Check for updates

Event-driven spectrotemporal feature extraction and classification using a silicon cochlea model

Ying Xu*, Samalika Perera, Yeshwanth Bethi, Saeed Afshar and
André van Schaik

International Centre for Neuromorphic Systems, The MARCS Institute for Brain, Behavior,
and Development, Western Sydney University, Kingswood, NSW, Australia

This paper presents a reconfigurable digital implementation of an event-based binaural cochlear system on a Field Programmable Gate Array (FPGA). It consists of a pair of the Cascade of Asymmetric Resonators with Fast Acting Compression (CAR-FAC) cochlea models and leaky integrate-and-fire (LIF) neurons. Additionally, we propose an event-driven SpectroTemporal Receptive Field (STRF) Feature Extraction using Adaptive Selection Thresholds (FEAST). It is tested on the TIDIGTIS benchmark and compared with current event-based auditory signal processing approaches and neural networks.

1. Introduction

In the human auditory pathway, information is extracted and conveyed through sequences of action potentials, or spikes. The spike streams form robust representations that are important for perception. The human sensory system achieves real-time, low-power, and noise-robust performance while operating in such an asynchronous "event"-based way. To mimic the efficiency of signal processing in the human auditory system, biologically inspired auditory sensors and algorithms have been implemented and investigated. For example, Liu et al. (2010, 2014) developed a $2 \times 64 \times 4$ channel dynamic audio sensor that used an analog cascade filter bank and pulse-frequency modulated circuits to emulate the peripheral auditory system and auditory nerve to generate spike streams; Yang et al. (2016) used a synchronised delta modulator to generate audio events; Singh et al. (2018) developed a digital multi-rate cochlea model on Field Programmable Gate Array (FPGA) where a digital leaky integrate-and-fire (LIF) neuron model with different thresholds was used to model auditory neurons of the human auditory system with different thresholds.

Such neuromorphic auditory sensors encode acoustic information into spikes in real-time at a low data rate, which makes them ideal solutions for real-world applications. In recent decades, efforts have been made to investigate neuromorphic sensing approaches to extract acoustic features from auditory spikes. For example, it has been argued that statistical features embedded in spike streams could be the mechanism for the precise encoding of auditory cues that are important for recognition (Gerstner and Kistler, 2002). Therefore rate-code based features (Neil and Liu, 2016), inter-spike interval distributions (Uysal et al., 2006;

Li et al., 2012) inter-spike velocity (Chakrabartty and Liu, 2010), and exponential features (Anumula et al., 2018) have all been investigated in speaker identification and speech recognition tasks. Rasetto et al. (2021) proposed a feature extraction approach to extract spectrotemporal features from a cochlea model built with "event"-based filters for a command recognition task.

In addition to neuromorphic auditory data processing, event-driven feature extraction algorithms have been more widely investigated in neuromorphic vision systems. With the increase in the adoption of neuromorphic vision sensors, various dense tensor representations for the sparse asynchronous event data have been proposed and investigated to learn the spatiotemporal features (Maqueda et al., 2018; Afshar et al., 2020b; Baldwin et al., 2022).

In Cohen et al. (2019) and Afshar et al. (2020b), the event-based time surface representations for event-based vision data have been used in extracting features for a range of tasks, such as object recognition on unmanned aerial vehicles (UAVs) (Zappa et al., 2020) and single photon avalanche diode (SPAD) sensors data processing (Afshar et al., 2020a).

In Afshar et al. (2020c), Feature Extraction using Adaptive Selection Thresholds (FEAST) was proposed for event-based vision data using the time surfaces representation. The FEAST method has been investigated for a range of applications such as object tracking (Ralph et al., 2022), event-based supervised learning (Bethi et al., 2022) and activity-driven adaptation in spiking neural networks (SNNs) (Haessig et al., 2020).

To investigate spectrotemporal representations for event-based auditory data, in Xu (2019), the FEAST method was investigated in audio for an isolated spoken digits recognition task and showed improved performance. In this work, we extend the work and propose to use FEAST to build a computational auditory cortical model – the Spectrotemporal Receptive Field (STRF) model. The proposed event driven STRF approach is applied to the binaural cochlear system for a multi-resolution spectrotemporal analysis.

