# COMBINING ROBUST SPIKE CODING WITH SPIKING NEURAL NETWORKS FOR SOUND EVENT CLASSIFICATION

*Jonathan Dennis, Tran Huy Dat, Haizhou Li*

Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632

## ABSTRACT

This paper proposes a novel biologically inspired method for sound event classification which combines spike coding with a spiking neural network (SNN). Our spike coding extracts keypoints that represent the local maxima components of the sound spectrogram, and are encoded based on their local time-frequency information; hence both location and spectral information are being extracted. We then design a modified tempotron SNN that, unlike the original tempotron, allows the network to learn the temporal distributions of spike coding input, in an analogous way to the generalized Hough transform. The proposed method simultaneously enhances the sparsity of the sound event spectrogram, producing a representation which is robust against noise, as well as maximises the discriminability of the spike coding input in terms of its temporal information, which is important for sound event classification. Experimental results on a large dataset of 50 environment sound events show the superiority of both the spike coding versus the raw spectrogram and the SNN versus conventional cross-entropy neural networks.

***Index Terms***— Neural spike coding, local spectrogram features, noise robust, sound event classification.

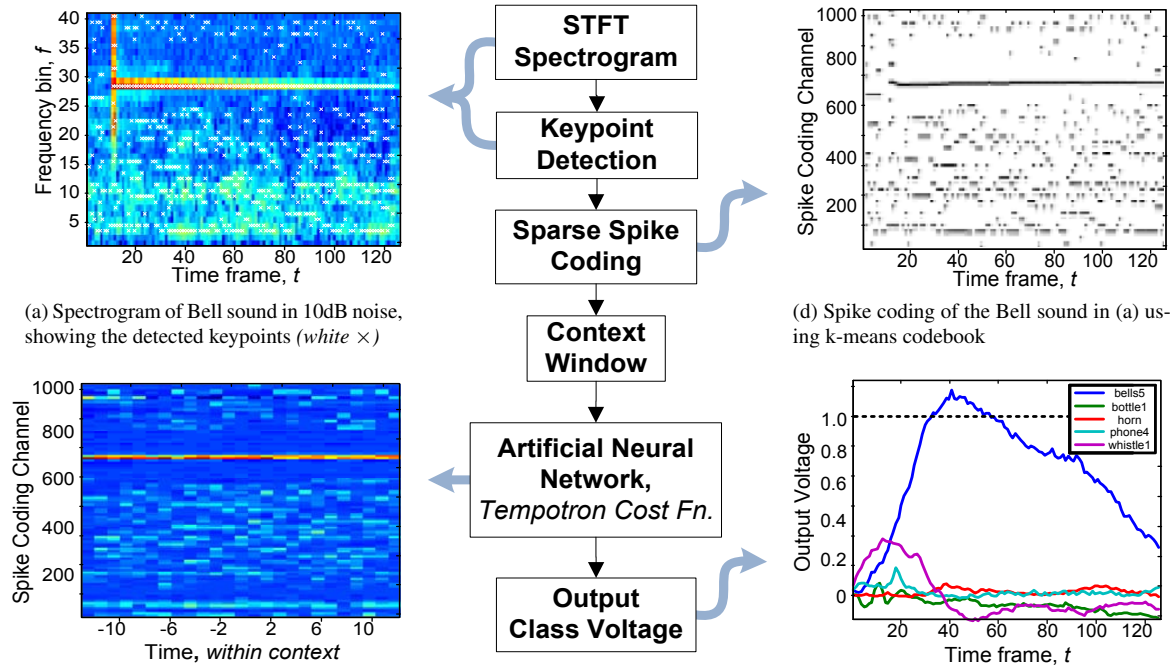
## 1. INTRODUCTION

The field of sound event classification has been receiving renewed interest in recent years due to the wide range of possible applications. These include acoustic surveillance [1], bioacoustic monitoring [2], environmental sound detection [3, 4], or general machine hearing [5]. While the task of sound event classification shares many similarities with the task of automatic speech recognition (ASR), there a number of fundamental differences that motivate the study of approaches designed specifically for sound events. In particular the problem of environmental mismatch is important, since the physical nature of many sounds means that the majority of the spectrogram may be dominated by the background noise, reducing the discrimination of the extracted feature representation.

To address this challenging task, we propose a novel spiking neural network (SNN) system that combines a robust spike coding of local spectrogram features, with an artificial neural network using a cost function that promotes a spiking detection-based output. The idea is that the neural network can learn a mapping from the input spike coding to produce a output spike at a particular moment in time which indicates the detection of the given sound class. There has been previous study into such biologically plausible recognition systems, and here we focus on the Tempotron learning approach [6], which we have previously employed in our preliminary work [7]. The Tempotron cost function reinforces the strongest output over the reference segment from the positive class until a spike is produced, while penalising negative classes that associate strongly with the given input. In this work we propose to generalise the Tempotron to allow the network to learn the distribution of the input spike coding over time and frequency by providing a context window of spikes at the input of the network. This is unlike the distribution function in the original Tempotron, which models the biologically inspired leaky integrate-and-fire activation as a function of two exponentials [6], and can be seen as a similar formulation to the Generalised Hough Transform (GHT) from image processing [8], which similarly learns the distribution of codebook activations.

