Neuromorphic Speech Recognition With Photonic Convolutional Spiking Neural Networks

Shuiying Xiang, Tianrui Zhang, Yanan Han, Xingxing Guo, Yahui Zhang, Yuechun Shi, and Yue Hao, *Senior Member, IEEE*

***Abstract*—Spiking neural network (SNN) have attracted lots of attention due to its event-driven nature and powerful computation capability. However, it is still limited to simple task due to the training difficulty. In this work, we propose a hybrid architecture of photonic convolutional spiking neural network (PCSNN) to realize the speech recognition task. In the PCSNN, the feature extraction is realized by a convolution SNN with unsupervised learning algorithm, the classification is realized by a photonic SNN with modified time-based supervised training algorithm. The TIDIGITS dataset is used to test the speech recognition performance of the proposed PCSNN, and the highest testing accuracy is 93.75%. The proposed PCSNN provides a solution for architecture and algorithm co-design for the speech recognition task, which is helpful for extending the applications of photonic SNN.**

***Index Terms*—Photonic convolutional spiking neural network, speech recognition, supervised learning.**

I. Introduction

With the rapid development of spiking neural network (SNN), the electronic hardware is no longer the only platform to run a SNN, the photonic neuromorphic platform becomes a promising candidate which is compatible with SNN [1]. Compared with the electronic SNN, the photonic SNN exhibits the advantages of low power consumption, high speed and low latency, which can overcome the shortcomings of electronic neuromorphic computing systems [2]. Thus, the photonic SNN is promising to realize the ultra-fast neuromorphic computing and information processing.

From the point of hardware perspective, in photonic neuromorphic systems, both synapses and spiking neurons are realized by photonic devices. On the one hand, the photonic spiking neurons based on vertical-cavity surface-emitting laser (VCSEL) with and without a saturable absorber have been demonstrated extensively [3], [4], [5], [6], [7], [8], [9]. Besides, there are many emerging photonic spiking neurons based on silicon micro-ring resonator (MRR) [10], [11], two-section Fabry-Peŕot laser [12], [13], two section distribution feedback laser [14], phase change material hybrid integrated with MRR [15], etc. On the other hand, the mainstream photonic synapses include the Mach-Zehnder Interferometer (MZIs) network [16], [17], [18], [19] and MRR network [20], [21], [22] which can perform vector-matrix multiplication.

From the point of algorithm perspective, the unsupervised and supervised learning algorithms for SNN have been studied [23], [24], [25], but the photonic SNN generally limited to solve relatively simple tasks such as spike sequence recognition [8], [9], XOR problem [25], optical character recognition [9], and Iris classification [26]. As far as we know, the photonic SNN has not been used to realize the speech recognition, which is an important practical tasks in the modern artificial intelligence era.

For speech recognition task, the approaches based on SNN can be divided into two categories. For the first type of solution [27], [28], [29], [30], [31], [32], [33], [34], the speech signal is converted to spectrogram. Each point on the spectrogram represents the speech energy of the speech signal at the corresponding time and frequency. According to the characteristics of energy distribution, the features can be extracted directly from the spectrogram, and then the extracted features are sent to the network as input layer for training, so as to realize the classification and recognition task. For the second type of solution [35], [36], [37], the acquired spectrogram is considered as an image. The convolution spiking neural network (CSNN) is used to extract the edge features of the image, and then the output of the fully connected layer of the CSNN is connected to the SNN or other classification networks for supervised training.

In this work, we propose a hybrid neural network architecture that combines the CSNN and photonic SNN, which we refer to as photonic convolutional SNN (PCSNN), and propose a

Manuscript received 15 December 2022; revised 18 January 2023; accepted 24 January 2023. Date of publication 27 January 2023; date of current version 8 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grants 2021YFB2801900, 2021YFB2801901, 2021YFB2801902, 2021YFB2801903, and 2021YFB2801904, in part by the National Outstanding Youth Science Fund Project of National Natural Science Foundation of China under Grant 62022062, in part by the National Natural Science Foundation of China under Grants 61974177 and 61674119, and in part by the Fundamental Research Funds for the Central Universities under Grant JB210114. *(Corresponding author: Shuiying Xiang.)*

Shuiying Xiang is with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China, and also with the State Key Discipline Laboratory of Wide Bandgap Semiconductor Technology, School of Microelectronics, Xidian University, Xi'an 710071, China (e-mail: syxiang@xidian.edu.cn).

Tianrui Zhang, Yanan Han, Xingxing Guo, and Yahui Zhang are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: trzhang98@gmail.com; 1356078782@qq.com; xidiangxx@126.com; 18332551054@163.com).

Yuechun Shi is with the Yongjiang Laboratory, Ningbo 315202, China (e-mail: yuechun-shi@ylab.ac.cn).

Yue Hao is with the State Key Discipline Laboratory of Wide Bandgap Semiconductor Technology, School of Microelectronics, Xidian University, Xi'an 710071, China (e-mail: yhao@xidian.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2023.3240248.

