Invited Paper

Network model of predictive coding based on reservoir computing for multi-modal processing of visual and auditory signals

Yoshihiro Yonemura$^1$ and Yuichi Katori$^{1,2a}$

$^1$ The School of Systems Information Science, Future University Hakodate
116-2 Kamedanakano-cho, Hakodate, Hokkaido, 041-8655, Japan

$^2$ The Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo 153-8505, Japan

a) katori@fun.ac.jp

Received December 9, 2020; Revised January 8, 2021; Published April 1, 2021

Abstract: We propose a hierarchical network model based on predictive coding and reservoir computing as a model of multi-modal sensory integration in the brain. The network is composed of visual, auditory, and integration areas. In each area, the dynamical reservoir acts as a generative model that reproduces the time-varying sensory signal. The states of the visual and auditory reservoir are spatially compressed and are sent to the integration area. We evaluate the model with a dataset of time courses, including a pair of visual (hand-written characters) and auditory (read utterances) signal. We show that the model learns the association of multiple modalities of the sensory signals and that the model reconstructs the visual signal from a given corresponding auditory signal. Our approach presents a novel dynamical mechanism of the multi-modal information processing in the brain and the fundamental technology for a brain like an artificial intelligence system.

Key Words: reservoir computing, predictive coding, multi-modal integration

1. Introduction

In the brain, the perception of the ever-changing world is achieved not only by a bottom-up flow of sensory information but also by a top-down flow of internally generated information. Coordination of top-down and bottom-up flow of information is necessary for robust and flexible perception and is critical for the integration of multi-modal sensory information [1–3], which is one of the essential abilities in our daily life. We can mentally visualize images from the corresponding auditory information and imagine the auditory or linguistic sequence from the related visual information. Each modality has different properties, e.g., the visual signal has large spatial dimensions, and the auditory signal has a complex temporal structure. Underlying mechanisms of integration, operation, and imagination among those multiple modalities of sensors are mostly unknown.

Sensory information is processed in a hierarchical structure of the brain. The hierarchical network
can be modeled with the framework of predictive coding [4–6], which processes the sensory signal with the top-down and bottom-up process. The higher area performs as a generative model that predicts the lower area’s internal states or sensory signal in the lowest area. In the lower area, the internal state and its predicted state are compared, and the prediction error sends back to the higher area. The higher area updates the prediction based on the prediction error sent from the lower processing area and sends it again to the lower area. The iterations of the prediction and the error-feedback rectify the state of the perception. The iterations provide the prediction of sensory signals. However, the conventional predictive coding model assumes a linear dynamical system as the generative model, and the abilities of the prediction are limited.

The ability to reconstruct and predict the time-series, we refer to this ability as expressibility, depends on the properties of the generative model. Rao’s model assumed a linear dynamical system as the generative model without recurrent connection among the neurons. Friston's model assumes that the dynamics of the target system is required to be predetermined, and thus, the expressibility of the model is limited to the presupposed signals. The expressibility of those conventional predictive coding models is limited by those dynamics of the generative model. In the brain, on the other hand, neurons have non-linear response property with recurrent connections among many neurons, which exhibit complex dynamics and leads to high expressibility of time-varying sensory signal. These rich dynamics on a high-dimensional system might be key to establish high expressibility.

Time-varying complex sensory signals can be modeled with reservoir computing (RC) [7, 8], which is a framework for constructing recurrent neural networks and can be implemented in many physical systems [9, 10]. In the RC framework, the recurrent connections are randomly and sparsely configured and are not necessary to be trained. The readout connections from the reservoir will be trained so that the network reproduces a given target time-series. Thus, the training of the network can be done with less computational cost. The RC is viewed as a model of the brain; the most of those RC are utilized for generation of motion [8, 11–13].

The predictive coding with reservoir computing (PCRC) model [14] utilizes a reservoir as the generative model of the predictive coding. The reservoir generates the multi-dimensional time-varying sensory signal, and the prediction error is sent back to the reservoir for rectifying the internal state of the network. This model can reconstruct and predict the time-varying sensory signal. The model's parameter values need to be adjusted to the given sensory signal so that the model captures the sensory signal’s dynamical properties. Signals coming from different modalities have different dynamical properties, and thus the parameter values, e.g., a time constant of the reservoir, are required to be configured for each modality. In the brain, signals coming from different sensory organs are firstly processed in distinctive areas, and then processed signals send to the area that integrates multiple sensory information. That hierarchical structure is also suggested to be necessary for multi-modal processing in the brain [1].