2. Background and related work

2.1. The event-based binaural CAR-FAC system on FPGA

In the previous work, we implemented a digital cochlea model, the Cascade of Asymmetric Resonators with Fast Acting Compression (CAR-FAC) cochlea model (Lyon, 2017) on a FPGA for sound localisation (Xu et al., 2021). This model approximates the physiological elements that make up the human cochlea, including the basilar membrane (BM), the outer hair cells (OHCs) and the inner hair cells (IHCs), as shown in **Figure 1**, and mimics its qualitative behaviour. The digital cochlea is reconfigurable in filter parameters and channel numbers. This work extends the cochlea model to an event-based binaural cochlear system. It includes a CAR-FAC cochlea pair and LIF neurons to generate auditory spike streams.

The architecture of the event-based binaural cochlear system is shown in **Figure 2**. Each "ear" in the system implements

the components of the CAR, the digital OHC (DOHC), the digital IHC (DIHC), the automatic gain control (AGC), the lateral inhibition (LI), and the LIF neuron. One "ear" can be switched off so that the system operates as a single CAR-FAC model. The FAC part that introduces nonlinearities can also be switched off so that the system operates as a linear CAR model. The details of the CAR-FAC module were described in Xu et al. (2016), Xu et al. (2018b), and (Xu et al., 2019).

2.2. Feature Extraction using Adaptive Selection Thresholds (FEAST)

The FEAST method in Afshar et al. (2020c) extracts spatiotemporal features for event-based vision data using real-valued exponentially decaying kernels and 2-D "neurons." The use of exponentially decaying kernels for event-based processing was described in Tapson et al. (2015) and called a "time surface" in Lagorce et al. (2015). The time surface is generated by applying an exponential decay with a time constant τ_v on a local (typically square) neighborhood centered on the current event. For example, an event e_i occurred at time t :

$$e_i = [x, y, t, p]^T \tag{1}$$

Where x, y represent the spatial location of the pixel with reference to the event-based sensor and $p \in \{-1, 1\}$ is the polarity of the event.

The time surface $S_i(\mathbf{u}, p)$ at the location ($\boldsymbol{u} = [\boldsymbol{x_i}, \boldsymbol{y_i}]^T$) of the event e_i at time t can be calculated as:

$$S_i(\mathbf{u}, p) = e^{-\frac{t_i - \Gamma_i(\mathbf{u}, p)}{\tau_v}} \tag{2}$$

Where $\Gamma_i(\mathbf{u}, p)$ is the timestamp of the latest event that occurred at the location u. The time surface of a pixel in the spatial neighborhood of size R around an event location is considered as an event context (E_C) [of size $(2 \times R + 1) \times (2 \times R + 1)$]. In event-based vision, an E_C represents the recent time history of events in the spatial neighborhood in 2-D.

Feature Extraction using Adaptive Selection Thresholds learns spatiotemporal features from the E_Cs through 2-D "neurons." Each neuron has randomly generated initial threshold and weights. The neurons act as feature extractors with individual adaptive thresholds via a competitive strategy. A similarity measure, $\cos(\theta)$, between the E_C and the neuron's weights is used as a metric to match the event contexts with the weights of each neuron:

$$similarity = \cos(\theta) = \frac{E\_C \cdot W_i}{||E\_C|| \cdot ||W_i||} \tag{3}$$

Where W_i denotes the weights of neuron i, and $\frac{E\_C}{||E\_C||}$ and $\frac{W_i}{||W_i||}$ are the normalisations of E_C and the weights. After normalisation, the similarity is calculated as a dot product of normalised E_C and weights.