Digital Object Identifier 10.1109/JSTQE.2023.3240248

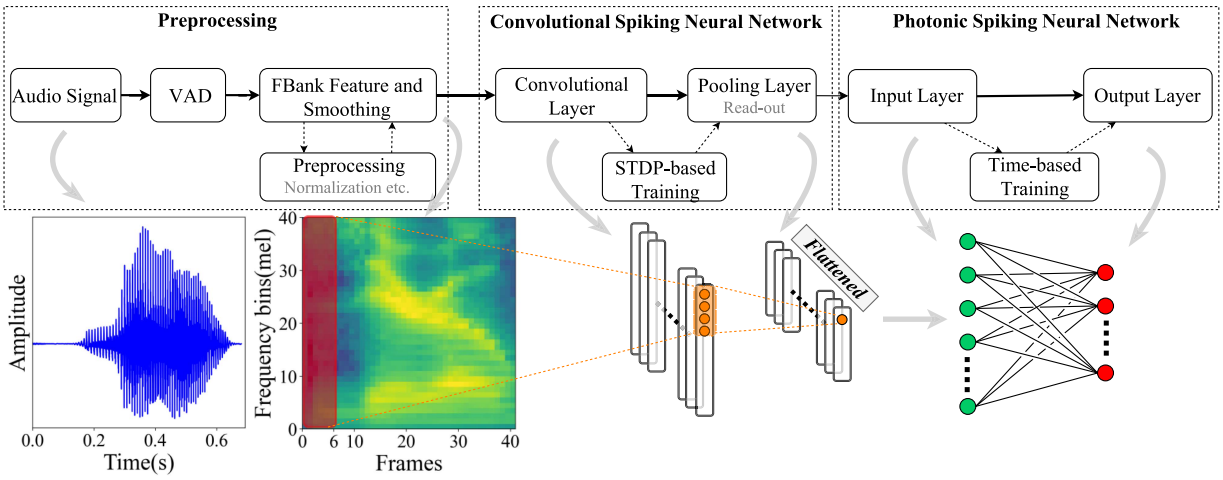

Fig. 1. The overall structure of the neuromorphic speech recognition network. It includes three functional parts, preprocessing, CSNN and photonic SNN. The CSNN is employed to achieve the feature extraction, and the photonic SNN is utilized to realize classification.

supervised learning algorithm modified from the STDP rule. More specifically, unsupervised feature extraction algorithm and supervised recognition algorithm are adopted [38]. In the speech signal feature extraction part, the CSNN [39] and [40] is used to extract the edge features of speech spectrograms in an unsupervised manner. In the photonic SNN, the proposed supervised learning algorithm is used to train and effectively realize the classification and recognition of speech signals. With the co-design of architecture and algorithm, we extend the photonic SNN to realize the speech recognition task successfully. The rest of the paper is organized as follows: in Section II, the overall architecture of the hybrid neuromorphic speech recognition network is presented and each functional part is described in detail. In Section III, the results for the speech recognition based on the PCSNN are presented. The hardware constraints with limited neuron nodes are also considered. Conclusions are provided in Section IV.

II. Network Structure and Principle

The overall structure of the neuromorphic speech recognition network is presented in Fig. 1. It includes three functional parts, preprocessing, CSNN and photonic SNN. For an audio signal, the first step is to perform the voice activity detection (VAD) to reduce the effect of background noise. Then, short-time Fourier transform is performed to obtain the speech spectrogram represented by the FBank feature. Meanwhile, smoothing processing is adopted to reduce the noise points. Then, the layer normalization is introduced to process the FBank feature, and the feature value is converted inversely proportional to the firing time of the integrator-and-fire neuron. In the CSNN, the STDP algorithm and correlation suppression strategy is used to train the network to achieve high-dimensional feature map. In the photonic SNN, the time-based supervised training algorithm is employed to classify the feature map. In the following, the network structure, principle and implementation details of three functional parts will be described in detail.

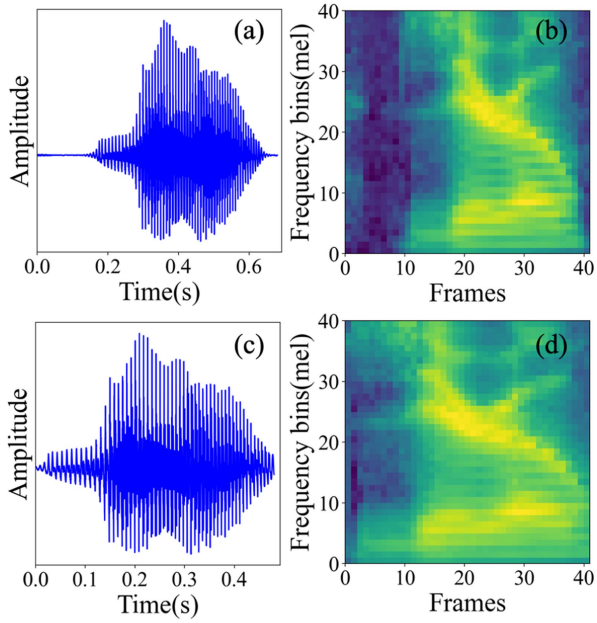

Fig. 2. Time series (left column) and spectrogram (right column) of speech signal. (a) and (b) correspond to the original speech signal, (c) and (d) correspond to the speech signal after the endpoint detection.

