) (Fig. 1, right panel). Originally, it referred to conveying a signal as a form of slowly changing amplitude, but in the context of auditory research, it is often used in the aforementioned sense.

Perceptual sensitivity to sound AM is considered an important property reflecting auditory perception because it quantifies the ability of the auditory system to detect subtle AM cues in the sound stimulus. It has been investigated under multiple experimental conditions in several independent studies [4–7] and has been shown to depend on stimulus parameters such as the AM rate, carrier bandwidth, and sound duration. In the previous experiments, a modulated and non-modulated sound are presented in succession to a human listener. When the listener is asked to identify which sound is modulated, the answer is generally correct when the modulated stimulus has a deep AM, but the discrimination is more difficult when the AM is shallower. Therefore, we can define an AM-detection threshold as the minimum depth required for discriminating a modulated sound from a non-modulated one.

Understanding this stimulus-parameter dependency

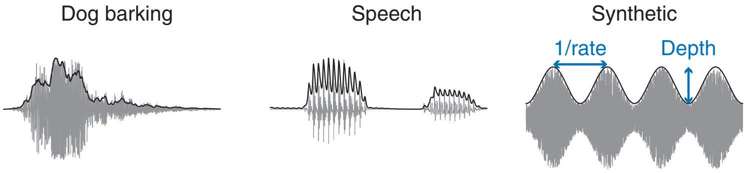

Fig. 1. Examples of sound AM, shown as black lines, in dog barking, speech, and synthetic sounds.

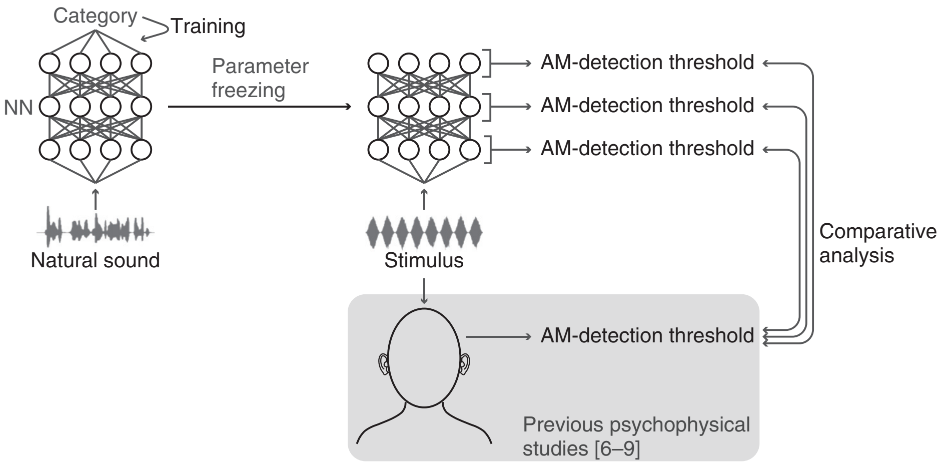

Fig. 2. Overview of the procedure of model construction and analysis.

is critical in the research of auditory perception because those parameters vary greatly in the sounds in our everyday environment and affect our sensitivity to AM cues, which in turn influences our perception. Therefore, it is important to consider the following fundamental questions: why such a stimulus-parameter-dependent sensitivity has emerged in the auditory system, and how it is actualized in our brain. The "why" question is scientifically important because it incorporates an understanding of the evolutionary and developmental process (or the "origin") of the auditory system. However, it requires considering the long time scale of our evolution and development, which makes it difficult (albeit not impossible) to address experimentally. In such a case, computational modeling can be effective. If we assume that an increase in evolutionary fitness is the primary factor in shaping the auditory system and that better sound-recognition performance yields better evolutionary fitness, we can computationally simulate this adaptation process using machine-learning techniques. After constructing a model by training it for a biologically relevant task such as sound recognition, we can simulate psychophysical or neurophysiological experiments and compare the emergent properties with those in the auditory system to gain insights into the effect of the adaptation to sound recognition on shaping the auditory properties. Several studies have been conducted along this line including ours that demonstrated the emergence of neuronal AM-related properties in an artificial neural network (NN) trained for sound recognition [8]. The "how" question has often been addressed by neuroscience, but this computational paradigm should help us understand those experimentally elucidated mechanisms from the perspectives of their emergence.

