INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Robust Sound Recognition: A Neuromorphic Approach

Jibin Wu[1], *Zihan Pan*[1], *Malu Zhang*[1], *Rohan Kumar Das*[1], *Yansong Chua*[2], *Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore

jibin.wu@u.nus.edu

Abstract

Humans perform remarkably well at sound classification that is used as cues to support high-level cognitive functions. Inspired by the anatomical structure of human cochlea and auditory attention mechanism, we present a novel neuromorphic sound recognition system that integrates an event-driven auditory front-end and a biologically plausible spiking neural network classifier (SNN) for robust sound and speech recognition. Due to its event-driven nature, the SNN classifier is several orders of magnitude more energy efficient than deep learning classifier, therefore, it is suitable for many applications in wearable devices.

Index Terms: automatic speech recognition, environmental sound recognition, auditory masking, spiking neural networks, neural threshold coding, aggregate-label learning

1. Introduction

The deep learning approaches have achieved remarkable success with the availability of a large amount of labeled training data, growing computational resources and effective network architectures. Automatic sound and speech recognition is no exception to it that have achieved human-level accuracies on a number of benchmark datasets. While being biologically-inspired, these artificial neural network models differ from the biological auditory system in many ways. Fundamentally, sound waves are encoded, transmitted and exchanged using asynchronous action potentials or spiking events in the human auditory system. These spikes are transmitted on the neural substrates at a speed that is several orders of magnitude slower than that of electrons on conventional silicon substrates. In spite of this, humans perform effectively on auditory perception tasks with highly parallel spiking neural networks and auditory attention mechanism.

In this work, we present an interactive neuromorphic sound recognition system with application to the environmental sound classification and speech recognition. Inspired by the anatomical structure of the human cochlea and psychological studies of human auditory attention mechanism, we introduce a novel neuromorphic auditory front-end. This front-end integrates the biologically plausible cochlear filter bank, auditory masking and neural threshold coding, to faithfully encode the spectral information into spatiotemporal spike patterns. Additionally, we apply our recently proposed membrane potential dependent aggregate label learning algorithm (MPD-AL) for training the event-driven SNN classifier to reliably recognize the underlying spike patterns [1]. To allow a better understanding of our system, a graphical user interface (GUI) is implemented for users to explore pre-recorded sound files[1], and an example of continuous spoken digit recognition is shown in Figure 1.

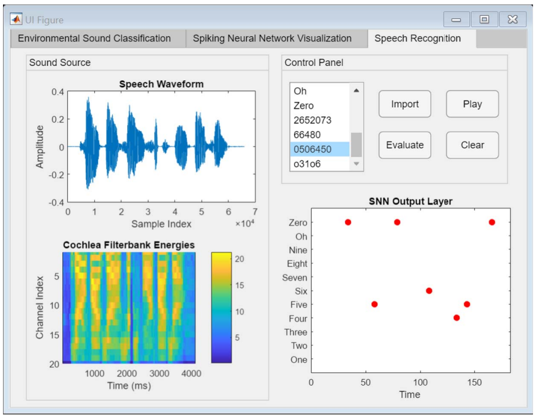

Figure 1: *Example of continuous spoken digit recognition with the proposed neuromorphic sound recognition system. The output neurons spike (highlighted in red) whenever the represented word class is presented in the input spike trains.*

2. System Description

As illustrated in Figure 2, our system consists of two main processing stages that are organized in a pipelined structure: a neuromorphic auditory front-end that effectively and efficiently encodes sound signals into spike trains, and a temporal classification mechanism that uses the event-driven SNN for decision-making.

2.1. Biologically Plausible Cochlear Filter Bank

The biologically-inspired cochlear filter banks (e.g., Mel-scaled filter bank and Gammatone filter bank) are widely used as a spectral analyzer in the conventional speech and sound recognition systems. However, they are not designed to match the event-driven nature of the spiking neural network, which is the most distinctive feature of a neuromorphic approach. Therefore, attention should be paid in reinvestigating the design of existing cochlear filter banks that are designed for synchronous computing machines. In our system, we introduce an asynchronous, time-domain cochlear filter bank [2] that can be implemented on the parallel neuromorphic hardware without any difficulty.