A. Preprocessing

To begin with, the dual-threshold endpoint detection algorithm is employed to preprocess the given speech signal as presented in Fig. 2(a). The speech signal is sampled and framed to obtain short-time energy and short-term average zero-crossing rate. For the short-time energy, a higher threshold is set to determine the beginning of speech segment, and a lower threshold is set to exclude the background noise. For short-time average zero-crossing rate, only a threshold is set in this paper. After judging by the threshold of short-time average, the threshold of short-time average zero-crossing rate is further used to ensure that no frame with low energy but high average zero-crossing rate is missed. In this case, the frame is considered to belong to the

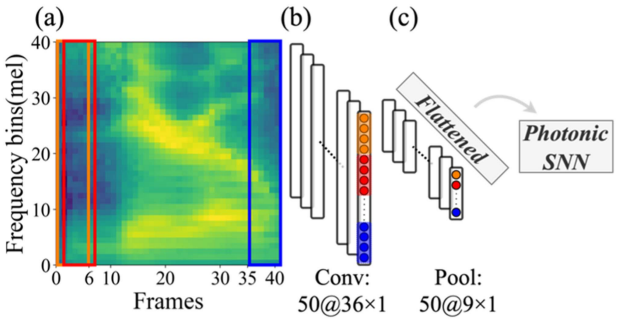

Fig. 3. The schematic diagram of the CSNN structure. (a) Spectrogram obtained in the preprocessing part, each feature value represents the firing time of the IF neuron at the corresponding position of the input layer. (b) Convolutional layer, the number of feature maps is 50, and the size of each feature map is 36×1. (c) Pooling layer, the number of feature maps is 50, and the size of each feature map is 9×1.

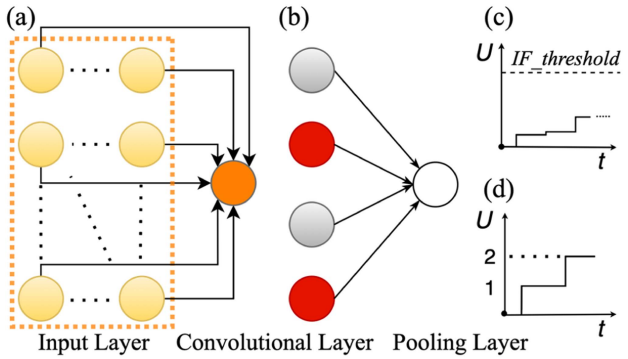

Fig. 4. The schematic diagram of connection (a) and (c) between the input layer and convolution layer, (b) and (d) between the convolution layer and pooling layer.

speech segment. The speech signal after the endpoint detection is presented in Fig. 2(c)

For the speech signal after the endpoint detection, the pre-emphasis, framing, time-domain windowing, and short-time Fourier transform are performed, and then the Mel-filter bank is used to obtain the FBank based on the Mel-scale [41]. Correspondingly, the spectrograms represented by the FBank for the original speech signal and the speech signal after the endpoint detection are displayed in Fig. 2(b) and (d).

B. CSNN

For the extracted FBank spectrogram, smoothing is performed, and min-max scaling is further employed to normalize the feature value to the range of [0, 1]. According to the time-to-first-spike encoding strategy, the feature map is inversely converted to the firing time of the corresponding neuron,

$$t = [1 - \sin(FBank \times \frac{\pi}{2})] \times T \tag{1}$$

Where T represents the simulation time window of the CSNN. Obviously, a larger FBank leads to a smaller firing time t.

In a conventional CNN, the image is usually local correlation, and the convolutional kernel with a small square matrix is adopted to extract the basic feature, while the deep structure is employed to further extract the high-level feature. But the FBank spectrogram is harmonic correlation, its correlation is along the whole frequency axis. Thus, similar to Refs. [36], [42], the convolutional kernel is modified accordingly to ensure that all frequency bands are included on the frequency axis, but the local connection property is still retained on the time axis. The size of the convolution kernel can be expressed as $(\Delta t, f) = (6, 40)$.

Note, in the time domain, adjacent speech signals have strong correlation [36], [42], [43], and its feature values are not randomly distributed like the features of conventional images, but concentrated in adjacent time periods. Thus, local weight sharing is considered. A feature map is divided into multiple local regions on the time axis, and the corresponding weights of different local regions are also different. As presented in Fig. 3(a) and (b), there are 41 frames in the time domain, the time duration of convolution kernel is 6, the step of sliding window is 1, thus, the total convolution operations is 36. Here, 4 convolution operations are considered as a local region, which share the same convolution kernel. Thus, 9 different convolution kernel are adopted here for a spectrogram. As indicated by the orange, blue, red rectangle in Fig. 3(a), the different colors correspond to different convolution kernel. For each feature map, there are 36 rows. Note, to obtain more feature maps, we consider 50 sets of convolution kernels, each set consists of 9 different convolution kernels. Thus there are 50 columns of feature maps in Fig. 3(b). Thus, in the convolution layer, the number of feature maps is 50, and the size of each feature map is 36.