In the present study, we propose a hierarchical PCRC model for the processing of multi-modal sensory signals. The hierarchical model is composed of multiple modules of the PCRC model. Specifically, in the present study, we use three PCRC modules: modules for sensory processing of visual and auditory signal and a module that integrate two modalities of sensory signals. In each sensory area, the reservoir reconstructs and predicts the given sensory signal in the prediction layer. The states of the reservoir in the two sensory reservoirs are compressed and are smoothed, and are sent to the input layer of the integration area. The reconstruction error of the compressed signal on the integration area network is fed back to the sensory area networks. In the following section, firstly, we describe the model of PCRC and the hierarchical PCRC model for multi-modal processing. We demonstrate that the single PCRC module reconstructs and predicts the high-dimensional time sensory signals. We then train the hierarchical PCRC model with pairs of visual and auditory signals and demonstrate that the model reconstructs the visual signal from the auditory signal.

2. Model
We propose a hierarchical network model based on PCRC [14] for processing the structured complex multi-dimensional time series. The overall network structure of the proposed model is shown in Fig. 1.
Fig. 1. Network structure of the PCRC models. (A) a module of the predictive coding based on reservoir computing. (B) PCRC based hierarchical model for the multi-modal processing of the visual and auditory processing.

The multi-modal processing model (Fig. 1(b)) is composed of PCRC modules (Fig. 1(a)). Firstly, we describe the single PCRC module, then connect three modules to construct the multi-modal processing model.

2.1 Predictive coding with reservoir computing module
The network in each module is composed of the prediction layer, the input layer, the prediction error layer, and the reservoir (Fig. 1(a)). In the module, the input signal on the input layer is reproduced on the prediction layer with the complex motion of the reservoir. The prediction error is fed back to the reservoir to reduce the prediction error. In the training phase, the connection from the reservoir to the prediction layer is modulated with first-order reduced and controlled error (FORCE) algorithm [11]. In the test phase, the model is operated with two different modes: the error-driven mode and the free-running mode. In the error-driven mode, the prediction error is fed back to the reservoir for reducing the prediction error. In the free-running mode, the prediction error is not sent to the reservoir, and the reservoir runs autonomously.

The PCRC module comprises the reservoir, the prediction layer, the input layer, and the prediction error layer. The notation of the module in each area is distinguished by the superscript ($i$). The membrane potential (or internal state) and the activities of the neurons in the reservoir are denoted...
represents strength of recurrent connection.

The states of the reservoir is updated by the following equations:

\[
m^{(i)}(n+1) = m^{(i)}(n) + \frac{1}{\tau^{(i)}} \{-m^{(i)}(n) + W_{\text{rec}}^{(i)} r^{(i)}(n) + W_{\text{back}}^{(i)} y^{(i)}(n) + \alpha^{(i)} W_e^{(i)} e^{(i)}(n) - b^{(i)}(n)\},
\]

\[
r^{(i)}(n+1) = \tanh(\beta_m^{(i)} m(n+1)),
\]

where \(W_{\text{rec}}^{(i)} \in \mathbb{R}^{N_x^{(i)} \times N_y^{(i)}}\) is matrix that represents recurrent connection. \(\tau^{(i)}\) is the time constant. \(n \in \mathbb{Z}\) is time steps. The activity of neurons is scaled by the parameter \(\beta_m^{(i)}\). The reservoir also receives inputs from the prediction layer \(y^{(i)} \in \mathbb{R}^{N_y^{(i)}}\) with the feedback connection \(W_{\text{back}}^{(i)} \in \mathbb{R}^{N_x^{(i)} \times N_y^{(i)}}\), the prediction error layer \(e^{(i)} \in \mathbb{R}^{N_y^{(i)}}\) with a coefficient \(\alpha_e^{(i)}\) that specifies the strength of the error feedback and switches the operation mode of the model, and further, the top-down input from higher area network \(b^{(i)}(n)\). The state of the prediction and prediction error layer are given by

\[
y^{(i)}(n) = \max(0, W_{\text{out}}^{(i)} r^{(i)}(n)),
\]

\[
e^{(i)}(n) = d^{(i)}(n) - y^{(i)}(n),
\]

The model has two operation modes: the error-driven mode and the free-running mode. In error-driven mode (\(\alpha_e^{(i)} = 1\)), the reservoir is updated with the prediction error, and the state of the prediction layer follows the state of the input layer. In free-running mode (\(\alpha_e^{(i)} = 0\)), the state of the reservoirs is updated by the internal dynamics independently from the sensory input.

The configuration and learning of the network are performed with the following procedure. The recurrent connection among the reservoir and the feedback connection from the prediction layer to the reservoir are configured randomly and sparsely, and its connectivity is non-necessary to be trained. In the training phase, the network is operated with the error-driven mode, the connection from the reservoir to the prediction layer is trained by the FORCE learning algorithm with a given time course dataset.

