A MULTI-SPIKE APPROACH FOR ROBUST SOUND RECOGNITION

Qiang Yu[1] *Yanli Yao*[1] *Longbiao Wang*[1] *Huajin Tang*[2] *Jianwu Dang*[1]

[1] Tianjin Key Laboratory of Cognitive Computing and Application
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] College of Computer Science, Sichuan University, Chengdu, China
{yuqiang, yaoyanli, longbiao_wang}@tju.edu.cn, htang@scu.edu.cn, jdang@jaist.ac.jp

ABSTRACT

The extraordinary performance of the brain on various cognitive tasks motivates the design of a biologically plausible system for the challenging task of environmental sound recognition. In this paper, we propose a novel approach based on multi-spike learning and key-point encoding. Our encoding extracts local temporal and spectral information from the sound and converts it into spatiotemporal spike pattern, which is further learned by the following spiking neural networks. Our experiments demonstrate the robustness and effectiveness of our approach across a variety of noise conditions, outperforming other conventional baseline methods in both mismatched and multi-condition scenarios.

Index Terms— Sound recognition, neural coding, multi-spike learning, neuromorphic computing

1. INTRODUCTION

As one of the major challenges in the acoustic processing field [1], environmental sound recognition has been receiving increasing interest in recent years not only because of its research value for addressing the chaotic and unstructured difficulties but also its importance in various applied developments such as bioacoustic monitoring [2], surveillance [3] and general machine hearing [4]. Successful sound recognition can provide a prompt description of a scene, which could be used to take any further actions. This audio-based approach is cheap and advantageous as is compared to a vision-based one considering poor lighting or visual obstruction. Additionally, the fundamental differences between the tasks of automatic speech recognition and sound recognition motivate studies designed specifically for sound [5–8]. Unlike structured speech or music signal, sounds are unstructured and present in continuous background noise, which reduces the discrimination of extracted features. Considering the extraordinary performance of the human brain on this challenging sound recognition task, a human-like or brain-inspired approach is demanded to effectively and efficiently detect sound events even under severe noise.

Like a typical recognition problem, sound recognition can be divided into three major functional parts: preprocessing, feature extraction and classification. In the stage of preprocessing, signals are modified in a way to facilitate the following phase of feature extraction where a proper representation of the signal is formed. The learning and recognition are involved in the final stage of classification. Different approaches can be categorized according to different employed methods in these three stages.

A conventional approach typically adopts a frame-based feature extraction where Mel-Frequency Ceptral Coefficients (MFCC) is one of the most popular choices, followed by Hidden Markov Model (HMM) for capturing the temporal information. There are two major drawbacks of such an approach. Firstly, MFCC features are more favorable for modeling single sound sources but not for environmental sounds which typically contain a variety of sources [8]. These MFCC features are modeled based on the overall spectrum, making them vulnerable to noise. Secondly, HMM does not model explicitly the diverse temporal dependencies of environmental sounds. It relies on a first-order state transition to implicitly model the temporal coding of the signal. Efforts have been made to improve the system performance by either incorporating sophisticated feature representations, such as stabilized auditory image [9], spectrogram image features [10] and matching pursuit [8], or involving advanced machine learning techniques such as CNN [6] and DNN [7]. However, the temporal structure of the signal is still not well modeled in these approaches. Additionally, almost none of these take consideration of processing information in a more biologically plausible way where spikes are adopted as like in the brain [11].

To address the above challenges, we propose a novel KP-SNN approach for sound recognition, based on key-point (KP) feature encoding and spiking neural network (SNN) processing. In a previous work [5], we combined local spectrogram features (LSFs) with the tempotron learning rule [12]

This work was supported by the Natural Science Foundation of China (No. 61806139, 61771333), and the Natural Science Foundation of Tianjin (No. 18JCYBJC41700). Corresponding author: Qiang Yu.