As presented in Fig. 4(a) and (c), the input layer includes 40×6 IF neurons corresponding to the size of convolutional kernel. For the IF neuron, the membrane voltage can be expressed as

$$V(t) = \sum_{i \in [1,240]} \sum_{t_i < t} w_i \times s_i \tag{2}$$

where s_i and w_i represent the amplitude of spike generated at t_i and weight for the i-th IF neuron. Here, we suppose that each IF neuron can only fire a single spike. The firing threshold is set as $IF_{threshold} = 33$, which is determined by optimizing the test accuracy. When the membrane voltage of the convolutional layer IF neuron reaches the threshold, we record the firing time and membrane voltage as follows,

$$t_{conv} = t \text{and} v_{conv} = V(t), when V(t) \geq IF_{threshold} \tag{3}$$

For simplicity, we consider s_i=1, then the (2) can be simplifed as

$$V(t) = \sum_{i \in [1,240]} \sum_{t_i < t} w_i \tag{4}$$

For the connection between the convolution layer and pooling layer, the size of convolution kernel is 4×1. Thus, in the pooling layer, the number of feature maps is 50, and the size of each feature map is 9. Here, the weight is set as 1. Here, multipl spikes are allowed for the IF neuron in the pooling layer. As presented in Fig. 4(b) and (d), the red neuron indicates firing state while the gray neuron indicates rest state. Thus, two IF neurons in the convolution layer fire a single spike and each is

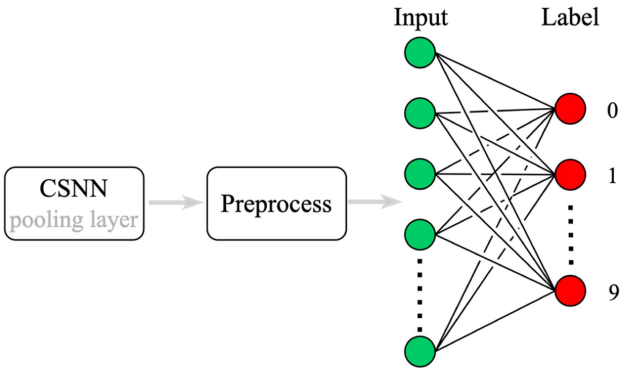

Fig. 5. The schematic diagram of photonic SNN. The input layer includes 450 neurons. The output layer includes 10 neurons to classify the speech class. Each neuron is a photonic spiking neuron based on VCSEL-SA.

transmitted to the pooling layer. The membrane voltage of IF neuron in the pooling layer is 2, which corresponds to 2 spikes. The pooling layer is flattened to an array whose size is 50×9. The number of firing spikes is then transmitted to the photonic SNN.

Note, in the convolution layer, inhibition strategy is introduced to extract more obvious feature. For each row, only one IF neuron is allowed to fire a spike, the others are inhibited. We consider the firing inhibition and update inhibition as follows. For the firing inhibition, the IF neuron with smaller t_{conv} is the winner. If multiple neurons have the same t_{conv}, the neuron with larger v_{conv} is the winner. For the update inhibition, during the training process, if multiple adjacent neurons in the same column fire spikes, we only choose the winner neuron and update its corresponding weight. Here, the range of adjacent positions is set as $\Delta t/2$. The firing inhibition ensures that different feature maps recognize different features. The aim of update inhibition is to effectively distinguish the corresponding features in different time periods when local weight sharing is applied.

The algorithm adopted to update the weight is based on a simpified spike-timing dependent plasticity (STDP) rule as follows [44],

$$\Delta\omega_{ij} = \begin{cases} \alpha^{+}\omega_{ij}\left(1-\omega_{ij}\right), & \text{if } t_i < t_j \\ \alpha^{-}\omega_{ij}\left(1-\omega_{ij}\right), & \text{else} \end{cases} \tag{5}$$

where t_i represents the firing time of the *i*-th neuron in the input layer, t_j denotes the firng time of *j*-th neuron in the convolution layer neuron. $\alpha^{+}= 0.004$ and $\alpha^{-}= -0.003$ represent the positive and negative learning rate, respectively.ω_{ij} is the connection weight. The initial weight in the convolution kernel is between 0 and 1.

C. Photonic SNN

The extracted feature from the flattened pooling layer of CSNN is inversely transformed to the firing time of the neuron in the input layer of photonic SNN, as shown in Fig. 5. The vertical-cavity surface-emitting laser with saturable absorber (VCSEL-SA) is employed as the photonic spiking neuron of the photonics SNN, for more detail on the model and parameters please refer to Refs. [5], [9].

We proposed a modified time-based supervised training algorithm to accomplish the speech classification task. The weight can be updated as follows,

$$\Delta\omega_{ij} = \begin{cases} 0, & n_j = 1, j = label \\ \alpha_1 K\left(T - t_i - t_{delay}\right), & n_j = 0, j = label \\ 0, & n_j = 0, j \neq label \\ \alpha_2 K\left(t_j - t_i - t_{delay}\right) & n_j = 1, j \neq label; t_j \leq t_{label} \\ 0, & n_j = 1, j \neq label; t_j \geq t_{label} + T_{thre} \\ \alpha_2\beta K\left(t_j - t_i - t_{delay}\right), & n_j = 1, j \neq label; t_{label} < t_j \leq t_{label} + T_{thre} \end{cases} \tag{6}$$

where j = 1,2,…,10 represents the output neurons. t_i reprents the firing time of *i*-th neuron in the input layer. t_{delay} denotes the transmission delay. t_j (n_j) is the firing time (number of spikes) of the *j*-th neuron of the output layer. *T* is the simulation time window and $T = 15$ns. When the output neuron does not fire a spike, the firing time is considered as the value of *T*. T_{thre}=4ns is a defined judgment reference when more than one output neurons fire spikes. If the earliest spike generated by the label neuron is sufficiently far away from the spikes generated by other output neurons, i.e, $t_j \geq t_{label} + T_{thre}$, then we assume that it will not affect the classification results and keep the weight unchanged. But if $t_{label} < t_j \leq t_{label} + T_{thre}$, the weight still need to be updated. α_1=0.02, α_2= −0.02 represent the learning rate. To accelerate the training convergence, we further introduce another learning rate as $\beta = \cos[\frac{\pi}{2} \times (t_j - t_{label})/T_{thre}]$.