The robustness of the proposed system is achieved through a sparse spike coding of the acoustic signature of the sound event in the spectrogram. As opposed to conventional frame-based features, we base the spike coding on local features from the spectrogram, an idea which has proved successful in our previous work on detecting overlapping sound events [9]. The use of local features is inspired by research into the human auditory system, which suggests that the processing may be based on the partial recognition of features that are local and uncoupled across frequency [10]. For the spectrogram, each local feature then captures a glimpse of the sound information in the spectrogram at a given time and frequency location [11], followed by recognition which proceeds by combining together the local evidence to form a decision. The advantage is that a sound can still be recognised even when a proportion of features is missing or corrupted.

To illustrate the idea, Fig. 1 gives an overview of the proposed system, which shows a bell ringing sound event being encoded and detected. It can be seen how the proposed system consists of two important steps: (1) the time-frequency spectrogram is transformed into a sparse spike coding by detecting local features in the spectrogram and matching them




(a) Spectrogram of Bell sound in 10dB noise, showing the detected keypoints *(white ×)*

(b) Spike distribution within the context window, learnt by the SNN for the Bell sound

(c) Steps in spike coding system

(d) Spike coding of the Bell sound in (a) using k-means codebook

(e) Output voltage from artificial neural network for five different classes over time

**Fig. 1**. Overview of the proposed spiking neural network system. The two key steps are as follows: (1) the spike coding of local features in the spectrogram, shown in (a) and (d); and (2) the spiking neural network with Tempotron cost function, which learns the distribution of the input spikes, shown in (b) and (e).

against a codebook of local feature information; and then (2) the spiking neural network learns the distribution of the input spikes over time and frequency to produce an output spike corresponding to a detection of the target class. These two aspects are described in detail over Sections 2 and 3, before Section 4 examines the performance of the proposed system on a large database of environmental sound events.

## 2. SPIKE CODING APPROACH

The proposed spike coding is generated by first detecting keypoints in the spectrogram, which localise the sparse high-energy peaks in the spectrogram, and then extracting the local spectral information to find the best matching entries in a codebook. The idea is that such peaks will still be present under mismatched noise, and the discriminative local feature information should provide a robust foundation for further processing. The final output can be considered as a novel sparse representation of the spectrogram, since the spike coding preserves the frequency information while outputting a sparse spike coding of the detected local features.

### 2.1. Local Spectrogram Feature Extraction

Starting from a conventional 40-dimension Mel-filtered spectrogram representation, $S(f,t)$, we first detect the high energy peaks which we refer to as "keypoints". To do this, we search for local maxima across the frequency dimension. The local spectral region surrounding the keypoint is then extracted and stored as follows:

$$K(f,t) = \{S(f \pm d_f, t \pm d_t)\} \text{ if } S(f,t) \geq S(f \pm 1, t) \quad (1)$$

where $d_f, d_t = 7$ are the range of the local patch across frequency and time respectively.

A further step is taken to reduce the number of keypoints by introducing a sparsity criterion to reject keypoints that do not represent significant maxima. For this, a threshold $\gamma$ is set based on the mean of the local spectral region. A keypoint is rejected if $S(f,t) - \mu(f,t) < \gamma$, where $\mu$ is calculated as:

$$\mu(f,t) = \text{mean}\left[K(f,t)\right]. \quad (2)$$

In our experiments, we fix $\gamma = 1$, although we found that the performance was not highly sensitive to this parameter.

### 2.2. Robust Spike Coding

To produce a spike coding of this representation, a codebook dictionary is first generated that contains entries representing the variation of local feature information. Here we use k-means clustering to produce the codebook, $C$. However other approaches are possible, such as sparse coding [12]. For the

each entry where $z = 1 \dots Z$ in the codebook, $C_z$, the Euclidean distance is used as the similarity measure, with a local missing feature mask introduced to reduce the influence of the noise on the codebook matching. The distance between the codebook and a local feature at position $(f, t)$ is therefore calculated as follows:

$$y_z = \parallel C_{z_r} - K(f,t)_r \parallel \tag{3}$$

where the reliable local feature indices, $r$, are calculated as $r = K(f,t) > \mu(f,t)$, and $\mu(f,t)$ is the local mean (2).