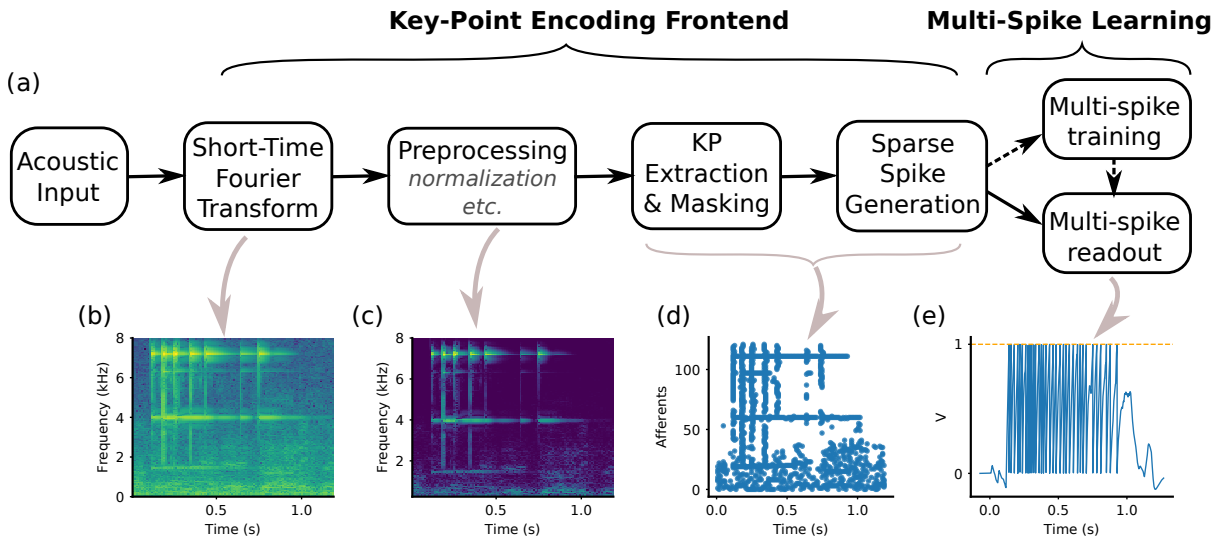

Fig. 1. Framework demonstration of the multi-spike based sound recognition system. (a) information processing framework; (b)-(d) mid-level representations of a sample sound under 10 dB noise; (e) target neural dynamics in response to the sound.

where neurons make decisions by firing one spike or keeping silent. The temporal coverage of the tempotron is limited to a single local feature in time. Here, our purpose is to utilize multi-spike learning [13, 14] to improve the performance by covering the whole time window of the signal with more spikes, which is one significant departure from our previous work. Another major difference with our previous work is that we simplify the encoding by directly using the extracted KPs without taking any extra steps for feature clustering. Such simplification would be beneficial for low-power and efficient on-line processing, especially when the system is to be deployed in devices like wearables or cameras. Additionally, we enhance the masking freedom of our encoding to generalize its potential for broader situations.

Our motivation is to develop a sound recognition system that is inspired by the human brain and thus inherits its advantages, such as robustness, effectiveness and efficiency, to a certain extend. It has suggested that the auditory systems utilize the local time-frequency regions with high signal-to-noise ratio (SNR) to process noise corrupted signals [15]. This supports the idea of using KPs. In addition, neurons in the brain communicate with each other with spikes [11], which inspires the idea of using SNNs for processing sound signals. Neurons typically evolve with the dynamic environment and elicit as many spikes as necessary whenever a firing condition is reached. This makes a multi-spike learning more favorable.

The rest of this paper is organized as follows: Section 2 describes our proposed methods including the feature encoding and multi-spike learning algorithms. Section 3 introduces the experiments used for evaluating our approach, as well as the results, followed by Section 4 which concludes our work.

2. METHODOLOGY

Following a typical pattern recognition system, we construct our sound recognition framework by combining a feature extraction and a classifier. In order to apply SNNs, a proper encoding is necessary to convert external information into spikes [16]. The idea is to represent and process information in a sparse spike form. Fig. 1(a) gives an overview of the proposed system where key-points are used for encoding and multi-spike learning is adopted as the decision-making part.

2.1. Key-point spike encoding

The key-points (KPs) are detected by localizing the sparse high-energy peaks in the spectrogram. These peaks are inherently robust to mismatched noise due to the property of local maximum. As detected from the spectrogram, each KP carries a local time-frequency information which will be sufficient to construct a spatiotemporal spike pattern for further processing with SNNs. This thus forms the idea of a sparse and robust encoding with KPs.

