“Psychophysical” modulation transfer functions in a deep neural network trained for natural sound recognition

TAKUYA KOUMURA[1,*], HIROKI TERASHIMA[1], AND SHIGETO FURUKAWA[1]

[1] *NTT Communication Science Laboratories 3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198 Japan*

Representation of amplitude modulation (AM) has been characterized by neurophysiological and psychophysical modulation transfer functions (MTFs). Our recent computational study demonstrated that a deep neural network (DNN) trained for natural sound recognition serves as a good model for explaining the functional significance of neuronal MTFs derived physiologically. The present study addresses the question of whether the DNN can provide insights into AM-related human behaviours such as AM detectability. Specifically, we measured “psychophysical” MTFs in our previously developed DNN model. We presented to the DNN sinusoidally amplitude-modulated white noise with various AM rates, and quantified AM detectability as d' derived from the model’s internal representations of modulated and non-modulated stimuli. The overall d' increased along the layer cascade, with human-level detectability observed in the higher layers. In a given layer, the d' tended to decrease with increasing AM rates and with decreasing AM depth, which is reminiscent of a psychophysical MTF. The results suggest that a DNN trained for natural sound recognition can serve as a model for understanding psychophysical AM detectability. Since our approach is not specific to AM, the present paradigm opens the possibility of exploring a broad range of auditory functions that can be evaluated by psychophysical experiments.

BACKGROUND

Amplitude modulation (AM) is an important physical dimension for natural sound recognition. For example, humans can recognize speech and other natural sounds with a deteriorated temporal fine structure if the amplitude envelopes of the sounds are preserved (Shannon *et al.*, 1995; Gygi *et al.*, 2004).

Numerous neurophysiological studies have sought to reveal how the auditory system represents AM. They have found that the spike synchrony to the stimulus AM and the average spike rate in neurons in the auditory system exhibit tuning to the AM rate. Tuning to the AM rate is often characterized by a modulation transfer function (MTF), which is defined as the spike synchrony or average spike rate as a function of the AM rate. Interestingly, peak AM rates and the upper cutoff frequencies of the MTFs

*Corresponding author: koumura@cycentum.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

systematically transform along the cascade of the brain regions in the auditory system (Joris *et al.*, 2004).

In our previous study, we asked an alternative question: why does the auditory system represent AM in such ways (Koumura *et al.*, 2019)? To explore the functional significance of the systematically transforming MTFs, we built a computational model that can perform a behaviourally meaningful task, namely natural sound recognition. Specifically, we trained a deep neural network (DNN) for the task and analysed the AM representation in it. A DNN is suitable for modelling the auditory system in two ways. First, it can perform natural sound recognition, which is one of the most important functions of the auditory system. Functions such as vocal communication and sound localization are of similar importance, but in this study we only focused on natural sound recognition. Second, it consists of a cascade of layers, which is similar to the cascade of brain regions in the auditory system (see Fig. 30-12 in Kandel *et al*. (2000)).

To directly compare the AM representation in the DNN with that revealed by neurophysiological studies, we performed single-unit recording in the trained DNN. We found that similar transformation of MTFs along the layer cascade in the DNN emerged as a result of the training for natural sound recognition. The similarity gradually increased in the course of the training. The results suggest that AM tuning in the auditory system might also be a result of optimization for natural sound recognition in animals in the course of evolution and development.

While neurological studies have investigated the neural representation of AM, psychophysical studies have sought to characterize behavioural responses to it. They have found the dependency of sensitivity to AM on AM rates, which is characterized by a psychophysical MTF defined as the AM detection threshold as a function of the AM rate. For example, when broadband noise is used as the stimulus carrier, an MTF takes the form of a low-pass filter, with the detection threshold decreasing about 3 dB per octave (Viemeister, 1979).

The present study addresses the question of whether such psychophysical properties also emerge in the DNN trained for natural sound recognition and to what extent they are similar to those observed in humans. We conducted a psychophysical AM detection experiment in our DNN (Fig. 1). To characterize AM sensitivities in the DNN, we calculated sensitivity index d' based on the representation of the stimuli in each layer and each unit, and defined the detection threshold as the minimum AM depth required to yield a certain value of d'. The MTFs in the middle layers were similar to those in humans, whereas an untrained DNN was not as sensitive to AM as humans are. The results suggest that not only neurophysiological MTFs but also psychophysical MTFs can be compared between the auditory system and DNNs to better understand why the MTFs have specific forms.

