frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 20 March 2023
DOI 10.3389/fnins.2023.1132269

Explaining cocktail party effect and McGurk effect with a spiking neural network improved by Motif-topology

Shuncheng Jia[1,2†], Tielin Zhang[1,2*†], Ruichen Zuo[3] and Bo Xu[1,2,4*]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3]School of Information and Electronics, Beijing Institute of Technology, Beijing, China, [4]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

EDITED BY
Malu Zhang,
National University of Singapore, Singapore

REVIEWED BY
Pengfei Sun,
Ghent University, Belgium
Zihan Pan,
Institute for Infocomm Research (A*STAR), Singapore

*CORRESPONDENCE
Tielin Zhang
✉ tielin.zhang@ia.ac.cn
Bo Xu
✉ xubo@ia.ac.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Neuromorphic Engineering, a section of the journal Frontiers in Neuroscience

RECEIVED 27 December 2022
ACCEPTED 03 March 2023
PUBLISHED 20 March 2023

CITATION
Jia S, Zhang T, Zuo R and Xu B (2023) Explaining cocktail party effect and McGurk effect with a spiking neural network improved by Motif-topology. *Front. Neurosci.* 17:1132269. doi: 10.3389/fnins.2023.1132269

Network architectures and learning principles have been critical in developing complex cognitive capabilities in artificial neural networks (ANNs). Spiking neural networks (SNNs) are a subset of ANNs that incorporate additional biological features such as dynamic spiking neurons, biologically specified architectures, and efficient and useful paradigms. Here we focus more on network architectures in SNNs, such as the meta operator called 3-node network motifs, which is borrowed from the biological network. We proposed a Motif-topology improved SNN (M-SNN), which is further verified efficient in explaining key cognitive phenomenon such as the cocktail party effect (a typical noise-robust speech-recognition task) and McGurk effect (a typical multi-sensory integration task). For M-SNN, the Motif topology is obtained by integrating the spatial and temporal motifs. These spatial and temporal motifs are first generated from the pre-training of spatial (e.g., MNIST) and temporal (e.g., TIDigits) datasets, respectively, and then applied to the previously introduced two cognitive effect tasks. The experimental results showed a lower computational cost and higher accuracy and a better explanation of some key phenomena of these two effects, such as new concept generation and anti-background noise. This mesoscale network motifs topology has much room for the future.

KEYWORDS

spiking neural network, Motif topology, reward learning, cocktail-party effect, McGurk effect

1. Introduction

Spiking neural networks (SNNs) are considered the third generation of artificial neural networks (ANNs) (Maass, 1997). The biologically plausible network architectures, learning principles, and neuronal or synaptic types of SNNs make them more complex and powerful than ANNs (Hassabis et al., 2017). It has been reported that even a single cortical neuron with dendritic branches needs at least a 5-to-8-layer deep neural network for finer simulations (Beniaguev et al., 2021), whereby non-differential spikes and multiply-disperse synapses make SNNs powerful on tools for spatially-temporal information processing. In the field of spatially-temporal information processing, there has been much research progress significant amounts of research into SNNs for auditory signal recognition (Shrestha and Orchard, 2018; Sun et al., 2022) and visual pattern recognition (Wu et al., 2021; Zhang M. et al., 2021).

This paper highlights two fundamental elements of SNNs and the main differences between SNNs and ANNs: specialized network design and learning principles. The SNNs encode spatial information using fire rate and temporal information using spike timing, providing hints and inspiration that SNNs can integrate into visual and audio sensory data.

For the network architecture, specific cognitive topologies developed *via* evolution are highly sparse and but efficient in SNNs (Luo, 2021), whereas equivalent ANNs are densely recurrent. Many researchers attempt have tried to understand the biological nature of efficient multi-sensory integration by focusing on the visual and auditory pathways in biological brains (Rideaux et al., 2021). These structures are adapted for some specific cognitive functions, e.g., efficient actions. For example, an impressive sparse network filtered from the *C. Elegans* connectome can outperform other dense networks during reinforcement learning of the Swimmer task. Some biological discoveries can further promote the research development of structure-based artificial operators, including but not limited to lateral neural interaction (Cheng et al., 2020), the lottery hypothesis (Frankle and Carbin, 2018), and meta structure of network motif (Hu et al., 2022; Jia et al., 2022). ANNs using these structure operators can then be applied in different spatial or temporal information processing tasks, such as image recognition (Frankle et al., 2019; Chen et al., 2020), auditory recognition, and heterogeneous graph recognition (Hu et al., 2022). Furthermore, when only focusing on the learning of weight, the weight agnostic neural network (Gaier and Ha, 2019; Aladago and Torresani, 2021) is a representative of the methods that only train the connections instead of weights.

For the learning principles, SNNs are more tuned affected by learning principles from biologically plausible plasticity principles, such as spike-timing dependent plasticity (STDP) (Zhang et al., 2018a), short-term plasticity (STP) (Zhang et al., 2018b), and reward-based plasticity (Abraham and Bear, 1996), instead of by the pure multi-step backpropagation (BP) (Rumelhart et al., 1986) of errors in ANNs. The neurons in SNNs will be activated once the membrane potentials reach their thresholds, which makes them energy efficient. SNNs have been successfully applied on to visual pattern recognition (Diehl and Cook, 2015; Zeng et al., 2017; Zhang et al., 2018a,b, 2021a,b), auditory signal recognition (Jia et al., 2021; Wang et al., 2023), probabilistic inference (Soltani and Wang, 2010), and reinforcement learning (Rueckert et al., 2016; Zhang D. et al., 2021).

For the two classic cognitive phenomena, the cocktail party effect describes the phenomenon that in a high-noise environment (e.g., noise from the environment or other speakers), the listener learns to filter out the background noise (including music noise and sounds from other speakers) and concentrate on only the target speaker, as shown in Figure 1A. The McGurk effect introduces the concept that the voice may be misclassified when the auditory stimulus conflicts with visual cues. A classic example of the McGurk effect describes how the new concept [da] can be generated by the integration of specific auditory input [ba] and visual cues [ga], as shown in Figure 1B.

This work focuses on the key characteristics of SNNs in information integration, categorization, and cognitive phenomenon simulation. We analyzed Motifs (Milo et al., 2002) in SNNs to reveal the essential functions of key meta-circuits in SNNs and biological networks and then used Motif structures to build loop modules in SNNs. Furthermore, a Motif-topology improved SNN (M-SNN) is proposed for simulating cocktail party effects and McGurk effects. To the best of our knowledge, we are the first to solve the problem using combinations of highly abstract Motif units. The following are the primary contributions of this paper:

- Networks with specific spatial or temporal types of Motifs can improve the accuracy of spatial or temporal classification tasks compared with networks without Motifs, making the multi-sensory integration easier by integrating two types of Motifs.
- We propose a method to mix different Motif structures and use them to simulate cognitive phenomena, including cocktail party effects and McGurk effects. In addition, the Motif topologies are critical, and networks with Motifs could effectively simulate these two effects (higher accuracy and better cognitive phenomenon simulation). (We specifically picked the MNIST and TIDigits datasets to simulate audio-visual inputs due to the lack of audio-visual-consistent datasets for classification testing.)
- During the network training process for various simulation experiments, the M-SNN can achieve a lower training computational cost than other SNNs without using Motif architectures. This result demonstrates that the M-SNN can achieve more human-like cognitive functions at a lower computational cost with the help of prior knowledge of multi-sensory pathways and biologically inspired reward learning methods.

The remaining parts are grouped as follows: Section 2 reviews the research about on the architecture, learning paradigms, and two classic cognitive phenomena. Section 3 describes the pattern of Motifs, the SNN model with neuronal plasticity, and learning principles. Section 4 verifies the convergence, the advantage of M-SNN in simulating cognitive phenomena, and the computational cost. Finally, a short conclusion is given in Section 5.

2. Related works

For the architecture, the lateral interaction of neural networks, the lottery hypothesis, and the network motif circuits are novel operators in structure research. In the research on lateral interaction, most studies have taken the synapse as the basic unit, including the lateral interaction in the convolutional neural network (Cheng et al., 2020) or that in the fully connected network (Jia et al., 2021). However, these methods take synaptic connections as the basic unit and only consider learning effective structures without considering meta-structure composition.