The best matching codebook entry then generates a spike in the output "spatio-temporal" pattern, $P(x,t)$, where $x = (f-1) \times Z + z_{best}$ is the unique position for the given frequency and codebook activation, with the best matching unit (BMU) $z_{best}$ calculated as:

$$z_{best} = \arg\max_z y_z. \tag{4}$$

This is repeated a number of times, $nBMU = 3$, to allow for several similar codebook entries to be activated. This was found to produce a more robust spike coding pattern.

## 3. NEURAL SPIKE RECOGNITION SYSTEM

To recognise the spatio-temporal spike coding, we propose to use a neural network that can learn the mapping between the input features and output target using back propagation. In particular, we want to encourage the network to learn weights that represent the distribution of the information in the sparse input spike coding. The network can then perform a function analogous to a generalised Hough transform (GHT) [8], which similarly learns the distribution of local feature codebook matches [13]. The advantage of the SNN approach is that it becomes possible for the spike distribution to be learnt discriminatively by adjusting the weights to directly optimise the mapping.

Here we propose a SNN structure to achieve this mapping through an innovative combination of both spiking and conventional neural network architectures. The chosen SNN system is the Tempotron [6], which is a biologically plausible architecture that learns to produce an output spike representing a detection of a given class. The network proceeds by first passing the input spike coding through a integration kernel, $\Omega$, which captures the temporal distribution of spikes using a leak integrate-and-fire neuron model, as follows:

$$V(t) = \sum_x \omega_x (P(x,t) * \Omega) \tag{5}$$

which is the weighted summation over the convolution between the input spike pattern, $P(x,t)$ and $\Omega$, which is defined as follows:

$$\Omega(t) = V_0 \left( exp\left[-t/\tau\right] - exp\left[-t/\tau_s\right] \right) \tag{6}$$

where $\tau, \tau_s$ are decay time constants and $V_0$ normalises the kernel to have a maximum of 1. The cost function is defined based on the difference $V_{thr} - V(t_{\max})$, such that weight modification is only required for erroneous patterns at time $t_{\max}$. Erroneous patterns are defined as positive patterns with no spike produced, or negative patterns producing a spike.

In this work we remove the fixed distribution function of the Tempotron in (6), and instead expand the weight matrix to take an input context window of frames, as follows:

$$V(t) = \sum_{t'} \sum_x \omega_{x,t'} P(x, t - t') \tag{7}$$

where $t'$ represents the relative time within a context window. This allows the spiking neural network to discriminatively learn the distribution of the input spike coding within the context window from the data.

While this is similar to the context window used in DNN systems, here the distribution is stored with the temporal information captured within the context window relative to the output spike produced by the Tempotron network at time $t_{\max}$. This allows much sharper and more discriminative distribution weights to be learnt compared to performing classification of every frame against a hard pre-defined label. In addition, the classification decision is based on the output voltage over the segment, with the maximum spike voltage used as a measure for classification, as opposed to the frame-based probability distribution output by the DNN.

## 4. EXPERIMENTS

In this section, experiments are conducted to analyse the performance of the proposed spiking neural network system on a database containing a large number of environmental sounds.

**Sound Database:** A total of 50 sound classes are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [14], giving a selection of collision, action and characteristics sounds. The isolated sound event samples are around 0.5-3s in duration, have a high signal-to-noise ratio (SNR), and are balanced with silence either side of the sound. The selected categories cover a wide range of sound events, including wooden, metal and china impacts, friction sounds, and others such as bells, phones, and whistles. For each event, 50 files are randomly selected for training and another 30 for testing. The total number of samples are therefore 2500 and 1500 respectively.

**Noise Conditions:** For each experiment the classification accuracy is investigated in mismatched conditions, with most systems using only clean samples for training. The average performance for each method is reported in clean and at 20, 10 and 0 dB SNR for the "Speech Babble" noise environment obtained from the NOISEX'92 database [15]. For multi-conditional training, three different noise types, "Destroyer Control Room", "Factory Floor" and "Jet Cockpit", from the NOISEX'92 database are added at 10dB SNR to the training.

| Method | Training Data | Input Features | Hidden Layers | Cost Function | Clean | 20dB | 10dB | 0dB | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| DNN | Multi | Mel-Spectrum | 4 hidden | Cross Entropy | **99.5** | 96.2 | 84.4 | 46.5 | 81.7 |
| CNN | | | 1 conv + 3 hidden | | 97.3 | 93.1 | 86.7 | 50.9 | 82.0 |
| **Proposed** | Clean | **Spike Coding** | 1 hidden | **Tempotron** | 98.1 | 58.1 | 26.5 | 10.2 | 48.2 |
| | | | **no hidden** | | 99.1 | **98.9** | **96.5** | **82.1** | **94.1** |
| | | | 1 hidden | | 99.3 | 96.8 | 93.7 | 77.0 | 91.7 |
| Bio-NN | | | no hidden | | 97.0 | 95.8 | 91.8 | 75.5 | 90.0 |
| DNN | | | 1 hidden | Cross Entropy | 96.3 | 92.5 | 86.3 | 70.0 | 86.3 |

**Table 1**. Experimental results comparing the classification accuracy of the baseline and proposed SNN system in clean and mismatched babble noise at 20/10/0dB SNR. *Note:* Bio-NN = Tempotron system using a fixed biologically-inspired distribution.