In order to detect KPs, external acoustic signal is first converted into spectrogram by Short-Time Fourier Transform. The resulting spectrogram, $S(t, f)$, describes the power spectral density of the sound signal over both time and frequency dimension. The spectrogram is then normalized to unit peak amplitude, followed by a logarithm step to convert it into log scale through $log(S(t, f) + \epsilon) - log(\epsilon)$ with $\epsilon = 10^{-5}$. Another normalization is conducted before sending the spectrogram to KP detection, and we still use the notation of $S(t, f)$ at this stage for simplicity.

KPs are detected on the resulting spectrogram by search-

ing local maxima across either time or frequency, as follows:

$$P(t,f) = \left\{ S(t,f) \middle| S(t,f) = max \left\{ \begin{array}{l} S(t \pm r_t, f) \text{ or} \\ S(t, f \pm r_f) \end{array} \right\} \right\} \quad (1)$$

Here, $r_{t,f} = 0, 1, ..., R_{t,f}$. $R_t = R_f = 4$ denotes the region size for KP detection, which we found was big enough for a sparse representation, but small enough to extract important peaks.

In order to further improve the sparsity of KPs, we introduce two different masking schemes: absolute-value masking and relative-background masking. In the absolute-value scheme, we will discard those KPs that satisfies the criteria of $P(t,f) < \beta_{\mathrm{a}}$. In the relative-background masking scheme, we compare the contrast of each KP and background mean value to remove those according to $P(t,f) * \beta_{\mathrm{r}} < mean\{S(t \pm d_t, f \pm d_f)\}$. β_{a} and β_{r} are the two hyper-parameters that control the level of reduction on the number of KPs. In our experiments, we set $\beta_{\mathrm{a}} = 0.15$ and $\beta_{\mathrm{r}} = 0.85$.

In [5], Local Spectrogram Features (LSFs) are further extracted based on the KPs, and modeled by a Self-Organizing Map (SOM) to generate a spike pattern. We find these steps are unnecessary for spike encoding, thus increasing the complexity of the system. The KPs contain both temporal and spectral information, which are sufficient enough to form a spatiotemporal spike pattern. In this paper, we will examine this idea of simplified encoding. Fig.1(b-d) illustrates an encoding example.

2.2. Multi-spike learning

The spike learning rule is used to process spike-based patterns and train neurons to adapt their weights for a favorable response to input signals. The tempotron rule introduces an efficient plasticity algorithm to train a neuron to elicit a single spike in response to target patterns while keep silent to others [12]. This rule was employed in our previous work [5] which has successfully demonstrated the strength of SNNs for the given task. However, the tempotron rule is designed to constrain neurons to have binary response only, namely either a single spike or none. Considering the multi-spike behavior in the brain, this binary firing dynamics is a departure from approaches directing to biological plausibility. Moreover, each spike decision is made based on a single local temporal region with the rest not being fully utilized. Therefore, a learning rule with multi-spike capability is more favorable than a binary one for the sound recognition task. We have proposed a group of multi-spike learning algorithms [13, 14, 17] which can train neurons to fire a desired number of spikes in response to a target pattern, and we adopt such a multi-spike learning algorithm in this study to fully explore its benefits for the sound recognition task.

The spiking neuron continuously integrates afferent spikes into its membrane potential, and generates as many spikes as necessary whenever a firing condition is reached. To be specific, the neuron dynamics is as follows:

$$V(t) = \sum_{i=1}^{N} w_i \sum_{t_i^j < t} K(t - t_i^j) - \vartheta \sum_{t_{\mathrm{s}}^j < t} \exp\left(-\frac{t - t_{\mathrm{s}}^j}{\tau_{\mathrm{m}}}\right) \quad (2)$$

where w_i denotes the synaptic efficacy. t_i^j and t_{s}^j are the afferent input spike and neuron's output spike, respectively. ϑ and τ_{m} represent the firing threshold and neuron's time constant, respectively. The kernel K models the postsynaptic potentials, and is given in the following form:

$$K(t - t_i^j) = V_0 \left[\exp\left(-\frac{t - t_i^j}{\tau_{\mathrm{m}}}\right) - \exp\left(-\frac{t - t_i^j}{\tau_{\mathrm{s}}}\right) \right], \quad (3)$$

where V_0 is a constant parameter that normalizes the peak of the kernel to unity, resulting the amplitude is governed by synaptic efficacy.