Network motifs (Milo et al., 2002; Prill et al., 2005) use primary n-node circuit operators to represent the complex network structures. The feature of the network (e.g., visual or auditory pathways) could be reflected by the number of different Motif topologies, which is called Motif distribution. To calculate the Motif distribution, the first Motif tool is mfinder, which implements the algorithm of full enumeration (randomly picking the edges from

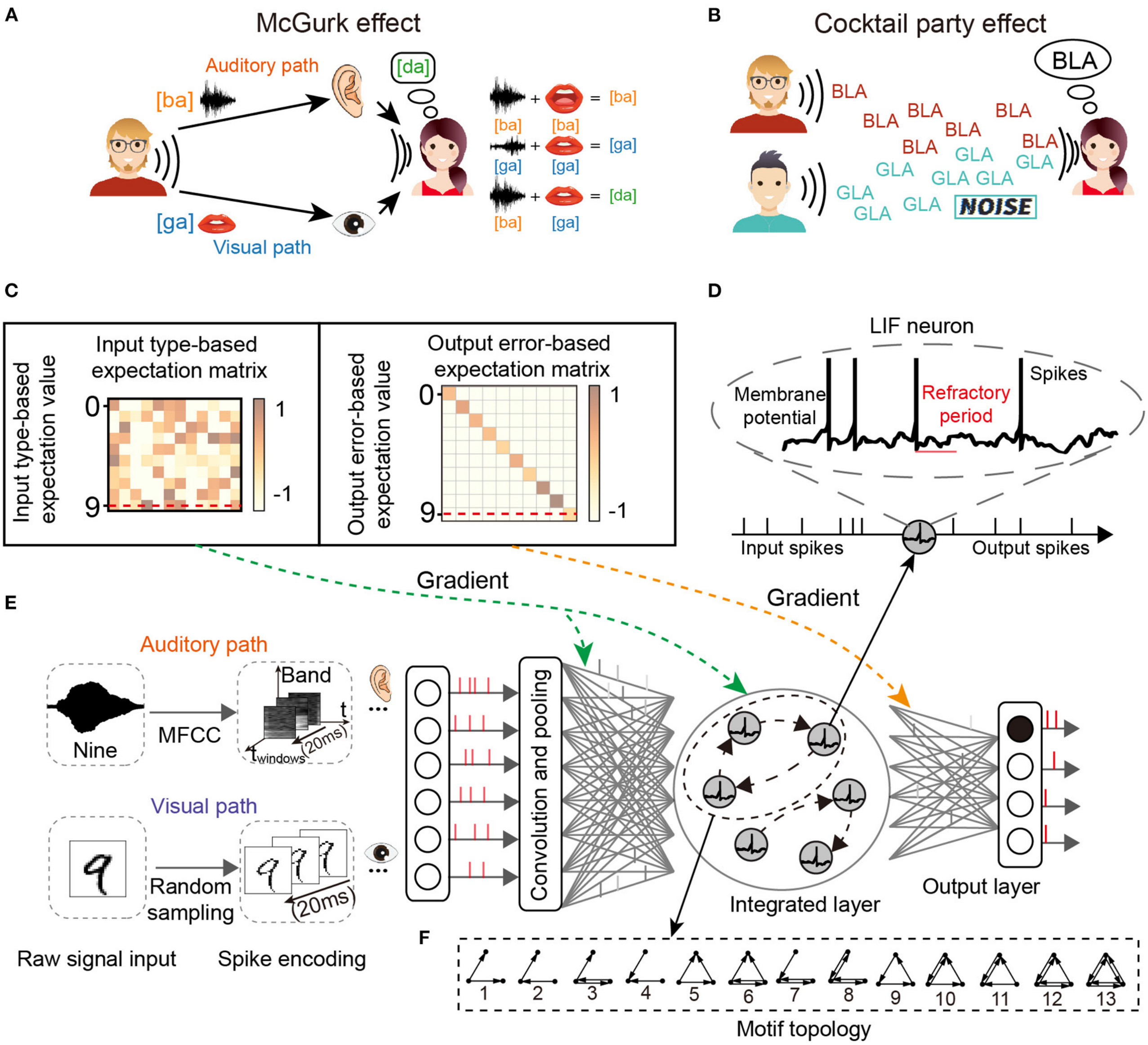


FIGURE 1
The network structure for multi-sensory integration and two cognitive phenomena. **(A)** McGurk effect. New concepts arise when the receiver receives different audio-visual input information. **(B)** Cocktail party effect. When the receiver's brain focuses on one speaker, it filters out the sounds and noise from others. **(C)** Input and output transformation matrix of reward learning. **(D)** The spiking neuron has a variable membrane potential. **(E)** The M-SNN network for single-sensory or multi-sensory integration tasks. **(F)** The example of 3-node Motifs.

the graph and counting the probability of *n*-node subgraphs). Then the FANMOD (Wernicke and Rasche, 2006) was introduced as a more efficient tool for finding reliable network motifs.

For learning paradigms, there are many methods have been proposed, such as the ANN-to-SNN conversion (i.e., directly training ANNs and then equivalently converting to SNNs; Diehl et al., 2015), proxy gradient learning (i.e., replacing the non-differential membrane potential at firing threshold by an infinite gradient value; Lee et al., 2016), and the biological-mechanism inspired algorithms [e.g., the SBP (Zhang et al., 2021a) which was inspired by the synaptic plasticity rules in the hippocampus, the BRP (Zhang et al., 2021b), which was inspired by the reward learning mechanism, and the GRAPES, that inspired by the synaptic scaling (Dellaferrera et al., 2022)]. Compared to other learning algorithms, biologically inspired algorithms are more similar to the process of how the human brain learns.

For the cocktail party effect, many effective end-to-end neural network models have been proposed (Ephrat et al., 2018; Chao et al., 2019; Hao et al., 2021; Wang et al., 2021). However, the analysis of why these networks work is very difficult since the functional structures in these black-box models are very dense without clear function diversity. As a comparison, the network motif constraint in neural networks might resolve this problem to some extent, which until now and as far as we know, however this has not yet been well-introduced.

For the McGurk effect, only a limited number of research papers have discussed the artificial simulation of it, partly caused by the simulation challenge, especially on the conflict fusion of visual and auditory inputs (McGurk and MacDonald, 1976; Hirst et al., 2018), e.g., self-organized mapping (Gustafsson et al., 2014).

3. Methods

3.1. Spiking dynamics

The leaky integrated-and-fire (LIF) neuron model is biologically plausible and is one of the simplest models to simulate spiking dynamics. It includes non-differential membrane potential and the refractory period, as shown in Figure 1D. The LIF neuron model simulates the neuronal dynamics with the following steps.

First, the dendritic synapses of the postsynaptic LIF neuron will receive presynaptic spikes and convert them to a postsynaptic current (I_{syn}). Second, the postsynaptic membrane potential will be leaky or integrated, depending on its historical experience. The classic LIF neuron model is shown as the following Equation (1).

$$\tau_m \frac{dV_t}{dt} = -\left(V_t - V_L\right) - \frac{g_E}{g_L}\left(V_t - V_E\right) + \frac{I_{syn}}{g_L}, \tag{1}$$

where V_t represents the dynamical variable of membrane potential with time t, dt is the minimal simulation time slot (set as 0.01ms), τ_m is the integrative period, g_L is the leaky conductance, g_E is the excitatory conductance, V_L is the leaky potential, V_E is the reversal potential for excitatory neuron, and I_{syn} is the input current received from the synapses in the previous layer. We set values of conductance (g_E, g_L) to be 1 in our following experiments for simplicity, as shown in Equation (3).

Third, the postsynaptic neuron will generate a spike once its membrane potential V_t reaches the firing threshold V_{th}. At the same time, the membrane potential V will be reset as the reset potential V_{reset}, shown as the following Equation (2).

$$\text{if } \left(V_t > V_{th}\right) \begin{cases} V_t = V_{reset} \\ T_{ref} = T_0 \end{cases}, \tag{2}$$

where the refractory time T_{ref} will be extended to a larger predefined T_0 after firing.