METHODS

Training of the DNN

As a model, we used the DNN we built in our previous study (Koumura *et al.*, 2019). Here we briefly explain the model and the training procedure. The DNN consists of 13 temporally dilated convolutional layers, and each layer consists of 128 units. All layers operate with 44.1 kHz sampling frequency.

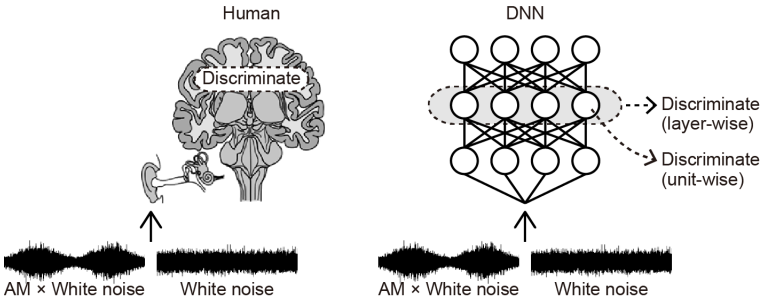

Fig. 1: Our approach. In an AM detection experiment, humans try to discriminate modulated stimuli from unmodulated stimuli (**left**). We simulated this procedure by trying to discriminate modulated and unmodulated stimuli from DNN's representation at the level of single layers or single units (**right**).

The DNN was trained for natural sound recognition. The input was a 0.19 s segment of natural sound, and the DNN's task was to estimate the category of the input sound. The sound data was a subset of ESC-50 (Piczak, 2015). The original dataset is divided into 5 folds. We used folds 1–4 for training and fold 5 for validation. The classification accuracy for the validation set was 45.1%. Due to space limitations, the details of the hyper-parameters and training procedures are not fully described here. They are provided in our previous paper (Koumura *et al.*, 2019).

Stimulus for AM detection experiment

As a stimulus, we used modulated and unmodulated white noise. The duration of the stimulus was 0.5 s, and raised cosine ramps of 50 ms were applied. The starting phase of AM was fixed at 0. The duration of the stimulus and ramps and the starting phase were the same as in Viemeister (1979). The overall amplitude was scaled so that the root mean square (RMS) of the stimulus was equal to the average root mean square (RMS) of the training data. The amplitude was scaled before modulation was applied as in Viemeister (1979). All carrier white noises were independently sampled trial by trial.

RESULTS

Representation of modulated and unmodulated sounds in a single layer

First, we visualized the representation of modulated (AM rate = 32 Hz; depth = −10 dB) and unmodulated white noise in the 7th layer as an example. Response time

courses in the 128 units were recorded in a single layer (Fig. 2). Responses to 32 modulated and 32 unmodulated noises in all units were concatenated, and their dimension was reduced to 4 by principle component analysis (PCA). We confirmed that the results obtained with 2- or 8-dimensional PCA were similar to those with 4-dimensional PCA. Visualizing the first two principle components indicates that in the 7th layer, AM with 32 Hz and −10 dB depth was well discriminated from unmodulated noises (Fig. 3).

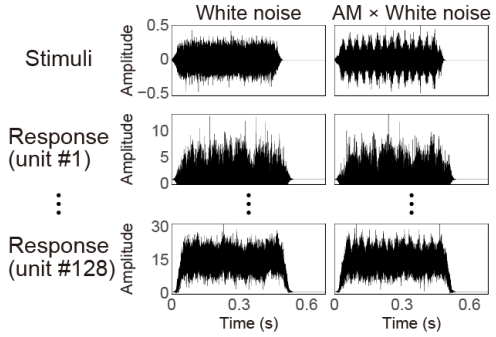

Fig. 2: Examples of stimuli and the DNN's responses. One sample of the noise stimulus is shown for each of the modulated and unmodulated stimuli. Responses in unit #1 and #128 in the 7th layer are shown.