The recurrent connections of each reservoir \(W_{\text{rec}}^{(i)}\) is configured with the following procedure. Firstly, generate zero matrix \(W_0\) and set non-zero values \(-1\) or \(1\) to the randomly selected \(\beta_d^{(i)} \times N_x \times N_x\) elements. Then, calculate spectral radius of \(W_0\): \(|\rho_0|\). Define \(W_{\text{rec}}^{(i)} = \alpha_r^{(i)} W_0 / |\rho_0|\), where \(\alpha_r^{(i)}\) represents strength of recurrent connection.

The feedback connections of each reservoir \(W_{\text{back}}^{(i)}\) and \(W_{\text{out}}^{(i)}\) are configured with the following procedure. Similarly to the \(W_{\text{rec}}^{(i)}\), generate zero matrix \(W_0\) and set non-zero values \(-1\) or \(1\) to the randomly selected \(\beta_b^{(i)} \times N_x \times N_y\) elements. Define \(W_{\text{back}}^{(i)} = \alpha_b^{(i)} W_0\), with the strength of the feedback connections \(\alpha_b^{(i)}\). Use same procedure to generate \(W_{\text{out}}^{(i)}\) with coefficient \(\alpha_e^{(i)}\).

The readout connections from the reservoir \(W_{\text{out}}^{(i)}\) is updated with FORCE learning algorithm [11] as follows:

\[
v^{(i)}(n) = P^{(i)}(n) r^{(i)}(n),
\]

\[
P^{(i)}(n+1) = P^{(i)}(n) - \frac{v^{(i)}(n)v^{(i)T}(n)}{1 + v^{(i)T}(n)r^{(i)}(n)},
\]

\[
W_{\text{out}}^{(i)}(n+1) = W_{\text{out}}^{(i)}(n) - \frac{e_{\text{out}}(n)v^{(i)T}(n)}{1 + v^{(i)T}(n)r^{(i)}(n)}.
\]

The initial value of \(P^{(i)}(n) \in \mathbb{R}^{N_x^{(i)} \times N_x^{(i)}}\) is

\[
P^{(i)}(0) = \frac{I}{\alpha_f},
\]

where, the matrix \(I\) is identity matrix and \(\alpha_f\) is scaling parameter.

After the training of the readout connection, the module is ready to reconstruct the given input on the error-driven mode and to predict the input on the free-running mode.
2.2 Multi-modal processing model

Figure 1(b) shows the multi-modal processing model based on PCRC model. The network is composed of three areas: visual, auditory, and integration area. The states of the network in each area are updated by Eqs. (1)–(8). In the following equations, the three areas: visual, auditory, and integration area are distinguished by the superscript $k \in \{V, A, I\}$, respectively.

The activities of the visual and auditory reservoir in the lower area network send to the integration area. Those activities, which are a high dimensional vector, is compressed by the dimension reduction matrices $U^{-1}_V$ and $U^{-1}_A$, which are described later. The compressed vector is concatenated into $\mathbf{d}^{(I)}$, and the top-down signal for the visual and auditory area $\mathbf{b}^{(V)} \in \mathbb{R}^{N_y^{(V)}}$ and $\mathbf{b}^{(A)} \in \mathbb{R}^{N_y^{(A)}}$ are given as follows.

$$\mathbf{d}^{(I)}(n) = [U^{-1}_V r^{(V)}(n), U^{-1}_A r^{(A)}(n)]^T, \quad \text{(9)}$$

$$\mathbf{b}^{(V)}(n) = \alpha^{(V)}_{td} U(V) e^{(IV)}(n), \quad \text{(10)}$$

$$\mathbf{b}^{(A)}(n) = \alpha^{(A)}_{td} U(A) e^{(IA)}(n). \quad \text{(11)}$$

where $e^{(IV)}$ and $e^{(IA)}$ are the concatenated prediction error on the integration area: $[e^{(IV)}, e^{(IA)}]^T = e^{(I)}$.

Concatenated vector $\mathbf{d}^{(I)}$ is smoothed with time-constant $\tau_d$:

$$\mathbf{d}^{(I)}(n) = \mathbf{d}^{(I)}(n - 1) + \left(\mathbf{d}^{(I)}(n) - \mathbf{d}^{(I)}(n - 1)\right) / \tau_d. \quad \text{(12)}$$

Note that the integration reservoir does not receive top-down signal ($\mathbf{b}^{(I)}(n) = 0$).