In our experiments, the three steps for simulating the LIF neurons were integrated into the Equation (3).

$$C\frac{dV_{i,t}}{dt} = g\left(V_{i,t} - V_{rest}\right)\left(1 - S_{i,t}\right) + \sum_{j=1}^{N} W_{i,j}X_{j,t}, \tag{3}$$

where C is the capacitance parameter, $S_{i,t}$ is the firing flag of neuron i at timing t, $V_{i,t}$ is the membrane potential of neuron i at timing t, V_{rest} is the resting potential, and $W_{i,j}$ represents the synaptic weight between the neuron i and j.

3.2. Motif topology

The n-node ($n \geq 2$) meta Motifs have been proposed in past research. Here, we use the typical 3-node Motifs to analyze the networks, which have been widely used in biological and other systems (Milo et al., 2002; Shen et al., 2012; Zhang et al., 2017). Figure 1F displayed all 13 varieties of 3-node Motifs. In previous studies, network topology had been transformed into parameter embeddings in the network (Liu et al., 2018). In our SNNs, the Motifs were used by the Motif masks and then applied into the recurrent connection at the hidden layer. The typical Motif mask is a matrix padded with 1 or 0, where 1 and 0 represent the connected and non-connected pathways, respectively. We introduce the Motif circuits into the hidden layer, and the Motif mask in the r-dimension hidden layer l at time t is represented as the $M_t^{r,l}$ as shown in Equation (4). As shown in Figure 2, we show some examples of Motifs (Figure 2A) and their corresponding Motif masks (Figure 2B). The Motif masks are generated by binary square matrices where only one (with connection) and zero (without connections) values are designed.

$$M_t^{r,l} = \begin{bmatrix} f(m_{1,2}) & \cdots & f(m_{1,r}) \\ \vdots & \ddots & \vdots \\ f(m_{r,1}) & \cdots & f(m_{r,r}) \end{bmatrix}, \tag{4}$$

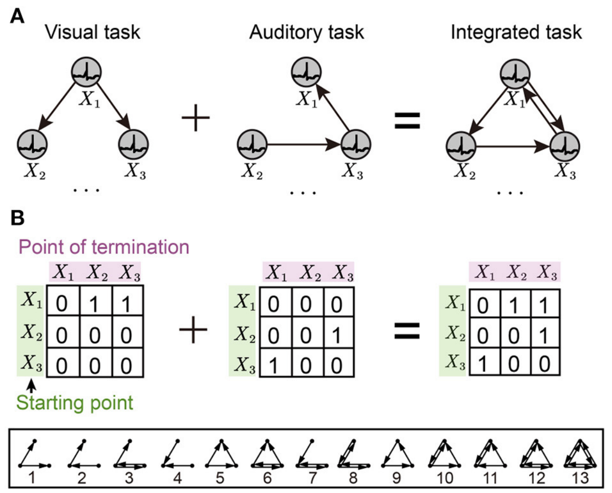
where $f(\cdot)$ is the indicator function. Once the variable in $f(\cdot)$ satisfies the conditions, the function value would be set as one; otherwise, zero. $m_{i,j}, (i,j = 1, \cdots r)$ are elements of synaptic weight $W_t^{r,l}$.

The network motif distribution is calculated by counting the occurrence frequency of network motif types. We enumerate every 3-node assembly (including Motifs and other non-Motif types) and only count the 13-type 3-node connected subgraphs of Motifs with the help of FANMOD (Wernicke and Rasche, 2006). In order to integrate the Motifs learned from different visual and auditory datasets, we propose a multi-sensory integration algorithm by integrating Motif masks with different types learned from visual or auditory classification tasks. Hence, the integrated Motif connections have both visual and auditory network patterns, as shown in Figure 2. Equation (5) shows the integrated equation with visual and auditory Motif masks.

$$M_t^{r,l} = M_t^{r,l}(s) \cup M_t^{r,l}(t), \tag{5}$$

where $M_t^{r,l}(s)$ is the spatial mask that learned from the visual dataset, $M_t^{r,l}(t)$ is the temporal mask that learned from the auditory dataset, and $M_t^{r,l}$ is the integrated mask. "∪" means the OR operation for every element of the visual Motif mask and auditory Motif mask.

For forming the network motifs in SNN, the Motif mask is used to mask the lateral connections in the neural network. The lateral and sparse connections between LIF neurons are usually designed to generate network-scale dynamics. As shown in Figure 1E, we design a four-layer SNN architecture, containing an input layer (for pre-encoding visual and auditory signals to spike trains), a convolutional layer, a multi-sensory integration layer, and a readout layer. The synaptic weights are adaptive while the Motif masks



$$\begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & 0 & 1 & 1 \\ X_2 & 0 & 0 & 0 \\ X_3 & 0 & 0 & 0 \end{array} + \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & 0 & 0 & 0 \\ X_2 & 0 & 0 & 1 \\ X_3 & 1 & 0 & 0 \end{array} = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & 0 & 1 & 1 \\ X_2 & 0 & 0 & 1 \\ X_3 & 1 & 0 & 0 \end{array}$$

FIGURE 2
Schematic diagram of an example for integrating Motif masks. **(A)** Schematic for Motifs of the M-SNN. **(B)** Schematic for Motif masks of the M-SNN.

are not. The membrane potentials in the hidden multi-sensory-integration layer are updated by both feed-forward potential and recurrent potential, shown in the following Equation (6):

$$\begin{cases} S_{i,t} = S_{i,t}^{f} + S_{i,t}^{r} \\ V_{i,t} = V_{i,t}^{f} + V_{i,t}^{r} \\ C\frac{dV_{i,t}^{f}}{dt} = g(V_{i,t} - V_{rest})(1 - S_{i,t}) + \sum_{j=1}^{N} W_{i,j}^{f} X_{j,t} \\ C\frac{dV_{i,t}^{r}}{dt} = \sum_{j=1}^{N} W_{i,j}^{r} S_{i,t} \cdot M_{t}^{r,l} \end{cases}, \tag{6}$$

where C is for capacitance, $S_{i,t}$ is the firing flag of neuron i at time t, $S_{i,t}^{f}$ and $S_{i,t}^{r}$ are the firing flags of neuron i in the feedforward process and recurrent process, respectively, $V_{i,t}$ denotes the membrane potential of neuron i at timing t, which includes feed-forward $V_{i,t}^{f}$ and recurrent $V_{i,t}^{r}$, V_{rest} is the resting potential, $W_{i,j}^{f}$ is the feed-forward synaptic weight from the neuron i to the neuron j, and $W_{i,j}^{r}$ is the recurrent synaptic weight from the neuron i to the neuron j. $M_{t}^{r,l}$ is the mask incorporating Motif topology to further alter feed-forward propagation further. The historical information is saved in the forms of recurrent membrane potential $V_{i,t}^{r}$, where spikes are created after the potential reaches a firing threshold, as illustrated in Equation (7).

$$\begin{cases} V_{i,t}^{f} = V_{reset}, S_{i,t}^{f} = 1 \quad if (V_{i,t}^{f} = V_{th}) \\ V_{i,t}^{r} = V_{reset}, S_{i,t}^{r} = 1 \quad if \left(V_{i,t}^{r} = V_{th}\right) \\ S_{i,t}^{f} = 1 \quad if \left(t - t_{s^f} < \tau_{ref}, t \in (1, T_1)\right) \\ S_{i,t}^{r} = 1 \quad if \left(t - t_{s^r} < \tau_{ref}, t \in (1, T_2)\right) \end{cases}, \tag{7}$$

where $V_{i,t}^{f}$, $V_{i,t}^{r}$, $S_{i,t}^{f}$, and $S_{i,t}^{r}$ are introduced in the previous Equation (6). V_{reset} is the reset membrane potential. τ_{ref} is the refractory period. t_{s^f} is the previous feed-forward spike timing and t_{s^r} is the previous recurrent spike timing. T_1 and T_2 are time windows.

3.3. Neuronal plasticity and learning principle

We use three key mechanisms during network learning: neuronal plasticity, local plasticity, and global plasticity.

