Information Maximization in a Feedforward Network Replicates the Stimulus Preference of the Medial Geniculate and the Auditory Cortex

Takuma Tanaka(✉)

The Center for Data Science Education and Research,
Shiga University, 1-1-1 Banba, Hikone, Shiga 522-8522, Japan
tanaka.takuma@gmail.com

Abstract. Central auditory neurons exhibit a preference for complex features, such as frequency modulation and pitch. This study shows that the stimulus preference for these features can be replicated by a network model trained to maximize information transmission from input to output. The network contains three layers: input, first-output, and second-output. The first-output-layer neurons exhibit auditory-nerve neuron-like preferences, and the second-output-layer neurons exhibit a stimulus preference similar to that of cochlear nucleus, medial geniculate, and auditory cortical neurons. The features detected by the second-output-layer neurons reflect the statistical properties of the sounds used as input.

Keywords: Information maximization · Auditory information processing · Auditory cortex · Pitch selectivity · Frequency modulation selectivity

1 Introduction

Neurons in the auditory system of the brain can detect multiple dimensions of the complex features of sounds. For example, human speech is composed of a series of combinations of features such as phoneme, tone, stress, length, and prosody. The precise timing and temporal variation of these features convey critical information about the content of speech. To correctly comprehend human speech, an auditory system must be able to detect the features precisely and encode these features in the firing patterns of neurons.

The neural encoding of sensory inputs has been intensively studied both experimentally and theoretically. Experiments on auditory information processing have revealed that auditory-nerve neurons respond to sine wave-like tones [1] and that central auditory neurons encode complex features such as sound intensity and pitch. More than half of medial geniculate neurons in cats exhibit non-monotonic rate–sound-intensity functions [2], which implies that these neurons can be interpreted as intensity-coding neurons. Bendor and Wang reported

A. Hirose et al. (Eds.): ICONIP 2016, Part IV, LNCS 9950, pp. 183–190, 2016.
DOI: 10.1007/978-3-319-46681-1_22

that neurons in the auditory cortex of marmoset monkeys responded to both pure tones and missing fundamental harmonic complex sounds and called them pitch-selective neurons [3]. This type of high-level feature selectivity is thought to have emerged as a result of the integration of simpler low-level feature selectivity, such as that exhibited by auditory-nerve neurons, in a circuit with hierarchical structure. Hierarchical structure has been widely used in the theoretical modeling of visual information processing [4–7]. Theoretical studies of visual processing have demonstrated that stimulus selectivity to complex features such as the boundary between two gratings emerges in a generative model of natural scenes [8,9] and in network models that maximize the amount of information conveyed by the output [7,10]. These studies suggest that feature representation in the auditory system can be understood on the basis of a similar framework. In fact, a previous model of auditory feature representation [11] showed that units representing complex auditory features emerged in a generative model of sounds. However, the "spikes" in this model were generated by the maximum likelihood estimation, and therefore whether neurons can perform such a computation remains unclear. Moreover, complex auditory feature detection has not been treated in terms of maximizing information transmission.

Therefore, this study examines auditory feature detection in a feedforward network of rate-coding neuron models using an algorithm based on the information maximization principle. The network consists of three layers: input, first-output, and second-output. Short waveforms from a natural sounds dataset and a human speech corpus are provided to the input layer. The first-output-layer neurons respond to the wavelet-like waveforms. The second-output-layer neurons encode more complex features, such as pitch, tone intensity, and upward and downward frequency modulation. The selectivity of these model neurons for these complex features is comparable to that of experimentally reported cochlear nucleus, medial geniculate, and auditory cortical neurons. These results suggest that the central auditory neurons can be understood in terms of information maximization and that an extended network model based on the information maximization principle could replicate the more complex feature detection of the auditory cortices.

2 Model

The Pittsburgh natural sounds dataset [12] was used as the natural-sound input, and the Priority Areas "Spoken Dialogue" Simulated Spoken Dialogue Corpus (PASD) was used as the human-speech input. The former was down-sampled to 11 kHz, and the latter was down-sampled to 8 kHz. The input time series was shifted and scaled to have zero mean and unit variance. Consecutive samples ($N = 200$) were randomly chosen from the input time series to be used as input to the network at each time step. The network model and learning algorithm described previously [7] were used. The network consists of input, first-output, and second-output layers, each containing N model neurons. The value of the i-th input at time step t is $x_i(t)$. The states of the neurons in the first- and second-output layers at time t are determined by

$$u_i(t) = f\left(\sum_{j=1}^{N} V_{ij} x_j(t)\right), \tag{1}$$

and

$$z_i(t) = f\left(\sum_{j=1}^{N} W_{ij}(|u_j(t)| - \overline{|u_j|})\right), \tag{2}$$