The configuration and learning of the multi-modal processing model is performed with the following procedure. The recurrent connection and the feedback connection are configured as the above-described procedure. The connection matrix between the lower and higher $U^{(A)}$ and $U^{(V)}$, whose inverse matrices are used for the dimension reduction, are defined by the following procedure. Firstly, run the lower area network (visual and auditory area) on the error-driven mode ($\alpha^{(V)}_e = 1$ and $\alpha^{(A)}_e = 1$) without top-down signal ($\tilde{\alpha}_{td} = 0$) and collect the time course of the reservoirs $r^{(V)}$ and $r^{(A)}$ in the state collecting matrix $R^{(V)}$ and $R^{(A)}$, respectively. Secondary, compute the dimension reduction matrices $U^{(V)}$ and $U^{(A)}$. Suppose that $T$ timesteps of reservoir’s states are collected in $R^{(i)}$ ($i \in \{V, A\}$), $R^{(i)}$ can be decomposed by principal component analysis (PCA) so that $R^{(i)} = S^{(i)} U^{(i)}$. Where $S^{(i)}$ is $T \times 20$ matrix, and $U^{(i)}$ is an $N^{(i)}_y \times 20$ matrix. The dimension reduction matrix $U^{-1}_{(i)}$ can be obtained as the pseudo-inverse matrix of $U^{(i)}$ only when $U^T_{(i)} U^{(i)}$ is full-rank. Then, connect the sensory modules (the visual and auditory module) and the integration with the obtained $U^{(i)}$, and run the whole network with error-driven mode ($\alpha^{(V)}_e = \alpha^{(A)}_e = 1$) and $\alpha_{td} > 0$. The $W^{(V)}_{out}$, $W^{(A)}_{out}$, and $W^{(I)}_{out}$ are obtained with FORCE learning.

3. Results

Here, we show that the single PCRC module reconstructs and predicts the high-dimensional time-varying signals. Then, we train the hierarchical PCRC model with paired time courses of visual and auditory signals and demonstrate that the model reconstructs the visual signal from the auditory signal.

3.1 Reconstruction and prediction of time series on single PCRC module

We demonstrate that the single PCRC module generates and predicts the multi-dimensional time series. In the training phase, the network model is trained with training data; we use the two datasets: the periodic and chaotic time series. Then, in the test phase, the network runs with a time series on test dataset with the error-driven mode to reconstruct the given time course, then switched into the free-running mode to predict the time course. Both datasets used here are $N^{(i)}_y = 20$ dimensional time series.

The periodic time series $x_i(n), i \in \{1, \cdots, N_y\}, n \in \mathbb{Z}$ is a complex sinusoidal curves generated by the following equation. See Fig. 2(C).
written numbers and read utterances of corresponding numbers. We use three images of hand-written

\[
x_i(n) = \frac{1}{2} \sum_{p=1}^{3} \exp \left( -\frac{(i - 5p)^2}{4} \right) \left( \sin 2\pi \left( \frac{n}{100} + \frac{p}{5} \right) + \sin 2\pi \left( \frac{n}{200} + \frac{p}{5} \right) \right),
\]

where the index \( p \) goes through from 1 to 3 in the summation and specifies the oscillation pattern.

The chaotic time series is generated with the diffusively coupled mm-model oscillators [18], which is described by the following equations:

\[
\frac{dx_i}{dt} = -y_i - \mu x_i^2(x_i - \frac{3}{2}) + I + J_i,
\]

\[
\frac{dy_i}{dt} = -y_i + \mu x_i^2,
\]

where \( \mu = 1.5, I = 0.004, \) and \( J_i \) is diffusive coupling between the oscillators are given by

\[
J_i = \begin{cases} 
G(x_2 - x_1) & (i = 1), \\
G(x_{i+1} + x_{i-1} - 2x_i) & (i = 2, \cdots, N_y - 1), \\
G(x_{N_y-1} - x_{N_y}) & (i = N_y), 
\end{cases}
\]

where \( G = 0.3 \) is the coefficient of the diffusive coupling. The Eq. (14) and (15) are numerically integrated, and the obtained time series is discretized with time interval \( \Delta t = 0.2 \).

We use the following parameter values. \( N_x = 300, N_y = 20, \tau_m = 6, \alpha_r = 1.0, \alpha_f = 1.0, \alpha_b = 0.4, \alpha_c = 0.3, \beta_r = \beta_b = \beta_c = 0.05 \). Note that the following result is not sensitive to these parameter values, and a similar result can be obtained from a broad range of the parameter values.

Figure 2 and 3 show the typical response of the proposed model after the training with the two datasets. The network runs firstly with the error-driven mode, and then it switches to the free-running mode. For both datasets, the model well reproduces the given input time courses in the prediction layer in the error-driven mode. In the free-running mode, the prediction error (the difference between the predictive and input layers) tends to get larger as time passes.

In the periodic time series case (Fig. 2), the prediction error maintains a lower level in both the error-driven and free-running modes. At the beginning of the error-driven mode, the initial state of the reservoir is randomly configured. Thus, the prediction layer does not match the given input time course, and the small fluctuation of the prediction layer occurs (Fig. 2(D)). This fluctuation of the prediction error drives and rectifies the reservoir to generate accurate generation. The prediction error is converged into zero, and the reservoir is virtually driven autonomously. Even in the free-running mode, the prediction error is maintained near zero.