The relation between neuron's output spike number and its threshold is characterized by a spike-threshold-surface (STS) [14, 18]. A learning rule can thus be derived to modify synaptic weights in such a way that resulted STS can lead to a desired number of spikes. Here, we adopt the TDP1 [14] as our multi-spike learning rule due to its simplicity and efficiency. The gradients of critical threshold ϑ_k^* with respect to weight w_i is given as:

$$\vartheta_i^{*'} = \frac{\partial V(t^*)}{\partial w_i} - \sum_{j=1}^{m} \frac{\partial V(t^*)}{\partial t_{\mathrm{s}}^j} \frac{1}{\dot{V}(t_{\mathrm{s}}^j)} \frac{\partial V(t_{\mathrm{s}}^j)}{\partial w_i} \quad (4)$$

where m denotes the number of output spikes occur before the time of a critical threshold, t^*. Based on Eq. 4, learning rules can thus be developed in a way to increase the number of output spikes if neurons fire less than a target, and to decrease it if more spikes are generated. Detailed descriptions about the learning can be referred in our previous work [14].

In the multi-class sound recognition task, we train one neuron corresponding to each category to fire multiple spikes. Therefore, the total number of output spikes is used for classification.

3. EXPERIMENTS

In this section, we conduct experiments to examine the performance of our proposed system, i.e. KP-SNN, for the challenging task of sound recognition. Experimental results are present with discussions. We compare our approach to a variety of different standard ones that are used for sound recognition, including MFCC-HMM and spectrogram-based deep learning approaches (SPEC-DNN and SPEC-CNN). We also provide comparison to our previous spike-based approach, LSF-SNN, where binary-spike response is used for training. We will examine the effects of multi-spike approach on improving recognition performance by comparing both the multi-spike (mul) and binary-spike (bin) learning rules.

Table 1. Classification accuracy with mismatched training.

Methods	KP-SNN (mul)	KP-SNN (bin)	LSF-SNN	MFCC-HMM	SPEC-DNN	SPEC-CNN
Clean	**100%**	99.35%	98.5%	99.0%	100%	99.83%
20dB	99.5%	96.58%	98.0%	62.1%	94.38%	**99.88%**
10dB	98.68%	94.0%	95.3%	34.4%	71.8%	**98.93%**
0dB	**98.10%**	90.35%	90.2%	21.8%	42.68%	83.65%
-5dB	**97.13%**	82.45%	84.6%	19.5%	34.85%	58.08%
Avg	**98.68%**	92.54%	93.3%	47.3%	68.74%	88.07%

3.1. Experimental dataset

Following the selection in [5], we choose the following ten sound classes from the Real World Computing Partnership (RWCP) [19]: whistle1, ring, phone4, metal15, kara, horn, cymbals, buzzer, bottle1 and bells5. We select the first 80 files of each class to form our experimental dataset. In each experimental run, we randomly select half files of each class as training, and leave the rest as testing. The "Speech Babble" noise environment is obtained from NOISEX92 database [20] for evaluating the robustness of the sound recognition. The performance is averaged over 10 runs of the experiments.

3.2. Experimental setups

The experiments are designed to evaluate the contribution of multi-spike based approach to sound recognition. We primarily select the conventional frame-based MFCC-HMM as one of the baselines. We also use the deep learning approaches, DNN and CNN, which are widely applied to visual and auditory recognitions. The above three baselines form the non-spiking based approaches for comparison with our spike-based ones. Additionally, the LSF-SNN is used to benchmark the contribution of our multi-spike based approach, i.e. KP-SNN (mul). Neurons corresponding to each category in our system are trained to elicit at least 20 spikes, and a decision is voted by the neuron with the most number of output spikes. We trained the system in a clean condition and evaluate it with different levels of noise: clean, 20, 10, 0 and -5 dB. This scenario is denoted as mismatched training. In order to further increase the performance, we conducted a multi-condition training scenario which is commonly used for deep learning. Neurons are trained with random levels of noise imposed on each sample sound under this case.

3.3. Results and discussion

Table 1 shows the recognition accuracies of different approaches under the mismatched condition. As can be seen from the table, the conventional machine learning approaches achieve a high accuracy of over 99% for clean environment, but their performance will decrease rapidly with the increasing noise, resulting in an average accuracy of 47.3% (MFCC-HMM), 68.74% (SPEC-DNN) and 88.07% (SPEC-CNN).