respectively, where V_{ij} is the connection weight from the input neuron j to the first-output-layer neuron i, W_{ij} is the connection weight from the first-output-layer neuron j to the second-output-layer neuron i, $\overline{|u_j|}$ is the average of $|u_j(t)|$ over time, and

$$f(x) = 2 \arctan \tanh \frac{x}{2} \tag{3}$$

is the activation function. The integration time constant τ was set to 10^4 steps in all simulations. The weights V_{ij} were updated once every 1000 steps. The updates of the matrices $\mathbf{V}$ and $\mathbf{W}$ were performed 10^4 times with $\epsilon = \epsilon_0$, 1.9×10^5 times with $\epsilon = 10\epsilon_0$, and 8×10^5 times with $\epsilon = \epsilon_0$ using the Newton method described in [7], where $\epsilon_0 = 10^{-6}$ for natural-sound input and $\epsilon_0 = 10^{-5}$ for human-speech input.

3 Results

3.1 First-Output-Layer Neurons

The input and first-output layer can be regarded as a network performing independent component analysis [6]. Similar to a previous study of independent component analysis [13], the first-output-layer neurons exhibit selectivity to Gabor wavelet-like sound waveforms. Figure 1A shows the column vectors of $\mathbf{V}^{-1}$, sound waveforms that the first-output-layer neurons are selective for, of the network in which natural sounds are used as the input. They have Gabor function-like shapes with different frequencies, amplitudes, phases, dispersions, and center

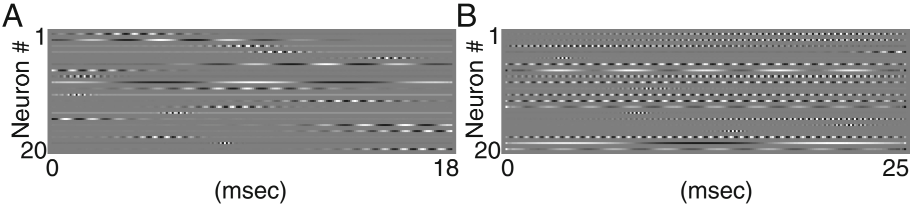

Fig. 1. Preferred sound waveform of 20 first-output-layer neurons in networks with (A) natural-sound input and (B) human-speech input.

positions. Figure 1B shows the selectivity of the first-output-layer neurons with human-speech input. The waves in Fig. 1B resemble sine waves without amplitude modulation because the majority of human speech is composed of vowels. Selectivity to these sine waves with and without amplitude modulation is similar to the stimulus preference of auditory-nerve neurons [1,13].

The response of these neurons to continuous sine waves is shown in Fig. 2. The horizontal and vertical axes represent the frequency and amplitude of sine waves, respectively. The density represents the maximal value of the first-output-layer neuron output obtained by varying the phases of the sine waves. Almost all neurons respond to a small-amplitude tone with the preferred frequency and to a wide range of tones if the amplitude is increased. This figure shows that the wavelet-like connection weights result in unimodal, V-shaped tuning curves with only one preferred frequency [14].

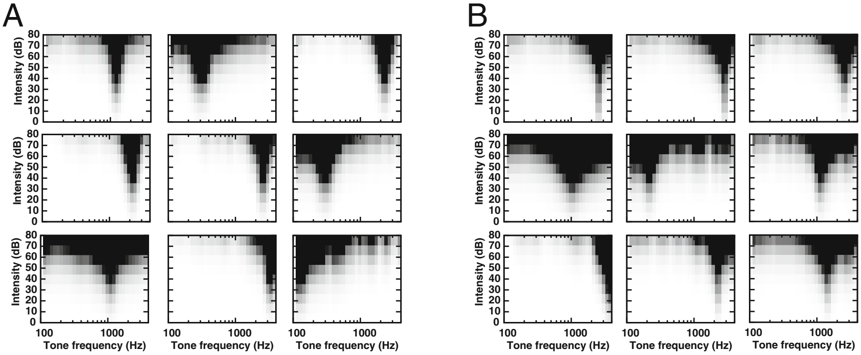

Fig. 2. Tuning curve of first-output-layer neurons in networks with (A) natural-sound input and (B) human-speech input.