Neuronal plasticity emphasizes spatially-temporal information processing by considering the inner neuron dynamic characteristics (Jia et al., 2021), different from traditional synaptic plasticities such as STP and STDP. The neuronal plasticity for SNNs approaches the biological network and improves the learning power of the network. Rather than being a constant value, the firing threshold is set by an ordinary differential equation shown as follows:

$$\frac{da_{i,t}}{dt} = (\alpha - 1)a_{i,t} + \beta(S_t^f + S_t^r), \tag{8}$$

where S_t^f is the input spikes from the feed-forward channel. S_t^r is the input spikes from the recurrent channel. $a_{i,t}$ is the dynamic threshold, which has an equilibrium point of zero without input spikes or $-\frac{\beta}{\alpha-1}$ with input spikes $S^f + S^r$ from the feed-forward and recurrent channels. Therefore, the membrane potential of adaptive LIF neurons is updated as follows:

$$C\frac{dV_{i,t}}{dt} = g\left(V_{i,t} - V_{rest}\right)\left(1 - S_t^f - S_t^r\right) + \sum_{j=1}^{N} W_{i,j} X_{j,t} - \gamma a_{i,t}, \tag{9}$$

where the dynamic threshold $a_{i,t}$ is accumulated during the period from the resetting to the membrane potential firing and finally attains a relatively stable value $a_{i,t}^{*} = \frac{\beta}{1-\alpha}(S_t^f + S_t^r)$. Because of the $-\gamma a_{i,t}$, the maximum firing threshold could reach up to $V_{th} + \gamma a_{i,t}$.

We set $\alpha = 0.9$ to guarantee that the coefficient of $a_{i,t}$ is -0.1, $\beta = 0.1$ to ensure that the spike has the same weight as $a_{i,t}$, and set γ to the common value of 1. Accordingly, the stable $a_t^{*} = 0$ for no input spikes, $a_t^{*} = 1$ for one input spike, and $a_t^{*} = 2$ for input spikes from two channels. When $a_{i,t} < (S_t^f + S_t^r)$, the threshold $a_{i,t}$ will increase, otherwise, the threshold $a_{i,t}$ will decrease. It is clear that the threshold will change in the process of the neuron's firing, and as the firing frequency of the neuron increases, the threshold will also elevate, or vice versa.

For local plasticity, the membrane potential at the firing time is a non-differential spike, so local gradient approximation (pseudo-BP) (Zhang et al., 2021b) is usually used to make the membrane potential differentiable by replacing the non-differential part with a predefined number, shown as follows:

$$Grad_{local} = \frac{\partial S_{i,t}}{\partial V_{i,t}} = \begin{cases} 1 \;\; if \left(\left|V_{i,t} - V_{th}\right| < V_{win}\right) \\ 0 \qquad\quad else \end{cases}, \tag{10}$$

where $Grad_{local}$ is the local gradient of membrane potential at the hidden layer, $S_{i,t}$ is the spike flag of neuron i at time t, $V_{i,t}$ is the membrane potential of neuron i at time t, and V_{th} is the firing threshold. V_{win} is the range of parameters for generating the pseudo-gradient. This approximation makes the membrane potential $V_{i,t}$ differentiable at the spiking time between an upper bound of $V_{th} + V_{win}$ and a lower bound of $V_{th} - V_{win}$.

For global plasticity, we used reward propagation, which has been proposed in our previous work (Zhang et al., 2021b). As shown in Figure 1C, the gradient of the hidden layer in training is generated from the input type-based expectation value and output error-based expectation value by transformed matrix (input type-based expectation matrix and output error-based expectation matrix), respectively, then the gradient signal will be directly given to all hidden neurons without layer-to-layer backpropagation, shown as follows:

$$\begin{cases} Grad_{R_l} = B^{f,l}_{rand} \cdot R_t - h^{f,l} \\ Grad_{R_L} = B^{f,L} \cdot e^{f,L} \\ \Delta W^{f,l}_t = -\eta^f \left(Grad_{R_l}\right) \\ \Delta W^{r,l}_t = -\eta^r \left(Grad_{t+1} + Grad_{R_l}\right) \cdot M^{r,l}_t \\ \Delta W^{f,L}_t = -\eta^f \left(Grad_{R_L}\right) \end{cases}, \qquad (11)$$

where $h^{f,l}$ is the current state of layer l and, R_t is the predefined input-type based expectation value. A predefined random matrix $B^{f,l}_{rand}$ is designed to generate the reward gradient $Grad_{R_l}$. $Grad_{R_L}$ is the gradient of the last layer, $B^{f,L}$ is the predefined identity matrix, and $e^{f,L}$ is the output error. $W^{f,l}_t$ represents the synaptic weight at layer l in feed-forward phase, $\Delta W^{r,l}_t$ is the recurrent-type synaptic modification at layer l which represents defined by both $Grad_{R_l}$ by reward learning and $Grad_{t+1}$ by iterative membrane-potential learning, and the $Grad_{t+1}$ means the gradient obtained at $t + 1$ moment (Werbos, 1990). The $M^{r,l}_t$ is the mask incorporating Motif topology to influence the propagated gradients further.

3.4. The learning procedure of M-SNN

The overall learning procedures of the M-SNN were shown in Algorithm 1, including the raw signal encoding, Motif structure integration, and cognitive effect simulation.

```
1. Initialize the network by resetting weights
and all related parameters. e.g., initial membrane
potential $V_i$, simulation time $T$, learning rates
$\eta = \eta^f = \eta^r$.
2. Encode raw numbers of datasets to spike trains.
3. Learn the synaptic weights $w_{ij}$ and Motif
masks $M^{r,l}_t$ by BP (Rumelhart et al., 1986) in two
single-sensory tasks to get the spatial mask $M^{r,l}_t(s)$
and temporal mask $M^{r,l}_t(t)$.
4. Synthesize Motif masks and train the synaptic
weight $w_{ij}$ on multi-sensory integration tasks.
4.1 Synthesize the integrated masks $M^{r,l}_t$ from
spatial and temporal masks, where $M^{r,l}_t = M^{r,l}_t(s) \cup M^{r,l}_t(t)$.
4.2 Initialize a new network and embed the Motif
mask $M^{r,l}_t$.
4.3 Only learn the synaptic weight $w_{ij}$ with local
Pseudo-BP and global reward learning (Zhang et al.,
2021b).
5. Test the performance of SNNs using these new
masks in the multi-sensory classification tasks
and simulate the cocktail party effect and McGurk
effect.
```

Algorithm 1. The M-SNN algorithm.

4. Experiments

4.1. Visual and auditory datasets

The MNIST dataset (LeCun, 1998) was selected as the visual sensory dataset. The MNIST dataset contains 60,000 28×28 one-channel grayscale images of handwritten digits from zero to nine for training, and there are also 10,000 of the same type of data for testing. The TIDigits dataset (Leonard and Doddington, 1993) was selected as the auditory sensory dataset, containing 4,144 spoken digit recordings from zero to nine. Each recording was sampled at 20 kHz for around one second and then transformed to the frequency domain with 28 frames and 28 bands by the Mel Frequency Cepstral Coefficient (MFCC) (Sahidullah and Saha, 2012). Some examples were shown in Figure 1E.

4.2. Experimental configurations

The SNNs were built in Pytorch, and the network architectures for MNIST and TIDigits were the same, containing one input encoding layer, one convolutional layer (with a kernel size of 5×5, and two input channels constructed by convolutional layer), one full-connection integrated layer (with 200 LIF neurons), and one output layer (with ten output neurons). Among the network, the capacitance C was 1μF/cm^2, conductivity g was 0.2 nS, time constant τ_{ref} was 1 ms, and resting potential V_{rest} was equal to reset potential V_{reset} with 0 mV. The learning rate was 1e-4, the firing threshold V_{th} was 0.5 mV, the simulation time T was set as 20 ms, and the gradient approximation range V_{win} was 0.5 mV.

As shown in Figure 1E, for the visual dataset, before being given to the input layer, the raw data were encoded to spike trains first by comparing each number with a random number generated from Bernoulli sampling at each time slot of the time window T. For the auditory dataset, the input data would first be transformed to the frequency spectrum in the frequency domain by the MFCC (Mel frequency cepstrum coefficient; Sahidullah and Saha, 2012). Then the spectrum would be split according to the time windows. Finally, the sub-spectrum would be converted into normalized value and randomly sampled with Bernoulli sampling to spike trains.