**Experimental Setup:** The experiments are designed to examine the performance of both the spike coding and neural spiking output aspects that contribute to the proposed spiking neural network system. Hence we primarily compare the results against competing artificial neural network systems, using different network structures and cost functions. Therefore we compare baseline deep/convolutional neural network (DNN/CNN) systems with multi-conditional training to improve the results in mismatched conditions. These networks use the cross-entropy cost function, and the best performance found using a context of 9 frames of 40 Mel-filterbank features as input.

The proposed spike coding has 25 codebook entries, occurring over the 40 frequency dimensions, giving each frame an input size of 1000 dimensions. Our proposed Tempotron network structure takes a context window over 25 frames of the input spike coding to learn a set of spike distribution functions, while the original Tempotron with fixed distribution function is denoted "Bio-NN". Across all systems, hidden layers have 1000 neurons, layers are added with layer-wise pretraining, and in the winning class is selected as the one that most strongly associates with the given input.

**Results and Discussion:** The results are shown in Table 1, where it can be seen that the best performing system combines the sparse input spike coding with the Tempotron cost function to give the proposed spiking neural network system. This method achieves 94.1% averaged over the four noise conditions, which is a strong result considering that only clean samples are required for training. This compares well to the multi-conditional DNN and CNN baselines, which both achieve around 82% when trained on conventional Mel-spectrum features. The results also break down the contributions from the various aspects of the system, as follows:

*Spike Coding vs. Mel-Spectrum:* while the Mel-spectrum performs well in clean conditions it is not robust to noise. However, the proposed spike coding only generates spikes at keypoints that are detected at sparse peaks in the signal. These areas carry the most robust and discriminative information, hence the spike coding system performs well even in mismatched conditions.

*Tempotron vs. Cross Entropy:* unlike the cross-entropy cost function, which attempts to classify each frame according to the training label, the Tempotron cost function is focussed on generating a spike when a given input pattern is detected. This allows it to learn a more precise mapping to capture the distribution of the input features surrounding the output spike, and does not require an accurate frame-level label for training. Table 1 shows that for the same spike coding input features, the DNN with cross entropy cost function performs consistently worse across the noise conditions.

*Learned vs. Bio-inspired (Bio-NN) Distribution:* the original Tempotron used a leaky integrate-and-fire neuron model, which is equivalent to assuming a fixed distribution of the input spike features. By removing this constraint, the performance of the system improves from 90.0% for the "bio-NN" to 94.1% for the proposed system. This highlights the importance of the time-frequency information captured by the spike coding features that can be easily learnt by the spiking artificial neural network proposed in this work.

## 5. CONCLUSION

This paper proposes a spiking neural network system for robust sound event recognition that combines a sparse spike coding of the spectrogram with the Tempotron cost function. The spike coding uses a keypoint detection step, such that spikes are only generated on the most robust and discriminative information in the spectrogram. The SNN can then learn the distribution of the these spikes to capture the sound event information over time and frequency. Unlike the conventional cross-entropy cost function, our system uses the Tempotron cost function, which produces a spiking output to indicate when the target class is detected. This enables it to learn a more precise mapping between the input spike coding and the output, which contributes to the strong experimental results, giving a significant improvement over even the multi-conditional DNN baseline.

## 6. REFERENCES


[1] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2007.

[2] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2011, pp. 337–340.

[3] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[4] B. Ghoraani and S. Krishnan, "Time-Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[5] R. F. Lyon, "Machine Hearing: An Emerging Field," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131–139, Sep. 2010.

[6] R. Gütig and H. Sompolinsky, "The tempotron: a neuron that learns spike timing-based decisions," *Nature Neuroscience*, vol. 9, no. 3, pp. 420–428, Feb. 2006.

[7] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, "Temporal Coding of Local Spectrogram Features for Robust Sound Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[8] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[9] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, Jul. 2013.

[10] J. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[11] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[12] O. Vinyals and L. Deng, "Are sparse representations rich enough for acoustic modeling?" in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[13] B. Leibe, A. Leonardis, and B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, Nov. 2008.

[14] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of the International Conference on Language Resources and Evaluation*, vol. 2, 2000, pp. 965–968.

[15] A. Varga and H. J. M. 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.