3.2 Second-Output-Layer Neurons

Figure 3 shows the connection weights from the first-output-layer neurons to the second-output-layer neurons in the networks with natural-sound and human-speech inputs. Each box corresponds to a second-output-layer neuron, and each line in the box corresponds to the connection weight from a first-output-layer neuron to the second-output-layer neuron. The vertical position of a line in the box represents the preferred frequency of the first-output-layer neuron obtained by fitting the row vector of $\mathbf{V}^{-1}$ with the Gabor function, that is, ω of

$$g(t) = \exp[-(t - t_0)^2/(2\pi\sigma^2)] \sin(\omega t + \phi). \qquad (4)$$

The horizontal position and the length of the line represent t_0 and σ, respectively. The color represents the value of the connection weight W_{ij}, with red and blue

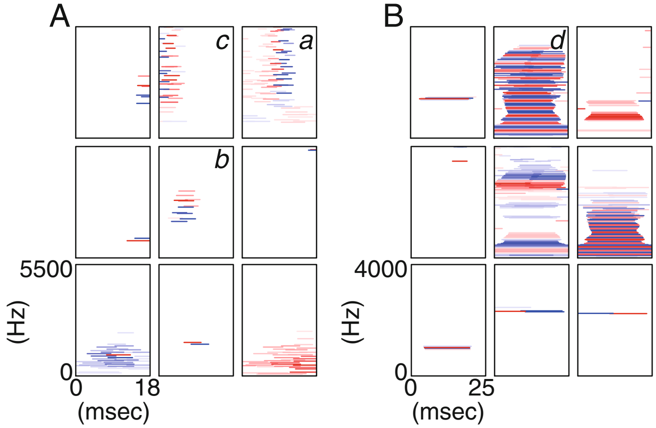

Fig. 3. Connection weights from first- to second-output-layer neurons for the networks with (A) natural-sound input and (B) human-speech input. The second-output-layer neurons are represented as boxes. Connections from the first-output-layer neurons correspond to the horizontal lines, whose vertical position, horizontal position, and length represent ω, t_0, and σ of the fitted Gabor function, respectively. Red and blue indicate positive and negative connection weights, respectively. (Color figure online)

indicating positive (excitatory) and negative (inhibitory) connection weights, respectively.

The stimulus selectivities exhibited by the first-output-layer neurons in the networks with natural-sound and human-speech inputs are substantially different from each other. Because the difference in the stimulus preference of the first-output-layer neurons affects the stimulus preference of the second-output-layer neurons, Figs. 3A and B show completely different types of stimulus preferences in these two networks. Therefore, the results of the two networks with different inputs are presented separately.

Natural Sounds

Frequency-Modulation Selectivity. Second-output-layer neuron *a* in Fig. 3A receives positive connection weights from first-output-layer neurons selective for low-frequency tones in the earlier half and for high-frequency tones in the latter half, while it receives negative connection weights from neurons selective for high-frequency tones in the earlier half and for low-frequency tones in the latter half. Therefore, this neuron is selective for a frequency change from a low tone to a high tone. Selectivity to frequency modulation is found in more than half of the neurons in the cochlear nucleus [15]. Neuron *a* appears to correspond to these cochlear nucleus neurons responsive to frequency modulation.

Intensity Tuning. Figure 4A shows the tuning curves of the second-output-layer-neurons trained with natural-sound input. These tuning curves are much more variable than those of the first-output-layer neurons (Fig. 2A). Tuning curve *b*

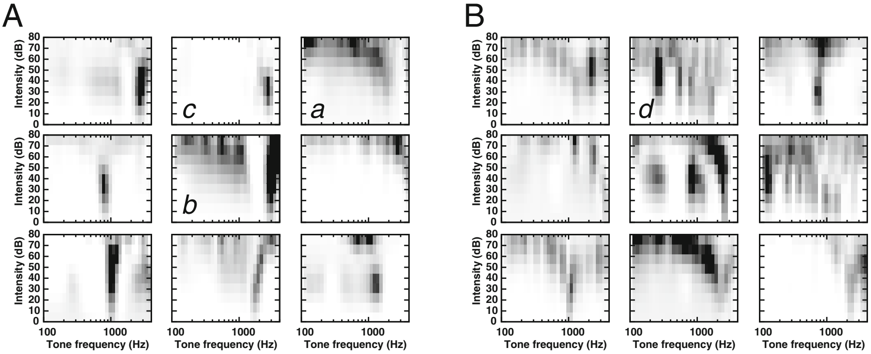

Fig. 4. Tuning curves of second-output-layer neurons with (A) natural-sound input and (B) human-speech input.

in Fig. 4A has two preferred frequencies; this suggests that this neuron receives strong positive connection weights from two first-output-layer neurons with different frequency preferences, which is evident in Fig. 3A (b). This is similar to auditory cortical neurons classified as double U-shaped tuning curves [16]. The tuning curve c (Fig. 3A) is similar to circumscribed neurons in the auditory cortex [16], which do not respond to increased amplitude of sine waves at any frequency. Therefore, this neuron is selective for a tone with a specific frequency and intensity level. Physiologically, this type of neuron is reported to compose approximately 20 % of the neurons in a cat's auditory cortex [16].