In the chaotic time series case (Fig. 3), the small fluctuation of the prediction error remains during the error-driven mode. The prediction error does not converge into zero, while the overall time course of the given input can be reproduced in the prediction layer. The reservoir autonomously generates a chaotic pattern immediately after switching into the free-running mode. In the free-running mode, the oscillation on the reservoir and the oscillatory pattern on the prediction layer are dumped.

Those results indicate that the small fluctuation of the prediction error utilized as the driver for the reservoir realizes the reproduction of the complex pattern of high-dimensional signal in the prediction layer. In the periodic case, the autonomous dynamics of the reservoir reproduce the periodic time series. Even in the chaotic case, a small perturbation for the reservoir reflecting the prediction error realizes the reproduction of the chaotic time series.

### 3.2 Multi-modal model of visual and auditory signal processing

In the hierarchical PCRC model for visual and auditory processing, the integration reservoir conducts on the reconstruction and the prediction of the compressed and concatenated state of the sensory reservoirs. Thus, the integration reservoir is expected to reconstruct information of one modality from the other modality information. The visual and auditory reservoirs are driven, in addition to the prediction error on each sensory module, by the prediction error on the integration reservoir.

The multi-modal model is evaluated with time-series data, including a pair of images of hand-written numbers and read utterances of corresponding numbers. We use three images of hand-written
Fig. 2. Typical responses of the single PCRC mode with the dataset of complex sinusoidal time series. (A) The state of the reservoir $m(n)$. (B) The state of the prediction layer $y(n)$. (C) The input for the module $d(n)$. (D) The prediction error $e(n)$. The region of whiteout indicates that the prediction error is converged into almost zero. The boundary of the error-driven mode and the free-running mode is indicated by the vertical dashed line.

numbers: “2”, “5”, and “9” from the MNIST dataset for the visual signal [17]. Each image is composed of $28 \times 28$ (784) grayscale pixels. The images are preprocessed by non-negative matrix factorization (NMF) and transformed into the 20-dimensional signal. Suppose that $V$ is $L \times 784$ matrix, each row of which represents a single image, and $V$ is a collection of $L$ images, NMF decomposes $V$ into two matrices: $V = HW$, where $H$ is $L \times 20$ coefficient matrix, and $W$ is $20 \times 784$ feature matrix. We use a transformed 20-dimensional vector as the input for the visual area network. The coefficient vector reconstructed by the PCRC module can be transformed into the images with $W$. We also use linguistic data containing the read utterance of corresponding numbers from T146 datasets for the auditory signal [15]. The dataset is an uncompressed audio dataset. Each dataset is preprocessed by a cochlear filter model [16], which is an auditory model showing the propagation of sound in the inner ear and conversion of acoustical energy into neural representations. The given auditory signals are transformed into 55 dimension signals. Figure 4 shows samples of the dataset. In the auditory signal, the initiation of the read utterance has jitter in the dataset. The utterance starts from 60 to 90 timesteps. The corresponding visual signals are presented from 80 to 160 timesteps without jitter.

In the training phase, modules in all three areas are operated in the error-driven mode, and 60
Fig. 3. Typical responses on the single PCRC mode with datasets of chaotic time series. (A) The state of the reservoir $m(n)$. (B) The state of the prediction layer $y(n)$. (C) The input for the module $d(n)$. (D) The prediction error $e(n)$.

Pairs of visual and auditory signals (including 20 pairs for each category) are utilized. We use the following parameter values for the training and the evaluation. $N_v^{(V)} = 200$, $N_x^{(A)} = 500$, $N_x^{(I)} = 400$, $N_y^{(V)} = 20$, $N_y^{(A)} = 55$, $N_y^{(I)} = 40$, $\tau^{(V)} = 1$, $\tau^{(A)} = 1$, $\tau^{(I)} = 5$, $\tau_d = 10$, $\alpha_f = 1.0$, $\alpha_v^{(V)} = \alpha_v^{(A)} = 0.01$, $\alpha_v^{(I)} = 0.15$, $\alpha_v^{(V)} = \alpha_v^{(A)} = 0.1$, $\alpha_v^{(I)} = 0.05$, $\beta_v^{(V)} = \beta_v^{(A)} = 0.1$, $\beta_v^{(I)} = 0.05$, $\beta_v^{(V)} = \beta_v^{(A)} = 0.1$, $\beta_v^{(I)} = 0.1$, $\beta_m^{(V)} = 0.1$, $\beta_m^{(A)} = 10$, $\beta_m^{(I)} = 5$.

After the training, the network is expected to reconstruct sensory information on one modality based on the signals coming from the other modality. In the following, we focus on reconstructing the visual information under the presentation of the corresponding auditory signals. Here, the auditory reservoir and the integration reservoir are in the error-driven mode, and the visual reservoir is in the free-running mode.