2.2. Auditory Masking

Humans possess outstanding auditory attention capability, whereby information processing is concentrated on the salient regions over auditory receptive fields. Inspired by the psychological studies of human auditory attention mechanism, we implemented both temporal and simultaneous masking [3] that occur in the time and frequency domains, respectively. These masking mechanisms can extract salient temporal and

[1] https://youtu.be/MIVvNb0sWOM

Copyright © 2019 ISCA

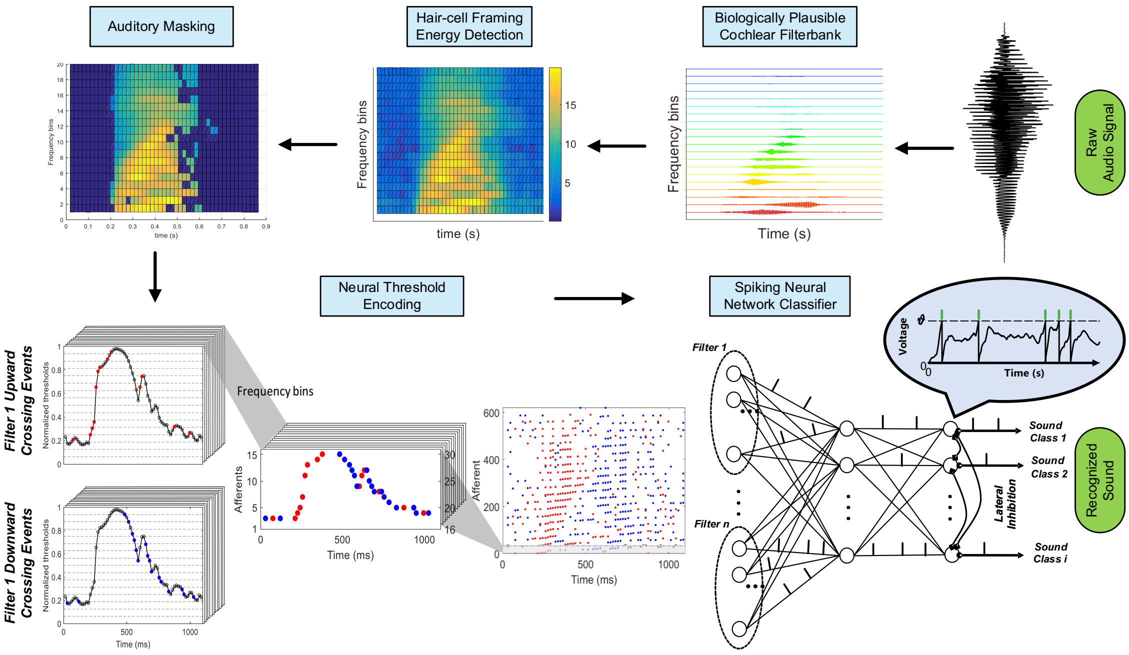

Figure 2: *Schematic diagram of the proposed neuromorphic sound recognition system.*

frequency features from the cochlear filter bank outputs, without any perceptible distortion, as well as reduce the overall spike rate.

2.3. Neural Threshold Coding

Over time, neuroscientists have revealed many neural coding schemes for the human auditory system, including rate and temporal codes (e.g., latency code and phase code). These neural codes could be used to faithfully encode stimuli, and to some extent, biologically plausible to the human auditory system. However, the usefulness for the backend SNN classifier and computational costs from a system perspective are usually ignored. Therefore, it is important to determine an effective and efficient neural coding scheme to generate discriminative spiking patterns for the downstream SNN classifier. In our system, we develop a neural threshold coding mechanism [1] that preserve the temporal dynamics of the filtered spectral information by only encoding the upward and downward crossing events. Such neural coding scheme significantly reduces the computational demand of the backend SNN classifier.

2.4. SNN Classifier with Multi-Condition Training

Spiking neurons exhibit rich temporal dynamics in their subthreshold membrane potential, making them well suited for processing temporally rich sound signals. However, due to the non-differentiable nature of spiking events, the powerful backpropagation algorithm is not directly applicable to SNNs. To resolve this predicament, our recently proposed MPD-AL algorithm with dynamic decoding scheme is applied to assign credits to those discriminative temporal features. Furthermore, we investigate training the proposed SNN model with both clean and noise-corrupted sound samples, as per multi-condition training strategy [4]. This strategy effectively improves the system noise robustness as can be observed from our demonstration[1].

3. Discussion and Conclusion

Automatic sound classification is required in many real-life applications. However, traditional pattern classification and deep learning techniques rely on high-performance computing that prevents such systems from large scale deployment. Furthermore, growing concern about information security and demand for personalized systems call for energy efficient solutions. Therefore, the neuromorphic approach introduced in this work offers an attractive solution to tackle all the aforementioned problems and represents an important milestone towards future neuromorphic computing machines.

4. Acknowledgements

This research is supported by Programmatic grant no. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain)

5. References

[1] M. Zhang, J. Wu, Y. Chua, X. Luo, Z. Pan, D. Liu, and H. Li, "MPD-AL: An efficient membrane potential driven aggregate-label learning algorithm for spiking neurons," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[2] Z. Pan, H. Li, J. Wu, and Y. Chua, "An event-based cochlear filter temporal encoding scheme for speech signals," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[3] T. S. Gunawan, E. Ambikairajah, and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech communication*, vol. 52, no. 5, pp. 381–393, 2010.

[4] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in Neuroscience*, vol. 12, p. 836, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00836