Human Speech

Pitch Selectivity. Figure 3B shows the connections from first- to second-output-layer neurons in the network with human-speech input. Some of these neurons receive positive inputs from first-output-layer neurons with preferred frequencies that are multiples of a value (neuron d). That is, these neurons respond to the pitch of a tone, and, consequently, have intensive multimodal tuning curves (Fig. 4, neuron d). This type of stimulus preference has been reported in auditory cortical neurons [16]. Because the present learning algorithm maximizes the information transmission from input to output, this result suggests that pitch selectivity in the auditory cortex emerges to encode human and other animal voices efficiently. Indeed, human voices are primarily composed of tones with a principal frequency and its higher harmonics. Owing to this statistical property of human voices, second-output-layer neurons acquire selectivity to tones with frequencies f, $2f$, $3f$, ... during training.

4 Discussion

This study presented the properties of model neurons in networks trained to maximize the amount of information transmitted to the output layers, using natural sounds and human speech as inputs. The first-output-layer neurons respond to wavelet-like stimuli and have unimodal tuning curves. This property is consistent with the experimentally reported properties of auditory-nerve neurons [1]. It is also consistent with previous theoretical studies [11,13], showing that wavelet-like functions are information-efficient in encoding natural sounds. The properties of the second-output-layer neurons are affected by the type of input. Natural sounds, which contain abundant abrupt changes in pitch, favor the emergence of second-output-layer neurons selective for pitch change. In contrast, human voices are dominated by continuous waves with higher harmonics, making pitch selectivity advantageous in encoding information. These properties are also consistent with previous experimental and theoretical results [3,11,15].

The training algorithm of the present model is based on the information maximization principle, that is, the information conveyed by the output neurons is maximized during training. The fact that neurons in the two output layers exhibit stimulus preferences similar to the cochlear nucleus, medial geniculate, and auditory cortical neurons suggests that the neurons in the central nervous system have evolved to encode as much information as possible by forming an information-efficient circuit. This is corroborated by previous studies which showed that the properties of simple and complex cells in the primary visual cortex can be replicated by the information maximization model [7,17]. If, as suggested by the present model, the sensory information processing in the central nervous system can be understood in terms of information maximization, a model with a larger number of layers would replicate and predict the properties of neurons in the higher sensory cortices.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Numbers 15H04266 and 16K16123.

References

1. de Boer, E., de Jongh, H.R.: On cochlear encoding: potentialities and limitations of the reverse-correlation technique. J. Acoust. Soc. Am. **63**, 115–135 (1978)
2. Rouiller, E., de Ribaupierre, Y., Morel, A., de Ribaupierre, F.: Intensity functions of single unit responses to tone in the medial geniculate body of cat. Hear. Res. **11**, 235–247 (1983)
3. Bendor, D., Wang, X.: The neuronal representation of pitch in primate auditory cortex. Nature **436**, 1161–1165 (2005)
4. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**, 193–202 (1980)
5. Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex **1**, 1–47 (1991)
6. Bell, A.J., Sejnowski, T.J.: The “independent components” of natural scenes are edge filters. Vis. Res. **37**, 3327–3338 (1997)
7. Tanaka, T., Nakamura, K.: Information maximization principle explains the emergence of complex cell-like neurons. Front. Comput. Neurosci. **7**, 165 (2013)

8. Karklin, Y., Lewicki, M.S.: A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. Neural Comput. **17**, 397–423 (2005)
9. Karklin, Y., Lewicki, M.S.: Emergence of complex cell properties by learning to generalize in natural scenes. Nature **457**, 83–86 (2009)
10. Tanaka, T., Aoyagi, T., Kaneko, T.: Replicating receptive fields of simple and complex cells in primary visual cortex in a neuronal network model with temporal and population sparseness and reliability. Neural Comput. **24**, 2700–2725 (2012)
11. Karklin, Y., Ekanadham, C., Simoncelli, E.P.: Hierarchical spike coding of sound. In: Advances in Neural Information Processing Systems, pp. 3032–3040 (2012)
12. Smith, E.C., Lewicki, M.S.: Efficient auditory coding. Nature **439**, 978–982 (2006)
13. Lewicki, M.S.: Efficient coding of natural sounds. Nature Neurosci. **5**, 356–363 (2002)
14. Kiang, N.Y.S., Sachs, M.B., Peake, W.T.: Shapes of tuning curves for single auditory-nerve fibers. J. Acoust. Soc. Am. **42**, 1341–1342 (1967)
15. Britt, R., Starr, A.: Synaptic events and discharge patterns of cochlear nucleus cells II. Frequency-modulated tones. J. Neurophysiol. **39**, 179–194 (1976)
16. Sutter, M.L.: Shapes and level tolerances of frequency tuning curves in primary auditory cortex: quantitative measures and population codes. J. Neurophysiol. **84**, 1012–1025 (2000)
17. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. **7**, 1129–1159 (1995)