Figure 5 shows typical time courses of the states of the module in each area. Firstly, the auditory signal is applied in the input layer of the auditory area (Fig. 5(B)), and it induces the fluctuation in the auditory reservoir (Fig. 5(A)). The state of the auditory reservoir is compressed and transferred to the input layer of the integration reservoir (blue curves in Fig. 5(E)), and it induces the oscillation in the integration reservoir (Fig. 5(C)). Then, the integration reservoir reconstructs the signal, including not only the auditory signal but also the corresponding visual signal (Fig. 5(D)). At this moment,
Fig. 4. Samples of datasets used in the training: time course of pairs of (A) visual signals (hand-written numbers) and (B) auditory signal (read utterance of the corresponding numbers). Each visual signal is presented with 100 timesteps. The visual signal is preprocessed by NMF, and a decomposed 20-dimensional signal is applied as the sensory input on the visual area network. The auditory signal is preprocessed with a cochlear filter and transform into the 55-dimensional signal. The signal is applied to the sensory input on the auditory area network. The period of the presentation of the visual signal is indicated by the horizontal bar.

the visual reservoir is in silence, and there is no input for the integration area from the visual area. Thus, a large prediction error occurs in channels of visual information in the integration reservoir (the arrow in Fig. 5(F)). This prediction error triggers oscillation in the visual reservoir (Fig. 5(G)) that reconstructs the corresponding visual signal in the prediction layer of the visual area (Fig. 5(H)). The oscillation in the visual reservoir is compressed and sent to the input layer of the integration area and results in the reduction of the prediction error (Fig. 5(F)). The different categories of auditory signals induce different patterns of oscillation in the reservoirs. The principal components of the orbits in the reservoir are well separated (Fig. 6).

Figure 7(A) shows a typical time course of reconstructed visual images in the visual area while the corresponding auditory signal is given in the auditory area. The image is based on the state of the prediction layer in the visual area. The state of the prediction layer is transformed with the feature matrix of NMF, and the brightness of the image in Fig. 7(A) corresponds to this transformed vector. The performance of the reconstruction is quantified with the zero-mean normalized cross-correlation (ZNCC) between the reconstructed image and the images in the training dataset:

\[ z = \frac{\sum_{x,y} (f(x,y) - \bar{f})(g(x,y) - \bar{g})}{\sqrt{\sum_{x,y} (f(x,y) - \bar{f})^2 \sum_{x,y} (g(x,y) - \bar{g})^2}}, \]  

(16)

where \( x, y \) are indexes of the pixel in the images, \( f(x,y) \) and \( g(x,y) \) are the brightness of the images, and the \( \bar{f} \) and \( \bar{g} \) are the mean of the images \( f(x,y) \) and \( g(x,y) \), respectively. We denote ZNCC at time step \( n \) between the reconstructed image and the image in a category \( c \) as \( z^{(c)}(n) \). We also define a reconstruction score \( R \) as follows:
Fig. 5. Typical time courses of the state of the network in the process of reconstruction of visual signal from a given auditory signal. (A), (B): The state of the reservoir (A) and input layer (B) in the auditory area. (C)–(F): The state of the reservoir (C), the prediction layer (D), the input layer (E), and the normalized prediction error (F) in the integration reservoir. The normalized prediction error is defined as $||e^{(I)}(t)||_2^2 / \max_t ||e^{(I)}(t)||_2^2$. (G),(H): The state of the reservoir (G) and the prediction layer (H). For the prediction, input, and error in the integration area, the concatenate vector of the auditory and visual signal are indicated by blue and red curves, respectively.

$$R = \frac{1}{N_{total}} \sum_{c \in C} \sum_{n \in N_c} \left( z^{(c)}(n) - \tilde{z}^{(c)}(n) \right),$$

where $C$ is the set of classification of signals, namely $C = \{"2", "5", "9"\}$, $N_c$ are sets of period that...
Fig. 6. Trajectories of principal components of the internal state of the auditory, visual, and integration reservoir. The three different categories of the signal, namely “two”, “five”, and “nine” are indicated by blue, green, and red curves, respectively.

Fig. 7. Typical time courses of the reconstructed signals in the process of reconstruction of visual signal from a given auditory signal. (A) The sequences of the image that inversely transformed from the feature vector in the prediction layer of the visual area. (B) The time courses of zero-mean normalized cross-correlation (ZNCC) between the reconstructed images and the images in the dataset. Three panels show the responses for the auditory signals “two”, “five”, and “nine”.

class $c$ is given as the auditory input, $N_{total}$ is total length of the period $N_{total} = \sum_c |N_c|$, $z^{(c)}(n)$ is ZNCC for the target (correct) class, and $\bar{z}^{(c)}(n)$ is average of ZNCC for the non-target (incorrect) class defined by the following equation:

$$z^{(c)}(n) = \frac{1}{|C| - 1} \sum_{c' \in C \setminus c} z^{(c')}(n) \quad (18)$$

Figure 7(B) shows the time course of ZNCC indicating that the visual images associated with the given auditory signal appear in the prediction layer of the visual area. The ZNCC corresponding to the given auditory signal increases during the presentation of the auditory signal.