There are two SNNs concluded in our experiment as follows:

- M-SNN. The Motif mask is generated randomly and then updated during the learning of synaptic weights in a Standard-SNN.
- Standard-SNN. The standard feed-forward SNN without Motif masks acts as the control algorithm for comparing M-SNN.

4.3. Analysis of spatial and temporal Motif topology during learning

The visual and auditory Motif masks were shown in Figure 3, which were trained from the MNIST and TIDigits datasets. After

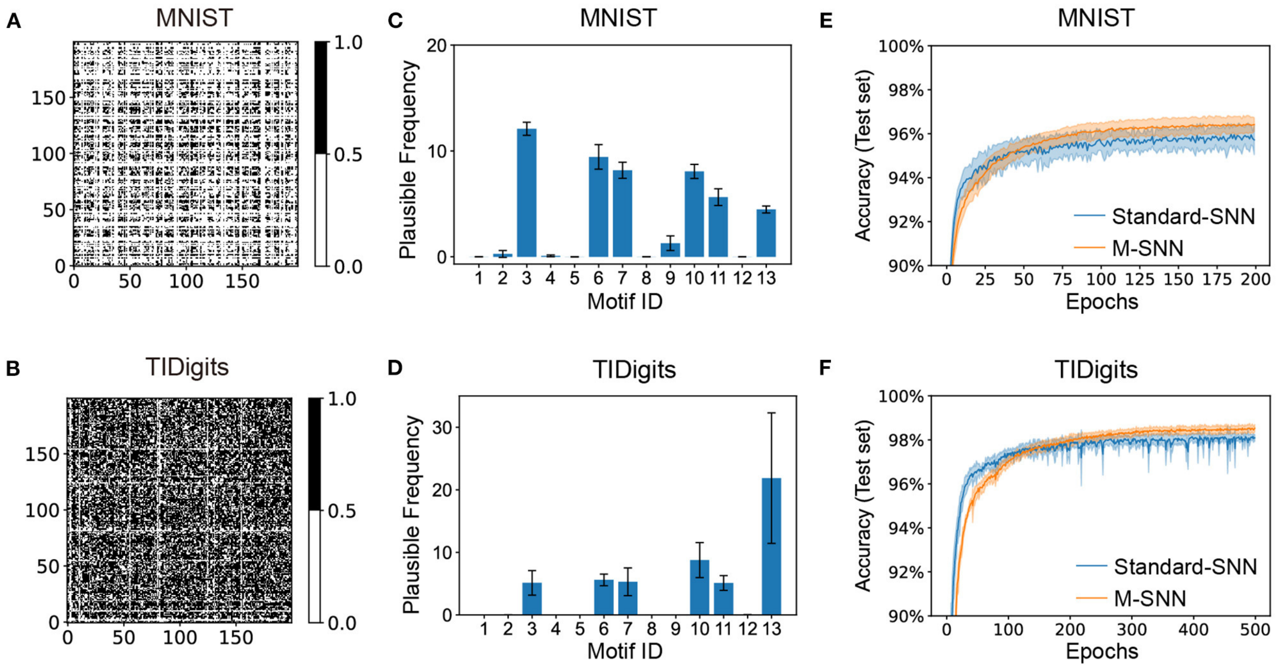


FIGURE 3
Network convergence with SNNs using Motifs in different datasets. **(A, B)** Motif masks of MNIST **(A)** and TIDigits **(B)** after training. **(C, D)** Plausible Frequency of Motif distributions of MNIST **(C)** and TIDigits **(D)** datasets after training. **(E, F)** Convergence curve of classification task of MNIST **(E)** and TIDigits **(F)** datasets.

training, the generated visual and temporal Motif masks were shown in Figures 3A, B, where the black dot in the visualization of the Motif mask indicated that there was a connection between the two neurons shown at the X-axis and Y-axis. The white dot meant there was not.

This result showed that the visual Motif mask connections were sparse, with only about half of the neurons being connected. Furthermore, the connection in the Motif mask is 64.39% for auditory TIDigits dataset, and 28.24% for visual MNIST dataset. For the temporal TIDigits dataset, the generated temporal Motif mask after training was shown in Figure 3B, where the learned Motif mask was denser than that on the visual MNIST in Figure 3A. It is consistent with the biological finding that temporal Motifs are denser than visual ones (Vinje and Gallant, 2000; Hromádka et al., 2008). These differences between spatial and temporal Motif masks indicated that the network needed a more complex connection structure to deal with sequential information. In addition, the connection points in the spatial and temporal Motif masks in Figures 3A, B seemed to be divided into several square regions, similar to the brain regions, which, to some extent, shows the similarity between artificial and biological neural networks at the brain region scale.

The information presented by Motif masks is relatively limited. For further analysis of the Motif structures by Motif distribution, we used the “Plausible Frequency” instead of the standard frequency to calculate the significant Motifs after comparing them to the random networks. The “Plausible Frequency” was defined by multiplying the occurrence frequency and $1 - P$, where the P was the P-value of a selected Motif after comparing it to 2,000 repeating control tasks with random connections. The “repeating control tasks” meant generating many matrixes (e.g., 2000) that each element was sampled from a uniformly random distribution. Furthermore, the P-value index showed the statistical significance of the concerning results, whereas a lower P-value indicated the more plausible result.

The Motif distributions corresponding to the Motif masks were shown in Figures 3C, D, where the spatial and temporal Motifs were distributed differently. For spatial Motifs, the 3rd, 6th, 7th, and 10th units were all prominent in spatial Motifs, while the 13th Motif was the most prominent in temporal Motifs. The abundant 3rd, 6th, 7th, and 10th Motifs in SNN revealed the balance of feedforward and recurrent connections for the spatial tasks. The Motif distribution reveals the difference in the abundance of micro-loops in different networks, indicating that temporal tasks require more complex network connections than spatial tasks. To some extent, the Motif distribution here can mitigate the “black box” problem of ANNs by clearly showing loop-level network differences. The plausible frequency eliminated the interference from the random connection. Figures 3E, F showed that M-SNN networks using Motif topologies can be convergent, where the accuracy of M-SNN was significantly higher than the accuracy of Standard-SNN after a few training epochs.

4.4. M-SNN contribute to solving the cocktail party effect

The cocktail party effect consists of two conditions. The first condition involves focusing on one person’s conversation and