2. Simulating psychophysical experiments in a neural network trained for sound recognition

In our present study, we applied this paradigm to the AM-detection threshold [9]. This study involved the following two steps: constructing a computational model of the auditory system and simulating psychophysical experiments in the model (**Fig. 2**). We used an artificial NN as the computational model. The

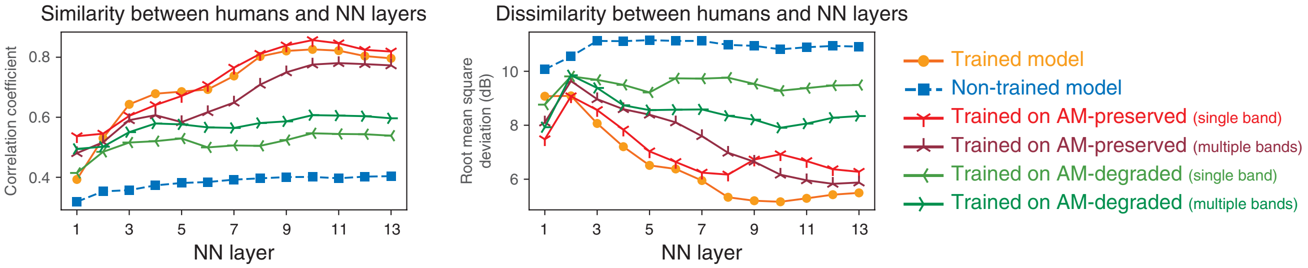

Fig. 3. Similarity and dissimilarity between the AM-detection threshold in humans and that in NN layers.

model architecture was almost the same as in our previous study [8]. We used a multi-layer (or deep) convolutional NN that takes a sound waveform as an input and outputs the estimated category of the input. The model parameters were adjusted to match the estimated categories to the true categories (Fig. 2, left). This process is called “training” in machine learning. The training objective was the classification of everyday sounds or of phonemes in speech sounds.

After the training, we froze the model parameters and simulated psychophysical experiments (Fig. 2, right). We delivered a modulated or non-modulated stimulus to the model and measured how accurately the model discriminated them. Specifically, we attempted to estimate whether the stimulus was modulated from the time-averaged model activity in response to the stimulus. The stimulus parameters were the same as in psychophysical studies [4–7]. This enabled the direct and quantitative comparison of our results with those in humans. The AM-detection threshold was defined as the depth at which the discrimination performance was 70.7% (again, same as in the psychophysical studies). This threshold was measured in each layer in the NN. We conducted the same simulation in the non-trained model (i.e. the model with random parameter values before training for sound recognition).

3. Emergent AM detection threshold in the model

We first quantified the similarity of the AM-detection threshold in the model and that in humans (**Fig. 3**, orange circles for the trained model, blue squares for the non-trained model). This comparison was done for each NN layer. We compared the stimulus-parameter-dependent AM-detection threshold in an NN layer and humans in terms of their overall patterns of stimulus-parameter dependency and their specific values. To compare their overall pattern, we quantified the relative similarity using the correlation coefficient (Fig. 3, left panel). To compare their specific values, we quantified the absolute dissimilarity using the root mean square deviation (Fig. 3, right panel). The upper layers in the trained model showed high similarity and low dissimilarity with humans, whereas the lower layers and the non-trained model showed low similarity and high dissimilarity. This indicates that the human-like detection threshold emerged in the upper layers as a result of training for sound recognition.

We then wanted to determine the essential factors for the emergence of the human-like detection threshold in the model. Since the similarity to humans was largely different between trained and non-trained models, we hypothesized that the training procedure is an important factor. To test this, we trained the model on degraded sounds and compared the emergent detection threshold with humans. We tested two types of degraded sounds: sounds with degraded AM components and those with degraded faster components (i.e. faster change in their amplitude than AM components). In the latter sounds, AM components were preserved. The model trained on AM-degraded sounds did not exhibit a human-like detection threshold (Fig. 3, green < and > shapes), whereas the model trained on AM-preserved sounds showed a detection threshold somewhat similar to humans (Fig. 3, red Y and inverse Y shapes). The results indicate that the AM structure in the training data is essential for the emergence of a human-like AM detection threshold.

Finally, to gain insights into how AM detection is conducted in the brain, we estimated the brain regions responsible for AM detection. Figure 3 indicates that layers around the 9th, 10th, and 11th layers exhibit the human-like AM detection threshold. We estimated the corresponding brain regions to these layers by using a method developed in our previous study for

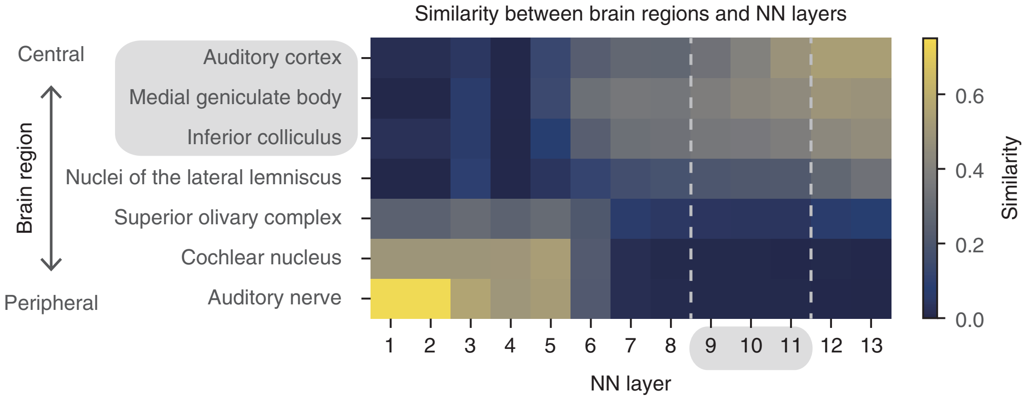

Fig. 4. The similarity of neuronal AM-related properties in the NN layers and those in the auditory brain regions.