In the learning phase, each neuron's unique threshold acts as a selection boundary. The neuron with the highest similarity that also crosses its selection threshold is picked as the winner neuron, which
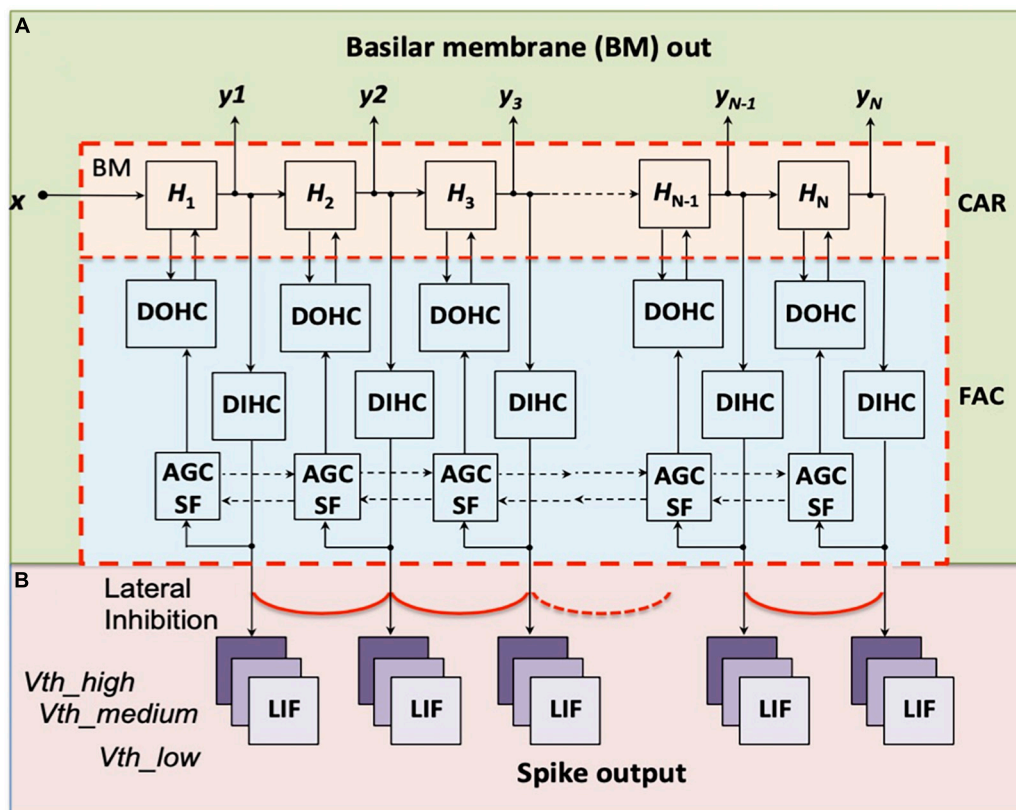
FIGURE 1
Structure of the CAR-FAC model **(A)** and the LIF neurons preceded by lateral inhibition **(B)**. x is the input sound, H_1 to H_N are the transfer functions of the CAR part, and y_1 to y_N represent the CAR-FAC Basilar Membrane (BM) output. The characteristic frequencies (CFs) of the CAR resonators decrease from left to right. The DOHC, the DIHC and the AGC loop comprise the FAC part. Each DIHC is connected to nine LIF neurons with three thresholds, *Vth_low*, *Vth_medium,* and *Vth_high*, after a Lateral Inhibition function between neighboring channels.

then emits a spike. The thresholds of the neurons are dynamic during learning, and are adapted based on two rules:

1. If there is a winner, then increase the threshold Vth_i of neuron i by a fixed amount ΔI.
2. If there is no winner, then decrease all the neurons' thresholds by a fixed amount of ΔE.

The E_C is then used to update the winning neuron's weights with a fixed mixing rate as follows:

$$W_{i\_update} = (1 - \eta) \times W_i + \eta \times E\_C \qquad (4)$$

Where the weights W_i of neuron i is updated according to the E_C that is successfully matched, and η is the mixing rate used to update the weights of the neuron. The weights of the neurons form features that cover the feature space of the input signals.

The use of a dynamic threshold Vth_i ensures that the rate of firing of all neurons is approximately equal across the dataset, as increasing the threshold on the matching feature serves to specialise each neuron from other neurons. If the weights are coding poorly for the incoming feature, then the global threshold decrease serves to expand the range of input features to which the neurons will respond. This learning process is dynamic and responsive to the statistics of the incoming data.