Here, *K* function represents the STDP curve.

$$K(t) = \begin{cases} A_{-}e^{-\frac{t}{\tau_{-}}}, t \leq 0 \\ A_{+}e^{-\frac{t}{\tau_{+}}}, t > 0 \end{cases} \tag{7}$$

where $A_{+} = 0.777$, $A_{-} = -0.777$, $\tau_{+} = 16.8$ns, $\tau_{-} = -16.8$ns.

After training convergence, the converged weights are used to realize the speech recognition. During the inference phase, the neuron in the output layer that fires earliest is considered as the notion of the speech class. For example, if the 1-st neuron fires earliest, the speech signal is considered as 0. Similarly, if the 2-nd (10-th) neuron fires earliest, the speech signal is considered as 1 (9).

III. Results and Discussions

In this paper, the TIDIGITS dataset [45] is used to test the speech recognition performance of the proposed hybrid PCSNN network architecture. The recorded speech content is the English pronunciation of the isolated numbers 0 to 9. We choose 4000 recordings from 200 different people, and the ratio of men to women among the 200 participants was about 1:1. 3600 recordings are considered as the training set, and the rest 400 recordings are considered as the testing set.

The spectrogram after the endpoint detection is presented in Fig. 6 for the samples 0-9. Compared with the original speech signal, the non-effective speech segments at the front and back

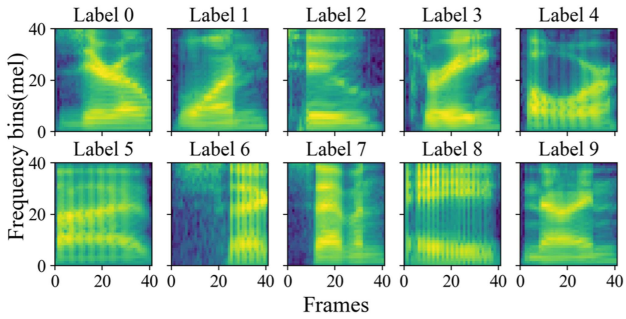

Fig. 6. The spectrogram or the samples 0-9 after the endpoint detection.

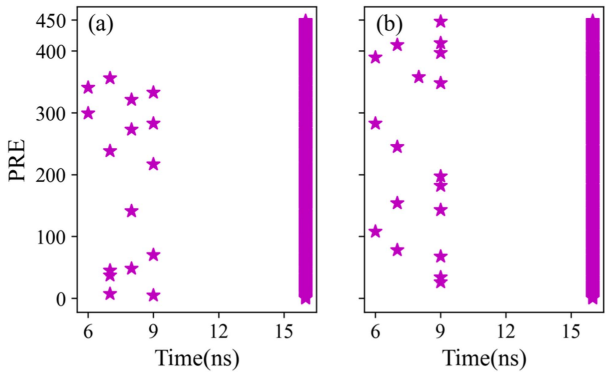

Fig. 7. The spatial-temporal encoding of the input layer neurons for label 0 (a) and 1 (b).

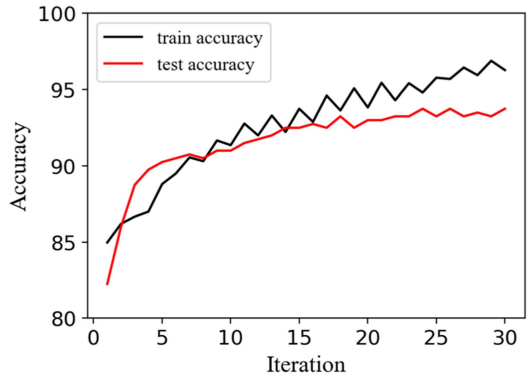

Fig. 8. The training and testing accuracy as a function of training iteration.

ends are truncated, and the effective speech segment in the middle part of the speech is retained, making the proportion of the effective speech segment larger. Label 0 to label 9 correspond to speech data 0 to 9.

The feature values extracted by pooling layer are integers ranging from 0 to 4. When the feature value is 0, it is encoded as T. When the feature value is 1, 2, 3, 4, the encoded time is 9ns, 8ns, 7ns, and 6ns. For instance, the spatial-temporal encoding of the input layer neurons for label 0 and 1 are presented in Fig. 7.