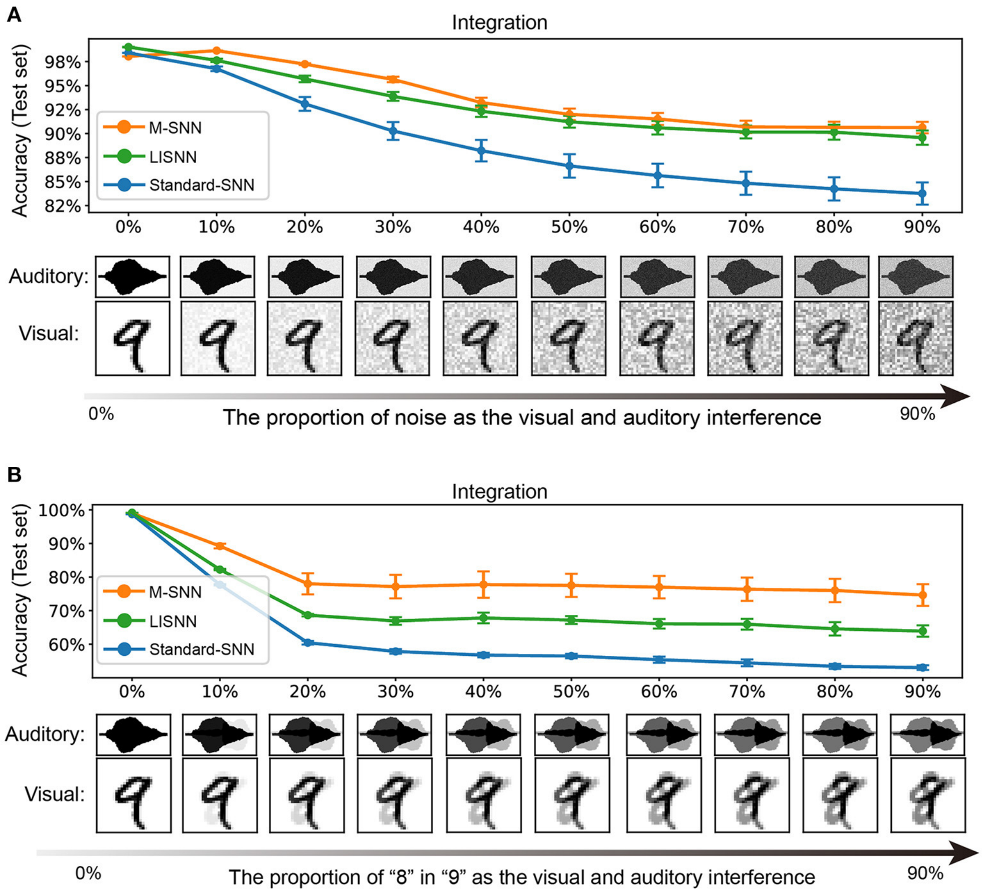


FIGURE 4
Simulation of cocktail party effect. **(A)** A simulation and results in which both visual and auditory inputs have interfered. **(B)** A simulation and results in which only the voice has interfered. All figures are averaged over five repeating experiments with different random seeds.

excluding other conversations or noise in the background. Second, it refers to the response of our hearing organs to a certain stimulus. The human attention mechanism has much to do with how the cocktail party effect happens. In our SNN, we simulated the first situation of the cocktail party effect. We used the MNIST dataset to represent the visual input and the TIDigits dataset for the phonetic input. We modeled two scenes to simulate the simplified cocktail party effect. The first scene was a simulation of the cocktail party effect, where both the visual and auditory inputs were messed up by random noise. The second scene simulated a cocktail party effect in which the visual and auditory inputs were simultaneously disrupted by the real image and voice.

4.4.1. Visual and auditory inputs are interfered with the stochastic noise

In our experiment, we trained the network with pure image and voice inputs and tested the network with input disturbed by stochastic noise. In the simulation process, we used the method of superimposing random numbers between [0, 1] into the image or speech input to simulate the interference effect of noise. With the different values of the added random numbers, different interference effects were formed, ranging from 0 to 90%, and the influence gradually increased. As shown in Figure 4A, when the influence of noise was relatively low, whether to adding Motifs into the network had little effect on the experimental results (99.00 ± 0.00% for the network with Motifs, 98.50 ± 0.22% for Standard-SNN, and 99.14 ± 0.03% for LISNN; Cheng et al., 2020). As shown in Figure 4A, with the increase of noise ratio, the recognition ability of the network to the input target signal decreased gradually. When the proportion of noise was increased to 60%, the accuracy of the M-SNN was 95.64 ± 0.29%, which was markedly higher than the accuracy of Standard-SNN (57.84 ± 0.68%) and was comparable with LISNN (93.88 ± 0.46%). The higher accuracy indicated that the Motifs in M-SNN had a positive effect on solving the cocktail party effect compared with Standard-SNN. Furthermore, LISNN with lateral interaction in the convolution layer could get a comparable effect with M-SNN.

4.4.2. Visual and auditory inputs are interfered with the real image and voice

We used the MNIST and TIDigits datasets without noise when training the network. We used "8" from the handwritten digital image and human voice in the simulation process instead of the stochastic noise as interference. As shown in Figure 4B, in the

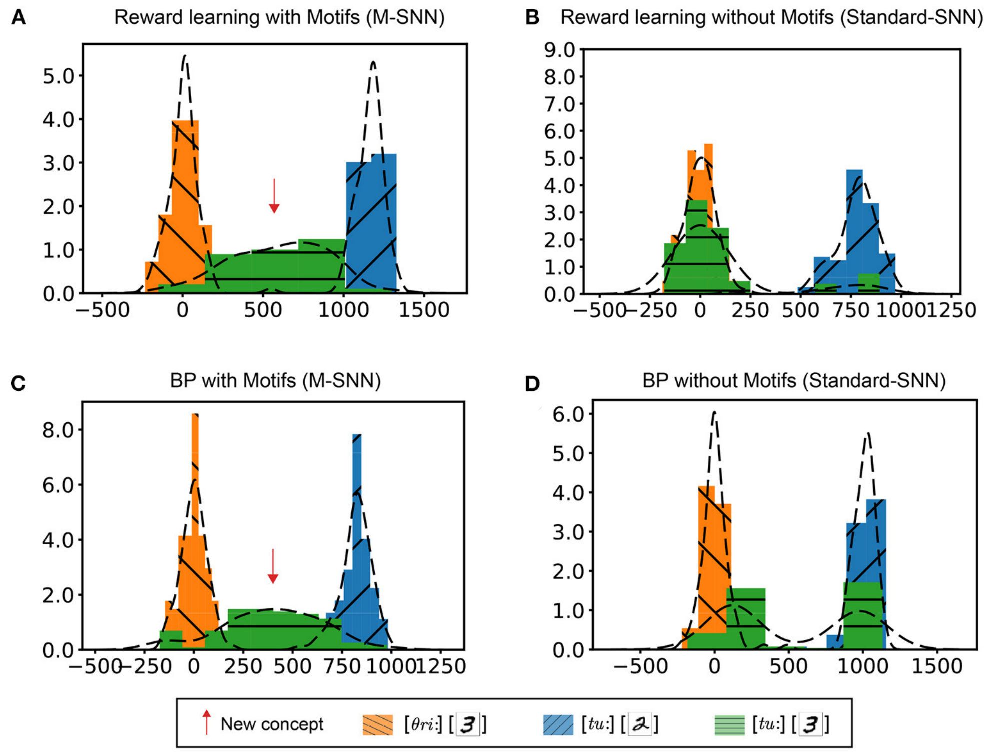


FIGURE 5

Simulation of McGurk effect. **(A, B)** Distribution in the integrated layer after reward learning with Motifs **(A)** and without Motifs **(B)** of different combinations of input. **(C, D)** Distribution in integrated layer after BP learning with Motifs **(C)** and without Motifs **(D)** of different combinations of input.

case of a few other interfering sounds, the effect of M-SNN on maintaining accuracy was insignificant. However, with the increase in the proportion of different interfering sounds, the impact of M-SNN on maintaining the recognition of the network was becoming more and more significant. When the noise ratio reached 50%, the recognition accuracy of M-SNN became 77.77 ± 3.94 %, while the Standard-SNN could only reach the an accuracy of 56.75 ± 0.67%, and the accuracy achieved by LISNN was 67.83 ± 1.58%. In these situations, the maximal increased accuracy was 7.5% when the proportion of "8" was 50%.

4.5. M-SNN for explainable McGurk effect

The McGurk effect described the psychological phenomenon that occurs when human speech input and image input are inconsistent, whereby most people would judge the input as neither a speech label nor a visual label but a novel concept. It had been shown that, for adults, the error rate in judging inconsistent audio-visual input as novel concepts was more than 90% (McGurk and MacDonald, 1976). For example, when the speech input was [ba] and the visual input was [ga], a new concept [da] was generated (Tiippana, 2014). During the simulation, we used handwritten digit images [2],[3] as the visual input, while speech digits [*tu*:],[*θri*:] were used to represent the corresponding pronunciation.

First, consistent audio-visual inputs were used to train the network weights. After training, the inconsistent audio-visual information would be fed into the network. In the integrated layer, we used TSNE (Maaten and Hinton, 2008) to reduce the dimension of the high-dimensional features. We conducted four experiments to verify the influence of learning rules and structures on the McGurk effect simulation: networks trained with reward learning with Motif (Figure 5A), networks trained with reward learning without Motif (Figure 5B), networks trained with BP learning with Motif (Figure 5C), and networks trained with BP without Motif (Figure 5D). As shown in Figure 5, the histogram showed the distribution of samples with different labels in the integration layer. The *x*-axis represents the distance between the feature point and the reference point on the 2D plane (using TSNE for clustering). For the Standard-SNN, there were two prominent feature distributions: [*θri*:,3] and [*tu*:,2]. However, for the learning results of M-SNNs, a clear feature distribution of [*tu*:,3] emerged between the distributions of [*θri*:,3] and [*tu*:,2]). This distribution corresponding to [*tu*:,3] characterized the new concept (McGurk effect). These results showed that Motifs in SNNs are important for generating the McGurk effect, and neither of these learning principles alone can produce the McGurk effect.