calculating the similarity between an NN layer and brain region in terms of neuronal AM-related properties [8]. Applying this method to the present model revealed that these layers were similar to the inferior colliculus, medial geniculate body, and auditory cortex (**Fig. 4**), which are included in the auditory midbrain and higher regions. This means the 9th, 10th, and 11th layers would respond to AM stimuli similarly to neurons in these brain regions. This suggests that, if we calculate the AM-detection threshold as in our analysis from the time-averaged neural activities in these brain regions, we would obtain an AM-detection threshold similar to that observed in the psychophysical experiments. A human brain might also use time-averaged neural activities in these brain regions when a human listener is performing AM detection.

4. Conclusions

We simulated psychophysical AM-detection experiments in an NN model trained for sound recognition. We observed the emergence of the human-like AM detection threshold in the upper layers in the trained model, suggesting that the detection threshold in humans might also be a result of the adaptation of the auditory system to sound recognition during evolution and/or development. This provides an answer to the question of why the auditory system exhibits the present form of the AM detection threshold. We demonstrated that the AM structure in the training data is an essential factor for the model to exhibit the human-like detection threshold. Mapping of the NN layer and brain regions suggests that psychophysical AM detection might be a result of neural activities in the auditory midbrain and higher brain regions. This provides an answer to the question of how AM detection is performed in the auditory system.

References

[1] P. X. Joris, C. E. Schreiner, and A. Rees, "Neural Processing of Amplitude-modulated Sounds," Physiol. Rev., Vol. 84, No. 2, pp. 541–577, Apr. 2004. https://doi.org/10.1152/physrev.00029.2003

[2] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech Recognition with Primarily Temporal Cues," Science, Vol. 270, No. 5234, pp. 303–304, Oct. 1995. https://doi.org/10.1126/science.270.5234.303

[3] B. Gygi, G. R. Kidd, and C. S. Watson, "Spectral-temporal Factors in the Identification of Environmental Sounds," J. Acoust. Soc. Am., Vol. 115, No. 3, pp. 1252–1265, Mar. 2004. https://doi.org/10.1121/1.1635840

[4] N. F. Viemeister, "Temporal Modulation Transfer Functions Based upon Modulation Thresholds," J. Acoust. Soc. Am., Vol. 66, No. 5, pp. 1364–1380, Nov. 1979. https://doi.org/10.1121/1.383531

[5] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling Auditory Processing of Amplitude Modulation. I. Detection and Masking with Narrow-band Carriers," J. Acoust. Soc. Am., Vol. 102, No. 5, pp. 2892–2905, Nov. 1997. https://doi.org/10.1121/1.420344

[6] C. Lorenzi, C. Soares, and T. Vonner, "Second-order Temporal Modulation Transfer Functions," J. Acoust. Soc. Am., Vol. 110, No. 2, pp. 1030–1038, Aug. 2001. https://doi.org/10.1121/1.1383295

[7] C. Lorenzi, M. I. G. Simpson, R. E. Millman, T. D. Griffiths, W. P. Woods, A. Rees, and G. G. R. Green, "Second-order Modulation Detection Thresholds for Pure-tone and Narrow-band Noise Carriers," J. Acoust. Soc. Am., Vol. 110, No. 5, pp. 2470–2478, Nov. 2001. https://doi.org/10.1121/1.1406160

[8] T. Koumura, H. Terashima, and S. Furukawa, "Cascaded Tuning to Amplitude Modulation for Natural Sound Recognition," J. Neurosci., Vol. 39, No. 28, pp. 5517–5533, July 2019. https://doi.org/10.1523/JNEUROSCI.2914-18.2019

[9] T. Koumura, H. Terashima, and S. Furukawa, "Human-like Modulation Sensitivity Emerging through Optimization to Natural Sound Recognition," J. Neurosci., Vol. 43, No. 21, pp. 3876–3894, May 2023. https://doi.org/10.1523/JNEUROSCI.2002-22.2023

Takuya Koumura

Researcher, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a B.E. in science in 2011, Master in arts and sciences in 2013, and Ph.D. in arts and sciences in 2016 from the University of Tokyo. From 2015 to 2016, he received a research fellowship for young scientists (DC2) from Japan Society for the Promotion of Science. In 2016, he joined NTT Communication Science Laboratories as a research associate and began studying auditory perception. He is a member of Acoustical Society of Japan and Japanese Neural Network Society.