The training and testing accuracy is presented in Fig. 8. It can be seen that, the training accuracy is above 0.95 after 20 iteration, and the testing accuracy is above 0.9 after 10 iteration. The highest testing accuracy is 93.75%.The firing times of the output neurons for ten different samples are further presented

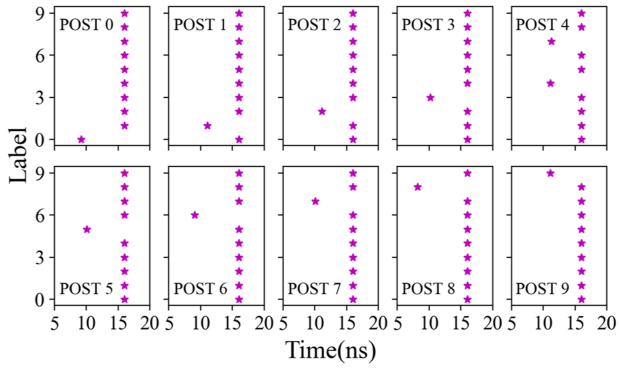

Fig. 9. The firing times of the output neurons for ten kinds of samples.

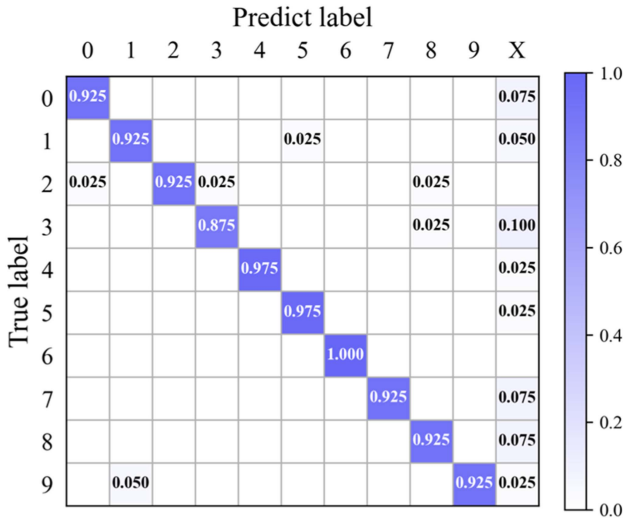

Fig. 10. Confusion matrix. 0∼9 correspond to 10 kinds of speech class. X represents undistinguishable case.

in Fig. 9. It can be seen that, each sample can be successfully recognized. For the input samples of label 7, we can find that, both POST4 and POST7 fire spikes, but the spike generated by POST4 (11.23ns) is latter than that generated by POST7 (10.07ns), and thus it still can be correctly recognized. The confusion matrix is further presented in Fig. 10.

Note, the number of photonic neurons is 450 in the above-mentioned scheme, but as can be seen in Fig. 7, the spike encoding is sparse. To further consider the hardware constraints, we also try to reduce the number of feature maps and frames, which can reduce the hardware node numbers of the photonic SNN. To obtain statistical significance, we run each test three times with different initial weights. Here, the number of frames is fixed at 41. As can be seen from Fig. 11(a), when the number of feature maps is 10, the minimum (maximum) accuracy can reach 82% (85.75%). For this case, the number of input neurons of photonic SNN is 90. With the increase of number of feature maps, the accuracy is increased firstly and almost saturated for larger number of feature maps. When the number of feature maps is 30, the minimum (maximum) accuracy can reach 90% (91.75%). When the number of feature maps reaches 50, the highest accuracy reaches 93.75%.

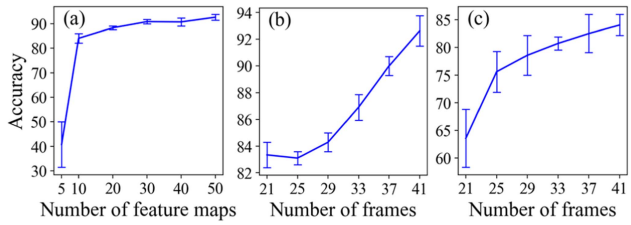

Fig. 11. Test accuracy for different (a) number of feature maps and (b) number of frames. Performance comparison of different (c) number of frame sizes over three runs, and the number of feature map is 10 in CSNN. The accuracy is calculated over three runs for different initial weights.

TABLE I
THE TRAINING AND TEST ACCURACY FOR DIFFERENT DATASETS

Dataset	Accuracy	
	Training set	Test set
TIDIGITS	95.85%	93.75%
TI 20-Word	98.50%	92.19%
FSDD	97.52%	84.00%

The accuracy for different numbers of frames is presented in Fig. 11(b). It can be seen that, when the number of frames is 21 or 25, the accuracy is lower than 84%. For even larger number of frames, the accuracy increases almost linearly.

We consider the number of feature map is 10, and present the accuracy for different number of frames in Fig. 11(c). It can be seen that, when the number of frames is 29, the accuracy is still above 80%. In this case, the number of input neurons of PSNN is reduced to 60, which is highly desirable for the hardware implementation.

Without loss of generality, we also considered some other speech recognition tasks. Here, the TI 20-Word dataset [46] and the Free Spoken Digit Dataset (FSDD) [47] are used. The TI 20-Word dataset contains the sounds of the numbers $0 \sim 9$ as well as 10 other words. Only the sounds of the numbers 0 to 9 are used here. This dataset was recorded by 8 males and 8 females. Each one recorded 26 pieces of data for each number, 16 of which are adopted for training, and the rest 10 pieces are used for testing. Therefore, the size of the training set was 2560, and the size of the test set was 1600. The FSDD dataset contains the sounds of the numbers $0 \sim 9$, recorded by 6 males. Each one recorded 50 pieces of data for each number, 45 pieces of which are used for training and 5 pieces of which are employed for testing. Therefore, the size of the training set is 2700, and the size of test set is 300. The results are presented in Table I, and are compared with the TIDIGITS dataset. It is shown that, high recognition accuracy can also be achieved for these datasets.