3. Materials and methods

3.1. The event-based binaural CAR-FAC system on FPGA

In this work, we implement the LIF neuron using:

$$LIF[s, t] = LIF[s, t-1] + c_{LIF} \times (IHC_{LI[s,t]} - LIF[s, t-1]) \qquad (5)$$

$$c_{LIF} = 1/(f_s \times \tau_{LIF}) \qquad (6)$$

Where $LIF[s, t]$ is the membrane potential of a neuron connecting to the CAR-FAC LI output, $IHC_{LI[s,t]}$ in channel s at time t, f_s is the sampling frequency, and τ_{LIF} is the time constant of the LIF neuron. When $LIF[s, t]$ is above a *threshold*, a spike, $spk[s, t]$, is generated at time t in channel s:

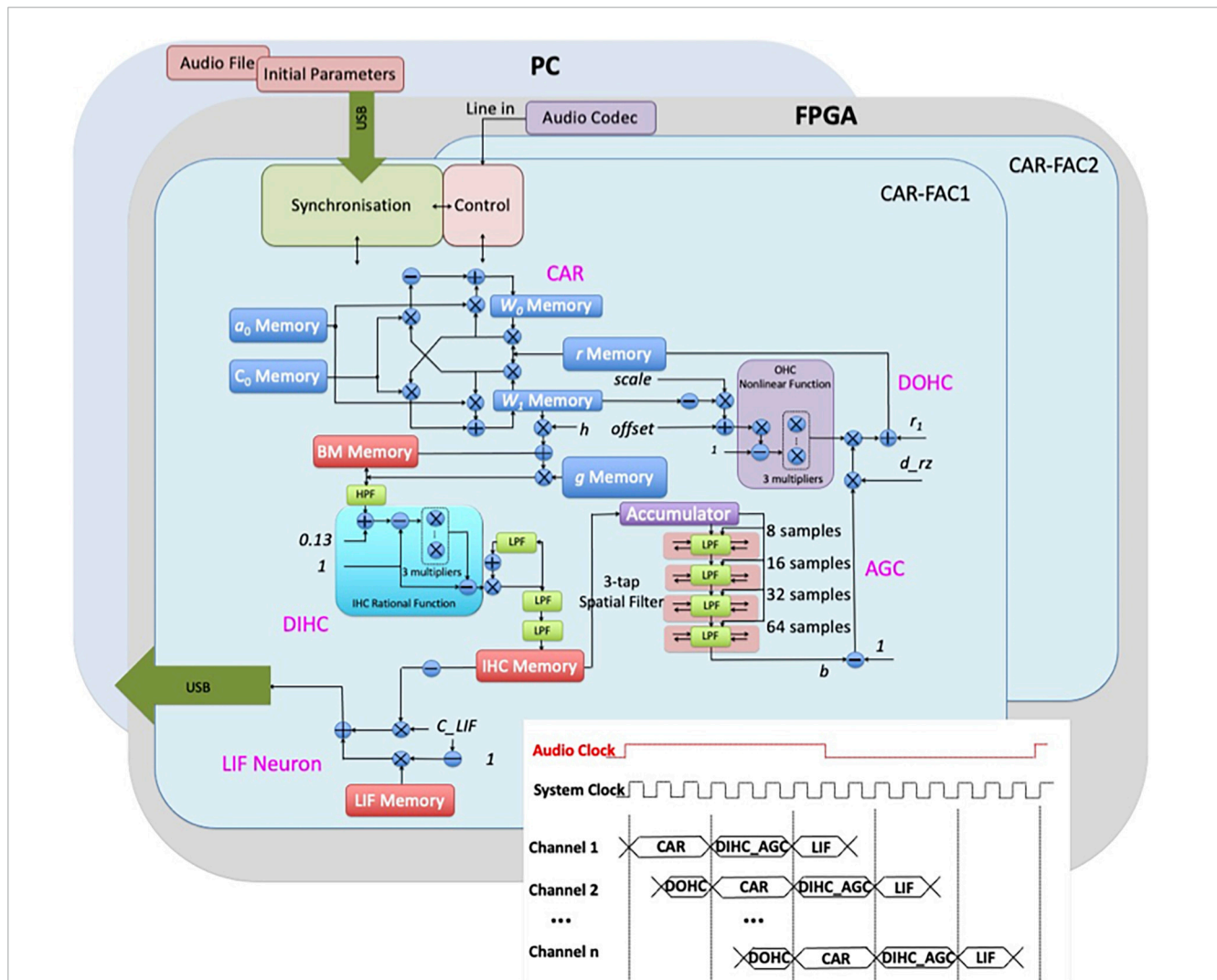$$\text{if } LIF[s, t] > threshold :$$
$$spk[s, t] = 1,$$
$$LIF[s, t] = V_{reset}$$
$$\text{else:}$$
$$spk[s, t] = 0 \qquad (7)$$