For comparing the stimulating effect of the McGurk effect, we compared additional algorithms as shown in Table 1. According to our knowledge, the SOM approach in the paper (Gustafsson et al., 2014) is the only unsupervised learning method that replicates the McGurk effect. In contrast, our M-SNN is the only supervised learning method.

4.6. Lower computational cost for M-SNN during training

We referred to the method in paper (Zhang et al., 2021a) to calculate the computational cost of the network during training for algorithm i, (i=1, 2), where the average training cost of the network was represented by the average epoch multiplied by the number of parameters of the network. A schematic for the mean epoch was shown in Figure 6A, and the equation was shown as follows:

$$Cost_i = \frac{1}{N}\sum_{l=1}^{N} \text{Argmin}_i\left(f_i(x) = Acc_l\right) \times O(n)_i, \qquad (12)$$

where $\text{Argmin}_i(\cdot)$ is the argument when $\cdot$ is the minimum, $f_i(x)$ is the accuracy function of training epoch x, Acc_l is the selected accuracy in $[f_1(x), f_2(x)]$, $O(n)_i$ is the algorithmic complexity of algorithm i, and N is the number of repetitions. The upper bound is Min[Max[$f_1(x)$, $f_2(x)$]] and the lower bound is Max[Min[$f_1(x)$, $f_2(x)$]], where Max and Min represents the maximum and the minimum, respectively. In our experiment, $N = 5$ and for the network with m, n, k input, hidden, and output neurons, respectively, the $O(n)$ of M-SNN is $(m \times n + n \times n + n \times k)$ and the $O(n)$ of Standard-SNN is $(m \times n + n \times k)$.

TABLE 1 Performance of different algorithms on simulating McGurk effect ("+" indicates that such a correspondence exists, while "−" indicates not).

	Standard -SNN	LISNN (Cheng et al., 2020)	SOM (Gustafsson et al., 2014)	M -SNN
Simulated McGurk	−	−	+	+
Supervised learning	+	+	−	+

We calculated the computational cost of training for different proportions of noise. The results of M-SNN and Standard-SNN computational costs were shown in Figure 6B, indicating that the increased noise ratio brought a higher computational cost to the network. In addition, the result showed that the Motifs in M-SNN could save on computational cost when network training (the training cost convergence curves of M-SNN was always below the convergence curves of Standard-SNN). When the noise ratio was 10%, M-SNN achieved the maximum cost-saving ratio of 72.6%. M-SNN achieved the most significant absolute cost savings (save 4.1 × 10^7) when the noise ratio reached 30%.

5. Conclusion

In this paper, we propose a model of Motif-topology improved SNN (M-SNN), exhibiting three main important features. First, M-SNN could improve recognition accuracy in multi-sensory integration tasks. Second, M-SNN could better simulate the cocktail party and McGurk effects than Standard-SNN. Compared with the common Standard-SNN and other SNN methods, M-SNN had a better function of filtering noise from other speakers in different proportions. Furthermore, compared with SNN without Motifs, M-SNN could better handle the McGurk effect with auditory and visual Motif topologies and visual ones. Third, compared with Standard-SNN, M-SNN has a lower computational cost during training in different noise ratios of the background, and the maximum computational cost-saving ratio is 72.6%.

A more profound analysis of the Motifs helps us understand more about the critical functions of the structures in SNNs. This inspiration from Motifs describes the sparse connection in the cell assembly that reveal the importance of the micro-scale structures. Motif topologies are patterns for describing the topologies of a system (e.g., biological cognitive pathways), including the n-node meta graphs that uncover the bottom features of the networks. We find that biological Motifs are beneficial for improving the accuracy

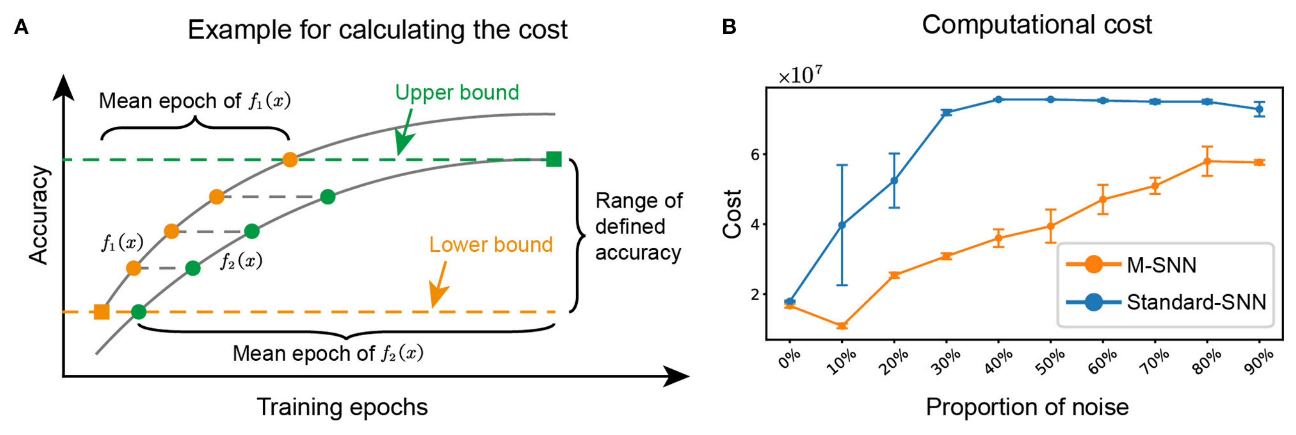


FIGURE 6
M-SNN in training for lower computational cost. **(A)** Schematic diagram depicting how to calculate the mean epoch during training. **(B)** The computational cost of network training under different proportions of noise.

of networks in visual and auditory data classification. Significantly, the 3-node Motifs are typical and concise, which could assist in analyzing the function of different network modules.

The research on the variability of Motifs will give us more ideas and inspiration toward buildings for a better network. The simulation of different cognitive functions by SNNs with biologically plausible Motifs has much in store to offer in future.

Code availability statement

The source code can be downloaded from https://github.com/thomasaimondy/Motif-SNN after the acceptance of the paper.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

TZ and BX came up with the idea. TZ, SJ, and RZ made the mathematical analyses and experiments. All authors wrote the paper together. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA27010404 and XDB32070100), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX), and the Youth Innovation Promotion Association CAS.

Acknowledgments

The authors would like to thank Duzhen Zhang for his previous assistance with the discussion.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abraham, W. C., and Bear, M. F. (1996). Metaplasticity: the plasticity of synaptic plasticity. *Trends Neurosci.* 19, 126–130. doi: 10.1016/S0166-2236(96)80018-X

Aladago, M. M., and Torresani, L. (2021). "Slot machines: discovering winning combinations of random weights in neural networks," in *ICML* (Virtual Event).