IV. CONCLUSION

In conclusion, we proposed a hybrid CSNN and photonic SNN architecture to accomplish the speech recognition task. The dual-threshold endpoint detection algorithm is employed to obtain the FBank spectrogram, the CSNN is then adopted to extract feature from the spectrogram with unsupervised learinging algorithm based on the STDP rule. The photonic SNN is further employed and a modified time-based supervised training algorithm is proposed to accomplish the speech classfication according to the feature value extrated by the CSNN. The TIDIGITS dataset is used to test the speech recognition performance, and the highest testing accuracy is 93.75%. In addition, high recognition accurcy can also be achieved for other datasets. This work provides a solution for extending the application of photonic SNN to the field of speech recognition, and thus, is interesting and valuable for the photonic neuromorphic computing and information processing.

REFERENCES

[1] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, no. 2, pp. 102–114, Jan. 2021.
[2] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, Nov. 2019.
[3] A. Hurtado, I. D. Henning, and M. J. Adams, "Optical neuron using polarisation switching in a 1550nm-VCSEL," *Opt. Exp.*, vol. 18, no. 24, pp. 25170–25176, Nov. 2010.
[4] S. Barbay, R. Kuszelewicz, and A. M. Yacomotti, "Excitability in a semiconductor laser with saturable absorber," *Opt. Lett.*, vol. 36, no. 23, pp. 4476–4478, Nov. 2011.
[5] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 5, Sep./Oct. 2013, Art. no. 1800212.
[6] T. Deng, J. Robertson, and A. Hurtado, "Controlled propagation of spiking dynamics in vertical-cavity surface-emitting lasers: Towards neuromorphic photonic networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 23, no. 6, Nov./Dec. 2017, Art. no. 1800408.
[7] J. Robertson, E. Wade, Y. Kopp, J. Bueno, and A. Hurtado, "Toward neuromorphic photonic networks of ultrafast spiking laser neurons," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7700715.
[8] S. Xiang et al., "STDP-based unsupervised spike pattern learning in a photonic spiking neural network with VCSELs and VCSOAs," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 6, Nov./Dec. 2019, Art. no. 1700109.
[9] S. Xiang et al., "Computing primitive of fully VCSEL-based all-optical spiking neural network for supervised learning and pattern classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2494–2505, Jun. 2021.
[10] J. Xiang, Y. Zhang, Y. Zhao, X. Guo, and Y. Su, "All-optical silicon microring spiking neuron," *Photon. Res.*, vol. 10, no. 4, pp. 939–946, Mar. 2022.
[11] A. Jha, C. Huang, H.-T. Peng, B. Shastri, and P. R. Prucnal, "Photonic spiking neural networks and graphene-on-silicon spiking neurons," *J. Lightw. Technol.*, vol. 40, no. 9, pp. 2901–2914, May 2022.
[12] S. Xiang et al., "Hardware-algorithm collaborative computing with photonic spiking neuron chip based on integrated Fabry–Pérot laser with saturable absorber," *Optica,* vol. 10, no. 2, pp. 162–171, Feb. 2023.
[13] D. Zheng et al., "Experimental demonstration of coherent photonic neural computing based on Fabry–Pérot laser with a saturable absorber," *Photon. Res.*, vol. 11, no. 1, pp. 65–71, Jan. 2023.
[14] H.-T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 6101715.
[15] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, May 2019.
[16] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jun. 2017.
[17] H. Zhou et al., "Chip-scale optical matrix computation for PageRank algorithm," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 2, Mar./Apr. 2020, Art. no. 8300910.
[18] Y. Tian et al., "Scalable and compact photonic neural chip with low learning-capability-loss," *Nanophotonics*, vol. 11, no. 2, pp. 329–344, Dec. 2021.
[19] Z. Song et al., "A hybrid-integrated photonic spiking neural network framework based on an MZI array and VCSELs-SA," *IEEE J. Sel. Topics Quantum Electron.*, vol. 29, no. 2, Mar./Apr. 2023, Art. no. 8300211.