FIGURE 2
Architecture of the binaural CAR-FAC FPGA system. The system consists of an audio codec and two "ears." Each of the ears includes a CAR-FAC module, a controller module, and an interface module. The FPGA board is hosted by a PC through a USB interface. The inset shows the system timing diagram demonstrating the pipelined CAR-FAC. With time multiplexing and pipeline techniques, a binaural real-time n-channel CAR-FAC system is built using only one CAR-FAC module and one LIF module for each ear.

and the membrane potential of the neuron is reset to value V_{reset}. The generated spike streams encode the amplitude of each channel response that is used in the following feature extraction.

## 3.2. Unsupervised feature extraction

In this work, we modify the FEAST and apply it on audio data to build an event driven STRF model. When the FEAST is applied to the event-based audio data, the E_C needs to be formed differently. **Figure 3** shows the construction of the E_Cs and the details will be illustrated in the next two sections. In this paper, we use the FEAST to build the event-based multi-resolution spectrotemporal analysis. The computational SpectroTemporal Receptive Field (STRF) model is inspired by psychoacoustical and neurophysiological findings in the early and central stages of the auditory system (Chi et al., 2005). The model provides a unified multi-resolution representation of the spectral and temporal

features likely critical in the perception of sound. It mimics aspects of the responses of higher central auditory stages, especially the primary auditory cortex. Functionally, it estimates the spectral and temporal modulation content of the auditory spectrogram via a bank of filters that are selective to different spectrotemporal modulation parameters ranging from slow to fast rates temporally, and from narrow to broad scales spectrally (Chi et al., 2005). Here we break the proposed event-based spectrotemporal feature extraction into two steps.