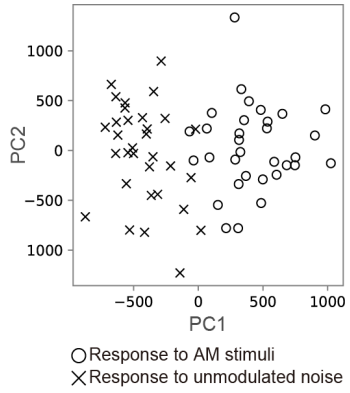

Fig. 3: Two-dimensional visualization of the AM representation in the 7th layer. Responses to 32-Hz modulated and unmodulated noises are shown. Since our model is deterministic, the only cause of the response variability is the variability of input noises. The horizontal and the vertical axes show first- and second-principle components, respectively.

Sensitivity index

As a measure of AM sensitivity, we calculated d' for AM detection based on the representations in each layer. The 4-dimensional representations of the 32 modulated and 32 unmodulated white noises were further projected onto a 1-dimensional axis with maximum detectability in terms of a linear discriminant analysis. The d' was calculated from the means and variances of the 1-dimensional representation (Averbeck and Lee, 2006).

Fig. 4 shows d' in the 7th layer for 32 Hz AM. The d' appeared constant and low with shallow AM, and at a certain AM depth it started to increase linearly on a logarithmic scale. This trend—sensitivity increasing with AM depth—is reasonable when considering the stimulus charactersitics. In theory, the shallower the AM, the more difficult it will be to detect it. Having observed this trend, we fitted a broken line with two segments to the d' on a logarithmic scale. One of the segments for the lower depth was assumed to be constant. The mean squared error of the fitted lines and measured logarithmic d' was 0.024 ± 0.013 (mean ± standard deviation over all AM rates and

all layers). From the fitted lines, we defined the detection threshold as the AM depth at $d' = 1.089$, which corresponds to 70.7% correct, assuming the responses follow a normal distribution. As in the standard psychophysical studies, an MTF was defined as the AM detection threshold as a function of AM rate.

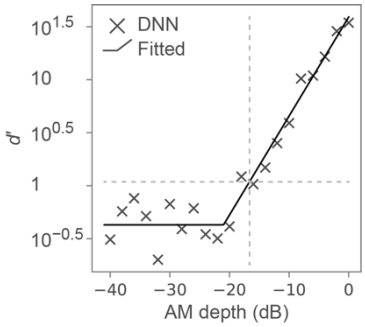

Fig. 4: The d' in the 7th layer for 32-Hz AM rate (crosses) and fitted lines. Horizontal and vertical dashed lines indicate $d' = 1.089$ and the detection threshold, respectively.

The d' was calculated for multiple AM rates and depths in each layer (Fig. 5, upper panels). The calculated values were large for deep and slow AM, and they were also large in the higher layers. The detection threshold in the DNN and humans are compared in the lower panels of Fig. 5. MTFs in the middle layer took the form of a low-pass filter with constant d' up to around 8 Hz and the slope of approximately 3 dB/octave, which is similar to those in humans, although there seems to be a constant discrepancy.

MTFs in a single unit

The above analysis was a comparison between human MTFs and an MTF in a single layer, calculated from the concatenated response timecourses in all units. Next, we calculated an MTF in each unit. MTFs varied among units (Fig. 6, upper panels). Interestingly, in the middle layer, their envelope aligned with the human MTFs. This is more clearly seen by connecting the most sensitive MTFs (Fig. 6, lower panels). In the middle layer, the envelope of the unit MTFs was very similar to that of human MTFs without a constant discrepancy.

AM sensitivity in the untrained DNN

The observed AM sensitivity could be a consequence of the training for natural sound recognition or could be explained by the architecture of the model with cascaded convolutions. To test these possibilities, we calculated the d' in an untrained DNN as a control experiment. The connection weights in the untrained DNN were randomly sampled from the normal distribution, and its activity bias was 0 (He *et al.*, 2015). The d' in the untrained DNN was much lower than those in the trained DNN, indicating that representations in the untrained DNN were not sensitive to the stimulus AM (Fig. 7). The results suggest that parameter optimization is necessary for AM sensitivity. It is worth noting again that our DNN is optimized for natural sound recognition, not for AM sensitivities in humans.