Table 2. Classification accuracy with multi-condition training.

Methods	KP-SNN (mul)	KP-SNN (bin)	SPEC-CNN
Clean	99.65%	99.13%	**99.89%**
20dB	99.83%	99.23%	**99.89%**
10dB	99.73%	99.1%	**99.89%**
0dB	**99.43%**	95.1%	99.11%
-5dB	**98.95%**	89.38%	91.17%
Avg	**99.52%**	96.38%	98.04%

The spiking-based approaches perform relatively well for each of the noisy conditions. Our proposed multi-spike approach, KP-SNN(mul), outperforms all the other approaches in severe noisy conditions, resulting in a strong result with average accuracy of 98.68%. We also evaluate the performance of binary-spike learning under our framework as compared to that in LSF-SNN [5]. The comparative performance of these two binary-spike learning frameworks demonstrates the effectiveness of our simplification on the encoding with KPs. In addition, we apply a multi-condition training scheme to further improve the performance of our system, and the results are presented in Table 2. The noisy training can effectively improve the performance of each approach. The proposed approach still dominates all the others over strong noise levels, with an average accuracy of 99.52%.

The superior performance of our proposed approach relies on the multi-spike dynamics which can reliably cover the whole duration of the sound presence (see Fig. 1(e)). Combing the robust and sparse KP spike encoding, our approach performs well in both mismatched and multi-condition scenarios.

4. CONCLUSION

In this paper, we proposed a novel framework with key-point encoding and multi-spike learning for robust sound recognition. We simplified the spike encoding method by directly converting the temporal and spectral information of KPs into spatiotemporal spikes, and enhanced the encoding freedom with two different masking schemes. Our multi-spike approach could outperform the conventional baselines under both mismatched and multi-condition cases.

5. REFERENCES

[1] Douglas OShaughnessy, "Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[2] Felix Weninger and Björn Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on*. IEEE, 2011, pp. 337–340.

[3] Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "On acoustic surveillance of hazardous situations," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 165–168.

[4] Richard F Lyon, "Machine hearing: An emerging field [exploratory dsp]," *IEEE signal processing magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[5] Jonathan Dennis, Qiang Yu, Huajin Tang, Huy Dat Tran, and Haizhou Li, "Temporal coding of local spectrogram features for robust sound recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 803–807.

[6] Ilyas Ozer, Zeynep Ozer, and Oguz Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, pp. 505–512, 2018.

[7] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[8] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[9] Thomas C Walters, *Auditory-based processing of communication sounds*, Ph.D. thesis, University of Cambridge, 2011.

[10] Jonathan Dennis, Huy Dat Tran, and Haizhou Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE signal processing letters*, vol. 18, no. 2, pp. 130–133, 2011.

[11] Peter Dayan and Laurence F Abbott, *Theoretical neuroscience*, vol. 806, Cambridge, MA: MIT Press, 2001.

[12] Robert Gütig and Haim Sompolinsky, "The tempotron: a neuron that learns spike timing–based decisions," *Nature neuroscience*, vol. 9, no. 3, pp. 420, 2006.

[13] Qiang Yu, Longbiao Wang, and Jianwu Dang, "Neuronal classifier for both rate and timing-based spike patterns," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 759–766.

[14] Qiang Yu, Haizhou Li, and Kay Chen Tan, "Spike timing or rate? neurons learn to make decisions for both through threshold-driven plasticity," *IEEE Transactions on Cybernetics*, 2018.

[15] Jont B Allen, "How do humans process and recognize speech?," *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 567–577, 1994.

[16] Qiang Yu, Huajin Tang, Kay Chen Tan, and Haizhou Li, "Rapid feedforward computation by temporal encoding and learning with spiking neurons," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 10, pp. 1539–1552, 2013.

[17] Qiang Yu, Longbiao Wang, and Jianwu Dang, "Efficient multi-spike learning with tempotron-like ltp and psd-like ltd," in *International Conference on Neural Information Processing*. Springer, in press, 2018.

[18] Robert Gütig, "Spiking neurons can discover predictive features by aggregate-label learning," *Science*, vol. 351, no. 6277, pp. aab4113, 2016.

[19] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition.," in *LREC*, 2000.

[20] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.