Figure 8 shows the dependency of the reconstruction score $R$ defined by Eq. (17) on the time-constant of the integration reservoir $\tau^{(I)}$ and the strength of the recurrent connection in the integration reservoir $\alpha_r^{(I)}$. The points and vertical bars in each plot represent the mean and standard
Fig. 8. Dependencies of the reconstruction score $R$ on (A) the time constant of the integration reservoir $\tau^{(I)}$ and on (B) strength of the recurrent connection in the integration reservoir $\alpha^{(I)}$.

deviation of the reconstruction score, which was quantified with randomly generated 20 different network configurations. The reconstruction score increases as the time-constant $\tau^{(I)}$ increases, and the $R$ reached maximum value near $\tau^{(I)} = 60$ (Fig. 8(A)). This result suggests that the processing on the integration area needs to be slower than the sensory area’s dynamics. In Fig. 8(B), $R$ reached the maximum value near $\alpha^{(I)}_{r} = 0.2$. In the range $\alpha^{(I)}_{r} > 0.2$, the reconstruction score decreased, and its variance increases. The weaker the recurrent connection strength, the faster the reservoir’s state damp, and conversely, the stronger the recursive recurrent connection, the slower the reservoir’s state damp. The strength of the recurrent connection is necessary to be in the appropriate range. If the strength is weaker than this range, the reservoir can not induce the appropriate amplitude of fluctuation on the prediction layer. If the recurrent connection is too strong, the reservoir’s state tends to be chaotic, and the chaotic behavior leads to the large deviation of the reconstruction score on different network configurations (Fig. 8(B)).

4. Conclusions

We proposed the hierarchical network model of predictive coding based on reservoir computing that processes multi-modal sensory information. The model has based on predictive coding, in which the internal network or generative model predicts sensory signal, and the prediction error is fed back into the internal network. We utilize reservoirs as the generative model of the predictive coding to predicts the complex time-varying pattern of the sensory signal. In the hierarchical network model composed of the sensory (visual and auditory) reservoirs and the integration reservoir, each reservoir learns spatio-temporal dynamics of the sensory signal, and the reservoir conducts the reconstruction and prediction. The model successfully reconstructs the visual signal from the given auditory signal with the reservoirs’ internal dynamics.

In the processing of the multi-dimensional complex time courses, the proposed hierarchical model stacks the mechanism of the accumulation of the temporal structure and the compression of spatial patterns. The incoming signal is reconstructed with the reservoir that accumulates the temporal structure of the signal in its high dimensional nonlinear dynamics. Then, the high dimensional state vector on the reservoir, including the information of a short history of the signal, is spatially compressed and transferred into the integration area network. The stacking of the accumulation and the compression leads to a higher-order abstraction of the complex time course.

In the present paper, we have focused on using the dataset including three categories of numbers, namely “2”, “5”, and “9”, which is distinguishable in the reconstructed images and the feature space of the images. We have confirmed that the network trained with five categories of numbers (from “0” to “4”) successfully reconstruct learned numbers. However, if we use the dataset including ten categories of numbers (from “0” to “9”), the reconstruction score was decreased; for a given auditory signal, an incorrect image but has a similar feature to the given auditory signal was reconstructed. The capacity of the reconstructable number of categories is depended on the similarity among the
categories. The evaluation of this capacity might be an important issue in the future.

This hierarchical network model proposed in this paper can be understood as a model of the brain. A possible mapping between the elements in the predictive coding model and the microcircuits on the laminar structure of the cortex is proposed [19]. Still, the several circuit details are yet to be verified. There has been increasing interest in the underlying neural computational mechanism of the psychosis [20]. One possible approach involves predictive coding suggested that a decreased precision of the prediction (or increased mismatches between the internally derived prediction and incoming sensory signals) results in psychosis [20]. In the future, our approach should be evaluated with a viewpoint from those neuroscience or physiology.

There are many possible directions for future researches. In the current study, we have focused on reconstructing the visual signal from the auditory signal. In the future, we improve the model to reconstruct the auditory signal from the given visual signal. The model should be evaluated with a variety of sensory signals, including signals recorded in a field of engineering. Our modeling study may contribute to a fundamental technology for brain-like artificial intelligence.

Acknowledgments

This paper is based on results obtained from a project, JPNP16007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was also supported by JSPS KAKENHI Grant Number 20H04258 and JST CREST Grant Number JPMJCR18K2, Japan.