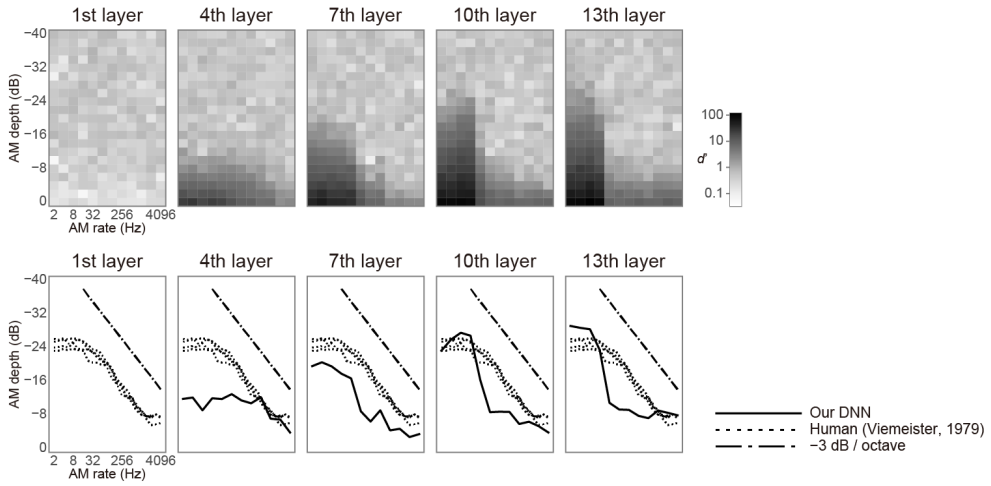

Fig. 5: Sensitivity index and MTF in each layer. The *d′* is colour-coded in the upper panels. The lower panels show MTFs in our DNN (solid lines) and in humans (dotted lines, Viemeister (1979)). We also plotted lines indicating – 3dB/octave (dashed-dotted lines) as in Viemeister (1979).

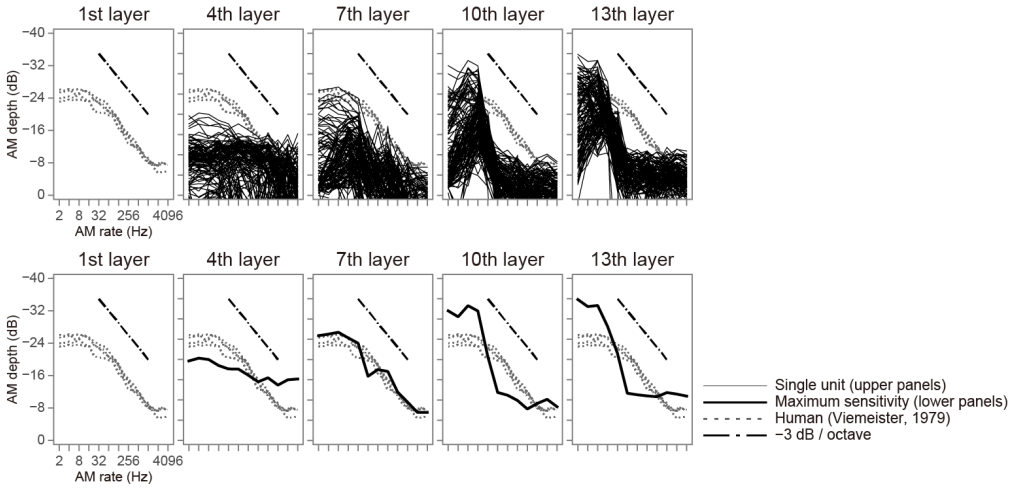

Fig. 6: MTFs in single units (upper) and their envelopes connecting the most sensitive MTFs for each AM rate (lower). Dotted lines and dashed-dotted lines are the same as Fig. 5.

DISCUSSION

We analysed AM sensitivity in the DNN trained for natural sound recognition and found that MTFs similar to those in humans emerged in the middle layers. The untrained DNN did not exhibit high sensitivity. These results, together with the neurophysiological analysis in our previous study, suggest that AM sensitivity in humans might be a result of optimization for natural sound recognition.