Beniaguev, D., Segev, I., and London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron* 109, 2727–2739.e3. doi: 10.1016/j.neuron.2021.07.002

Chao, G.-L., Chan, W., and Lane, I. (2019). Speaker-targeted audio-visual models for speech recognition in cocktail-party environments. *arXiv [Preprint]. arXiv:1906.05962*. doi: 10.48550/arXiv.1906.05962

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Carbin, M., et al. (2020). "The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA), 16301–16311. doi: 10.1109/CVPR46437.2021.01604

Cheng, X., Hao, Y., Xu, J., and Xu, B. (2020). "LisNN: improving spiking neural networks with lateral interactions for robust object recognition," in *IJCAI* (Yokohama), 1519–1525. doi: 10.24963/ijcai.2020/211

Dellaferrera, G., Woźniak, S., Indiveri, G., Pantazi, A., and Eleftheriou, E. (2022). Introducing principles of synaptic integration in the optimization of deep neural networks. *Nat. Commun.* 13:1885. doi: 10.1038/s41467-022-29491-2

Diehl, P. U., and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* 9:99. doi: 10.3389/fncom.2015.00099

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. (2015). "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *The 2015 International Joint Conference on Neural Networks (IJCNN-2015)* (Killarney), 1–8. doi: 10.1109/IJCNN.2015.7280696

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *CoRR, abs/1804.03619*. doi: 10.1145/3197517.3201357

Frankle, J., and Carbin, M. (2018). The lottery ticket hypothesis: finding sparse, trainable neural networks. *arXiv [Preprint]. arXiv:1803.03635*. doi: 10.48550/arXiv.1803.03635

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. (2019). Linear mode connectivity and the lottery ticket hypothesis. *arXiv [Preprint]. arXiv: abs/1912.05671*. doi: 10.48550/arXiv.1912.05671

Gaier, A., and Ha, D. (2019). "Weight agnostic neural networks," in *Advances in Neural Information Processing Systems* (Vancouver), 32.

Gustafsson, L., Jantvik, T., and Paplinski, A. P. (2014). "A self-organized artificial neural network architecture that generates the McGurk effect," in *2014 International Joint Conference on Neural Networks (IJCNN)* (Beijing), 3974–3980. doi: 10.1109/IJCNN.2014.6889411

Hao, Y., Xu, J., Zhang, P., and Xu, B. (2021). "Wase: learning when to attend for speaker extraction in cocktail party environments," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, ON), 6104–6108. doi: 10.1109/ICASSP39728.2021.9413411

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., and Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-018-30798-8

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol.* 6:e16. doi: 10.1371/journal.pbio.0060016

Hu, Q., Lin, W., Tang, M., and Jiang, J. (2022). Mbhan: motif-based heterogeneous graph attention network. *Appl. Sci.* 12:5931. doi: 10.3390/app12125931

Jia, S., Zhang, T., Cheng, X., Liu, H., and Xu, B. (2021). Neuronal-plasticity and reward-propagation improved recurrent spiking neural networks. *Front. Neurosci.* 15:654786. doi: 10.3389/fnins.2021.654786

Jia, S., Zuo, R., Zhang, T., Liu, H., and Xu, B. (2022). "Motif-topology and reward-learning improved spiking neural network for efficient multi-sensory integration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Virtual Event; Singapore), 8917–8921. doi: 10.1109/ICASSP43922.2022.9746157

LeCun, Y. (1998). *The Mnist Database of Handwritten Digits*. Available online at: http://yann. lecun. com/exdb/mnist/

Lee, J. H., Delbruck, T., and Pfeiffer, M. (2016). Training deep spiking neural networks using backpropagation. *Front. Neurosci.* 10:508. doi: 10.3389/fnins.2016.00508

Leonard, R. G., and Doddington, G. (1993). *Tidigits ldc93s10*. Philadelphia, PA: Linguistic Data Consortium.

Liu, H., Simonyan, K., and Yang, Y. (2018). Darts: differentiable architecture search. *arXiv [Preprint]. arXiv:1806.09055*. doi: 10.48550/arXiv.1806.09055

Luo, L. (2021). Architectures of neuronal circuits. *Science* 373:eabg7285. doi: 10.1126/science.abg7285

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

Maaten, L. v. d., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. doi: 10.48550/arXiv.2108.01301

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824–827. doi: 10.1126/science.298.5594.824

Prill, R., Iglesias, P., and Levchenko, A. (2005). Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol.* 3:e30343. doi: 10.1371/journal.pbio.0030343

Rideaux, R., Storrs, K. R., Maiello, G., and Welchman, A. E. (2021). How multisensory neurons solve causal inference. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2106235118. doi: 10.1073/pnas.2106235118

Rueckert, E., Kappel, D., Tanneberg, D., Pecevski, D., and Peters, J. (2016). Recurrent spiking networks solve planning tasks. *Sci. Rep.* 6:21142. doi: 10.1038/srep21142

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sahidullah, M., and Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Commun.* 54, 543–565. doi: 10.1016/j.specom.2011.11.004

Shen, K., Bezgin, G., Hutchison, R. M., Gati, J. S., and Mcintosh, A. R. (2012). Information processing architecture of functionally defined clusters in the macaque cortex. *J. Neurosci.* 32, 17465–17476. doi: 10.1523/JNEUROSCI.2709-12.2012

Shrestha, S. B., and Orchard, G. (2018). "Slayer: spike layer error reassignment in time," in *Advances in Neural Information Processing Systems* (Montréal), 31.

Soltani, A., and Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. *Nat. Neurosci.* 13, 112–119. doi: 10.1038/nn.2450

Sun, P., Zhu, L., and Botteldooren, D. (2022). "Axonal delay as a short-term memory for feed forward deep spiking neural networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Virtual Event; Singapore), 8932–8936. doi: 10.1109/ICASSP43922.2022.9747411

Tiippana, K. (2014). What is the McGurk effect? *Front. Psychol.* 5:725. doi: 10.3389/fpsyg.2014.00725

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wang, J., Lam, M. W., Su, D., and Yu, D. (2021). "Tune-in: training under negative environments with interference for attention networks simulating cocktail party effect," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35* (Virtual Event), 13961–13969. doi: 10.1609/aaai.v35i16.17644

Wang, Q., Zhang, T., Han, M., Wang, Y., and Xu, B. (2023). "Complex dynamic neurons improved spiking transformer network for efficient automatic speech recognition," in *Thirty-Seventh AAAI Conference on Artificial Intelligence* (Virtual Event).

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proc. IEEE.* 78, 1550–1560. doi: 10.1109/5.58337

Wernicke, S., and Rasche, F. (2006). Fanmod: a tool for fast network motif detection. *Bioinformatics* 22, 1152–1153. doi: 10.1093/bioinformatics/btl038

Wu, J., Chua, Y., Zhang, M., Li, G., Li, H., and Tan, K. C. (2021). A tandem learning rule for effective training and rapid inference of deep spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst* 34, 446–460. doi: 10.1109/TNNLS.2021.3095724

Zeng, Y., Zhang, T., and Xu, B. (2017). Improving multi-layer spiking neural networks by incorporating brain-inspired rules. *Sci. China Inform. Sci.* 60:052201. doi: 10.1007/s11432-016-0439-4

Zhang, D., Zhang, T., Jia, S., and Xu, B. (2021). "Multiscale dynamic coding improved spiking actor network for reinforcement learning," in *Thirty-Sixth AAAI Conference on Artificial Intelligence* (Virtual Event). doi: 10.1609/aaai.v36i1.19879

Zhang, M., Wang, J., Wu, J., Belatreche, A., Amornpaisannon, B., Zhang, Z., et al. (2021). Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 1947–1958. doi: 10.1109/TNNLS.2021.3110991

Zhang, T., Cheng, X., Jia, S., Poo, M. M., Zeng, Y., and Xu, B. (2021a). Self-backpropagation of synaptic modifications elevates the efficiency of spiking and artificial neural networks. *Sci. Adv.* 7:eabh0146. doi: 10.1126/sciadv.abh0146

Zhang, T., Jia, S., Cheng, X., and Xu, B. (2021b). Tuning convolutional spiking neural network with biologically plausible reward propagation. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 7621–7631. doi: 10.1109/TNNLS.2021.3085966

Zhang, T., Zeng, Y., and Xu, B. (2017). A computational approach towards the microscale mouse brain connectome from the mesoscale. *J. Integr. Neurosci.* 16, 291–306. doi: 10.3233/JIN-170019

Zhang, T., Zeng, Y., Zhao, D., and Shi, M. (2018a). "A plasticity-centric approach to train the non-differential spiking neural networks," in *The 32th AAAI Conference on Artificial Intelligence (AAAI-2018)* (Virtual Event). doi: 10.1609/aaai.v32i1.11317

Zhang, T., Zeng, Y., Zhao, D., and Xu, B. (2018b). "Brain-inspired balanced tuning for spiking neural networks," in *IJCAI* (Stockholm), 1653–1659. doi: 10.24963/ijcai.2018/229