References

[1] M.H. Giard and F. Peronnet, “Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study,” Journal of Cognitive Neuroscience., vol. 1, no. 5, pp. 473–490, 1999. DOI: 10.1162/089892999563544
[2] M.H. Giard, F. Peronnet, P.B. Badcock, K.J. Friston, M.J.D. Ramstead, A. Ploeger, and J. Hohwy, “The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior,” Cognitive, Affective, & Behavioral Neuroscience, vol. 19, no. 6, pp. 1319–1351, December 2019. DOI: 10.3758/s13415-019-00721-3
[3] B.E. Stein and T.R. Stanford, “Multisensory integration: current issues from the perspective of the single neuron,” Nature Review Neuroscience, vol. 9, no. 4, pp. 255–266, April 2008. DOI: 10.1038/nrn2331
[4] R. Rao and D. Ballard, “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects,” Nature Neuroscience, vol. 2, pp. 79–87, 1999. DOI: 10.1038/4580
[5] R.L. Gregory, “Perceptions as hypotheses,” Philosophical Transaction of the Royal Society B Biological Science, vol. 290, no. 1038, pp. 181–197, 1980. DOI: 10.2307/2395424
[6] K. Friston, “Hierarchical models in the brain,” PLoS Computational Biology, vol. 4, no. 11, pp. e1000211, 2008. DOI: 10.1371/journal.pcbi.1000211
[7] H. Jaeger, “Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the “Echo State Network” approach,” GMD Report, vol. 5, 2002.
[8] W. Maass, T. Natschläger, and H. Markram, “Real-time computing without stable states: a new framework for neural computation based on perturbations,” Neural Computation, vol. 14, no. 11, pp. 2531–2560, November 2002. DOI: 10.1162/089976602760407955
[9] K. Nakajima, “Physical reservoir computing – An introductory perspective,” arXiv, nlin.AO, 2005.00992, May 2020. DOI: 10.35848/1347-4065/ab8daf
[10] G. Tanaka, T. Yamane, J.B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, “Recent advances in physical reservoir computing: A review,” Neural Networks, vol. 115, pp. 100–123, July 2019. DOI: 10.1016/j.neunet.2019.03.005
[11] D. Sussillo and L.F. Abbott, “Generating coherent patterns of activity from chaotic neural networks,” Neuron, vol. 63, no. 4, pp. 544–557, August 2009. DOI: 10.1016/j.neuron.2009.07.018
[12] T. Yamazaki and S. Tanaka, “Computational models of timing mechanisms in the cerebellar granular layer,” *Cerebellum*, vol. 8, no. 4, pp. 423–432, December 2009. DOI: 10.1007/s12311-009-0115-7

[13] K. Tokuda, N. Fujiwara, A. Sudo, and Y. Katori, “Chaos may enhance expressivity in cerebellar granular layer,” arXiv:2006.11532v1 [q-bio.NC], June 2020.

[14] Y. Katori, “Network model for dynamics of perception with reservoir computing and predictive coding,” in *Advances in Cognitive Neurodynamics (VI)*, ed. J.M. Delgado-Garcia, X. Pan, R. Sanchez-Campusano, R. Wang, pp. 89–95, Springer Nature, Singapore, 2017. DOI: 10.1007/978-981-10-8854-4_11

[15] Visit https://catalog.ldc.upenn.edu/LDC93S9 Linguistic data consortium, TI 46-Word.

[16] R. Lyon, “A computational model of filtering, detection, and compression in the cochlea,” *Proceedings of the IEEE*, vol. 86, no. 11 pp. 2278–2324, 1998. *ICASSP ’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1282–1285, May 1982. DOI: 10.1109/ICASSP.1982.1171644

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11 pp. 2278–2324, 1998. DOI: 10.1109/5.726791

[18] I. Tsuda, H. Fujii, S. Tadokoro, T. Yasuoka, and Y. Yamaguti, “Chaotic itinerancy as a mechanism of irregular changes between synchronization and desynchronization in a neural network,” *J. Integr. Neurosci.*, vol. 3, no. 2, pp. 159–182, June 2004. DOI: 10.1142/S021963520400049X

[19] S. Shipp, “Neural elements for predictive coding,” *Front. Psychol.*, vol. 7, pp. 1792(21 pages), November 2016. DOI: 10.3389/fpsyg.2016.01792

[20] S. Sterzer, R.A. Adams, P. Fletcher, C. Frith, S.M. Lawrie, L. Muckli, P. Petrovic, P. Uhlhaas, M. Voss, and P.R. Corlett, “The Predictive Coding Account of Psychosis,” *Biol. Psychiatry*, vol. 84, no. 9, pp. 634–643, November 2018. DOI: 10.1016/j.biopsych.2018.05.015