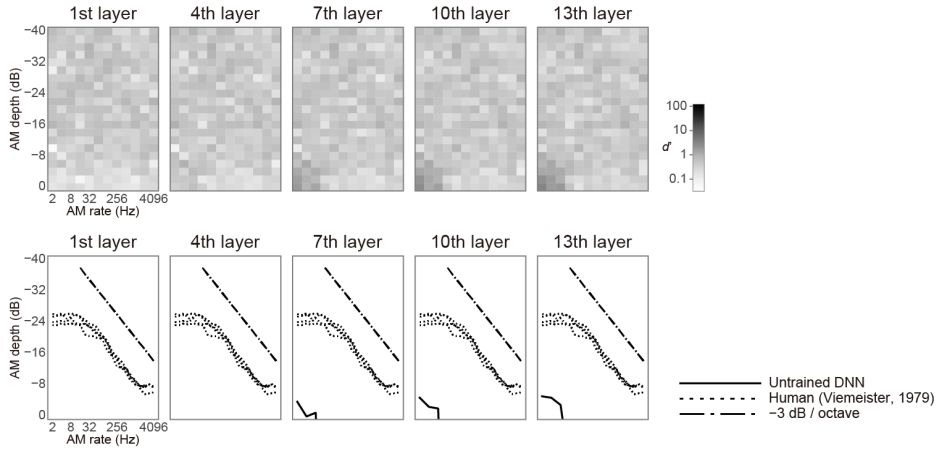

Fig. 7: MTFs in the untrained DNN. Other conventions, including the color scale of *d'*, are the same as in Fig. 5.

In our previous study, DNNs trained with speech signals exhibited neurophysiological MTFs similar to those trained with natural sounds. Therefore, we expect that they would also show psychophysical MTFs similar to those obtained in this study.

Comparison of MTFs in the DNN and in humans

Although the form of the MTF in the middle layer was similar to those in humans, there was a constant discrepancy of around 4 dB, indicating that humans are a little more sensitive to our layer representation. On the other hand, single units can be as sensitive as humans if units with maximum sensitivity are recruited for each AM rate. Thus, it may be possible to model human MTFs by combining MTFs with the most sensitive units. Our previous neurophysiological simulation suggested that unit activities in the middle layers of the DNN may be a model of neural activities in the brainstem. When taken together with the present results, it appears that humans might integrate outputs in the most sensitive neurons in the brainstem to yield responses in an AM detection task.

Future work

The present study only tested an AM stimulus with broad band noise carriers. Psychophysical MTFs have been measured using various carriers, such as those with narrowband noise (Dau *et al.*, 1997), and it has been shown that an MTF depends on the type of carrier. In addition, other types of modulation, such as second-order modulation, has been tested in humans (Lorenzi *et al.*, 2001). Testing other types of stimuli in our model remains as future work.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP15H05915 (Grant-in-Aid for Scientific Research on Innovative Areas "Innovative SHITSUKSAN Science and Technology").

REFERENCES

Averbeck, B.B., Lee, D. (**2006**). "Effects of Noise Correlations on Information Encoding and Decoding," J. Neurophysiol., **95**, 3633–3644. doi:10.1152/jn.00919.2005.

Dau, T., Kollmeier, B., Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am., **102**(5), 2892-2905. doi:10.1121/1.420344.

Gygi, B., Kidd, G.R., Watson, C.S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am., **115**, 1252–1265. doi:10.1121/1.1635840.

He, K., Zhang, X., Ren, S., Sun, J. (**2015**). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," Proceedings IEEE International Conference on Computer Vision (ICCV), 1026–1034 . doi:10.1109/ICCV.2015.123.

Joris, P.X., Schreiner, C.E., Rees, A. (**2004**). "Neural processing of amplitude-modulated sounds," Physiol. Rev., **84**, 541–577. doi:10.1152/physrev.00029.2003.

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (**2000**). "Principles of Neural Science," Fourth Edition. New York, NY: McGraw-Hill.

Koumura, T., Terashima, H., Furukawa, S. (**2019**). "Cascaded Tuning to Amplitude Modulation for Natural Sound Recognition," J. Neurosci., **39**, 5517–5533. doi:10.1523/JNEUROSCI.2914-18.2019.

Lorenzi, C., Soares, C., Vonner, T. (**2001**). "Second-order temporal modulation transfer functions," J. Acoust. Soc. Am., **110**, 1030–1038. doi:10.1121/1.1383295

Piczak, K.J. (**2015**). "ESC : Dataset for Environmental Sound Classification," In 23rd ACM International Conference on Multimedia, 1015–1018.

Shannon, R.V., Zeng, F-G., Kamath, V., Wygonski, J., Ekelid, M. (**1995**). "Speech Recognition with Primarily Temporal Cues," Science, **270**, 303–304. doi:10.1126/science.270.5234.303.

Viemeister, N.F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am., **66**, 1364–1380. doi:10.1121/1.383531.