[20] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightw. Technol.*, vol. 32, no. 21, pp. 4029–4041, Nov. 2014.
[21] A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Aug. 2017.
[22] C. Huang et al., "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photon.*, vol. 5, no. 4, Apr. 2020, Art. no. 040803.
[23] F. Ponulak and A. Kasiński, "Supervised learning in spiking neural networks with ReSuMe: Sequence learning, classification, and spike shifting," *Neural Computation*, vol. 22, no. 2, pp. 467–510, Feb. 2010.
[24] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Front. Comput. Neurosci.*, vol. 9, Aug. 2015, Art. no. 99.
[25] S. Xiang et al., "Training a multi-layer photonic spiking neural network with modified supervised learning algorithm based on photonic STDP," *IEEE J. Sel. Topics Quantum Electron.*, vol. 27, no. 2, Mar./Apr. 2021, Art. no. 7500109.
[26] Y. Han et al., "Delay-weight plasticity-based supervised learning in optical spiking neural networks," *Photon. Res.*, vol. 9, no. 4, pp. B119–B127, Mar. 2021.
[27] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1085–1093, Jul. 2013.
[28] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, "Temporal coding of local spectrogram features for robust sound recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 803–807.
[29] R. Xiao, R. Yan, H. Tang, and K. C. Tan, "A spiking neural network model for sound recognition," in *Proc. Int. Conf. Cogn. Syst. Signal Process.*, 2016, pp. 584–594.
[30] A. Tavanaei and A. S. Maida, "A spiking network that learns to extract spike signatures from speech signals," *Neurocomputing*, vol. 240, pp. 191–199, May 2017.
[31] G. Srinivasan, P. Panda, and K. Roy, "Spilinc: Spiking liquid-ensemble computing for unsupervised speech and image recognition," *Front. Neurosci.*, vol. 12, Aug. 2018, Art. no. 524.
[32] Y. Yao, Q. Yu, L. Wang, and J. Dang, "A spiking neural network with distributed keypoint encoding for robust sound recognition," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[33] R. Gütig and H. Sompolinsky, "Time-warp–invariant neuronal processing," *PLoS Biol.*, vol. 7, no. 7, 2009, Art. no. e1000141.
[34] Q. Yu et al., "Improving multispike learning with plastic synaptic delays," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 20, 2022, doi: 10.1109/TNNLS.2022.3165527.
[35] A. Tavanaei and A. Maida, "Bio-inspired multi-layer spiking neural network extracts discriminative features from speech signals," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 899–908.
[36] M. Dong, X. Huang, and B. Xu, "Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network," *PLoS One*, vol. 13, no. 11, Nov. 2018, Art. no. e0204596.
[37] Z. Zhang and Q. Liu, "Spike-event-driven deep spiking neural network with temporal encoding," *IEEE Signal Process. Lett.*, vol. 28, pp. 484–488, 2021.
[38] A. Tavanaei, Z. Kirby, and A. S. Maida, "Training spiking convnets by STDP and gradient descent," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.
[39] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "STDP-based spiking deep convolutional neural networks for object recognition," *Neural Netw.*, vol. 99, pp. 56–67, Mar. 2018.
[40] C. Lee, G. Srinivasan, P. Panda, and K. Roy, "Deep spiking convolutional neural network trained with unsupervised spike-timing-dependent plasticity," *IEEE Trans. Cogn. Devlop. Syst.*, vol. 11, no. 3, pp. 384–394, Sep. 2019.
[41] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*.
[42] V. Lostanlen and C.-E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," 2016, *arXiv:1605.06644*.
[43] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," 2017, *arXiv:1709.04396*.
[44] T. Masquelier and S. J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity," *PLoS Comput. Biol.*, vol. 3, no. 2, Feb. 2007, Art. no. e31.
[45] R. G. Leonard and G. Doddington, *Tidigits Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[46] Texas Instruments. 46-Word Speaker-Dependent Isolated Word Corpus (TI-46), NIST Speech Disc 7-1.1, NIST, 1991. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S9
[47] Z. Jackson, "Free spoken digit dataset (FSDD)," *Tech. Rep.*, 2016.

Shuiying Xiang was born in Jiangxi Province, China, in 1986. She received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2013. She is currently an Associate Professor with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China. Her research interests include vertical cavity surface-emitting lasers, neuromorphic photonic systems, brain-inspired information processing, chaotic optical communication, and semiconductor lasers dynamics.

Tianrui Zhang was born in Xi'an, China, in 1998. He is currently working toward the M.S. degree with Xidian University, Xi'an. His research interests include machine learning and spiking neural network.

Yanan Han was born in Ningxia Hui Autonomous Region, China, in 1996. She is currently working toward the Ph.D. degree with Xidian University, Xi'an, China. Her research interests include the dynamics and applications of semiconductor lasers, random number generators, and brain-inspired information processing.

Xingxing Guo was born in Ji'an, China, in 1993. She received the Ph.D. degree from Xidian University, Xi'an, China. Her research focuses on the dynamics and applications of semiconductor lasers.

Yahui Zhang was born in Zhangjiakou, China, in 1993. She received the Ph.D. degree from Xidian University, Xi'an, China. Her research interests include vertical cavity surface emitting lasers, neuromorphic photonic systems, brain-inspired information processing, and spiking neural networks.

Yuechun Shi was born in Nantong, China, in 1983. He received the Ph.D. degree from Nanjing University, Nanjing, China, in 2012. During 2016–2022, he was an Associate Professor with the College of Engineering and Applied Sciences, Institute of Optical Communication Engineering, Nanjing University. He is currently the Research Fellow with Yongjiang Laboratory, Ningbo, China. His research interests include semiconductor laser/ array, integrated photonics, photonic neuromorphic computing, and optical sensor.

Yue Hao (Senior Member, IEEE) was born in Chongqing, China, in 1958. He received the Ph.D. degree from Xi'an Jiao tong University, Xi'an, China, in 1991. He is currently a Professor with the State Key Discipline Laboratory of Wide Bandgap Semiconductor Technology, School of Microelectronics, Xidian University, Xi'an. His research focuses on wide forbidden band semiconductor materials and devices.