3.2.1. Temporal feature extraction – 1-D FEAST

The CAR-FAC model shows highly frequency-dependent gains and the connecting LIF neurons encode the amplitude of the channel responses. A similar amplitude coding is also used by Liu et al. (2014). **Figure 4** show the CAR-FAC response to a TIDIGITS utterance "o." In the middle frequency channel, 650 Hz, the response shows the highest gain in amplitude, and thus higher spike numbers than the higher (1000 Hz) and lower
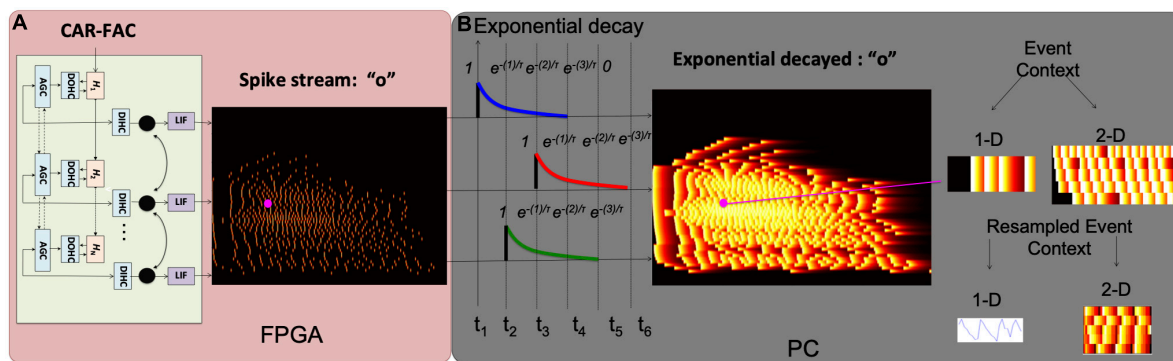
FIGURE 3
Construction of the auditory time surface window. **(A)** Spike streams generated from the binaural CAR-FAC system on FPGA; **(B)** when a spike occurs in a frequency channel, the value decays exponentially as t increases, forming the exponential decaying feature. A 1-D/2-D event context (E_C) is generated on a local neighborhood centered on the current event (in magenta). The 1-D E_C provides temporal feature extraction, and the 2-D E_C provides spectrotemporal feature extraction. The E_Cs are resampled to form a uniform size for all the events as described in the text.
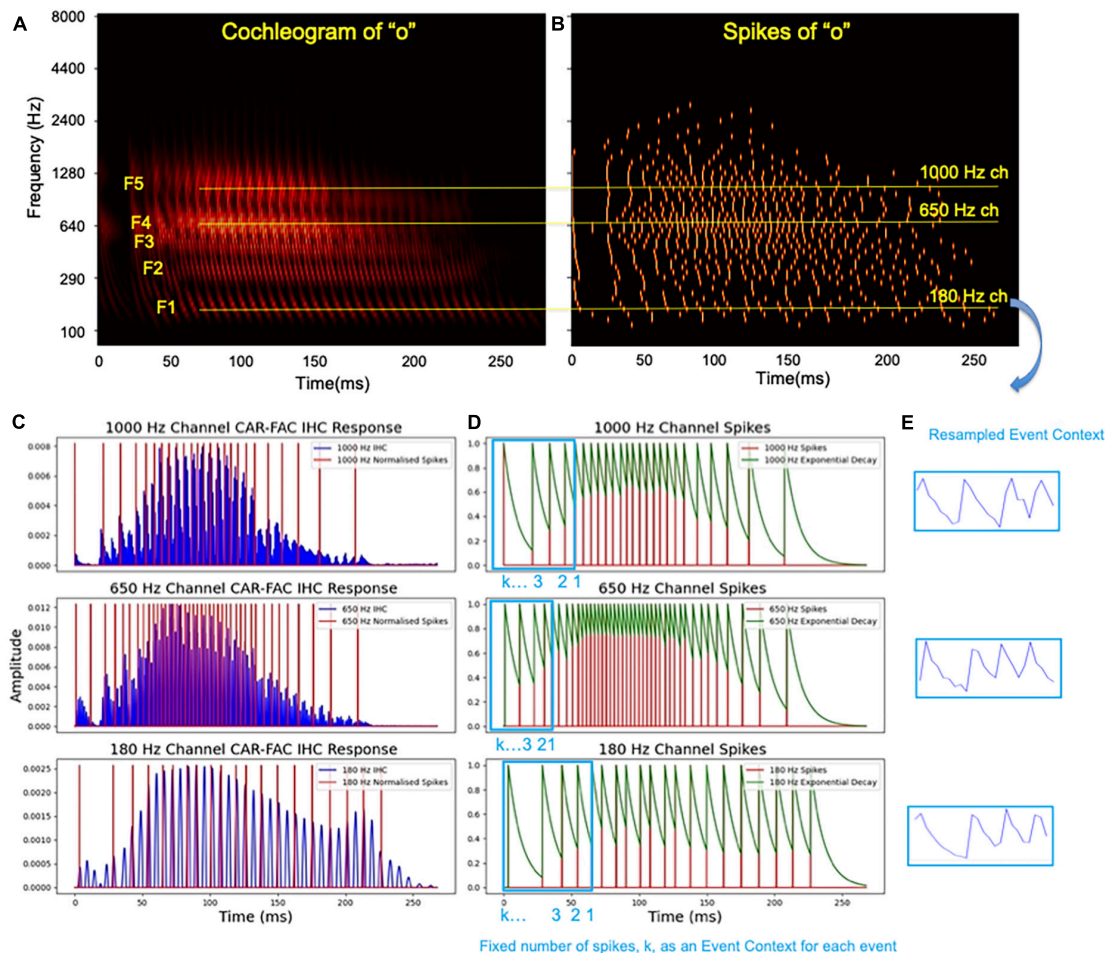
FIGURE 4
The CAR-FAC output of utterance "o," **(A)** formants (F1-F5) are labeled and three channels, 180 Hz, 650 Hz, and 1000 Hz, are highlighted in yellow; **(B)** the spikes of utterance "o"; **(C)** the IHC output from CAR-FAC and the generated spikes encoding amplitude of the three highlighted channels; **(D)** the time surfaces in these channels; **(E)** the resampled E_Cs that preserves temporal information.

frequency channels (180 Hz). Additionally, the inter-spike interval encodes the changes in amplitude. For example, for an increment in amplitude, the spike train shows a gradually decreasing inter-spike interval, whereas, for a decrease in amplitude, the spike train shows a gradually increasing inter-spike interval. In this way, the spike trains of each channel encode syllabic
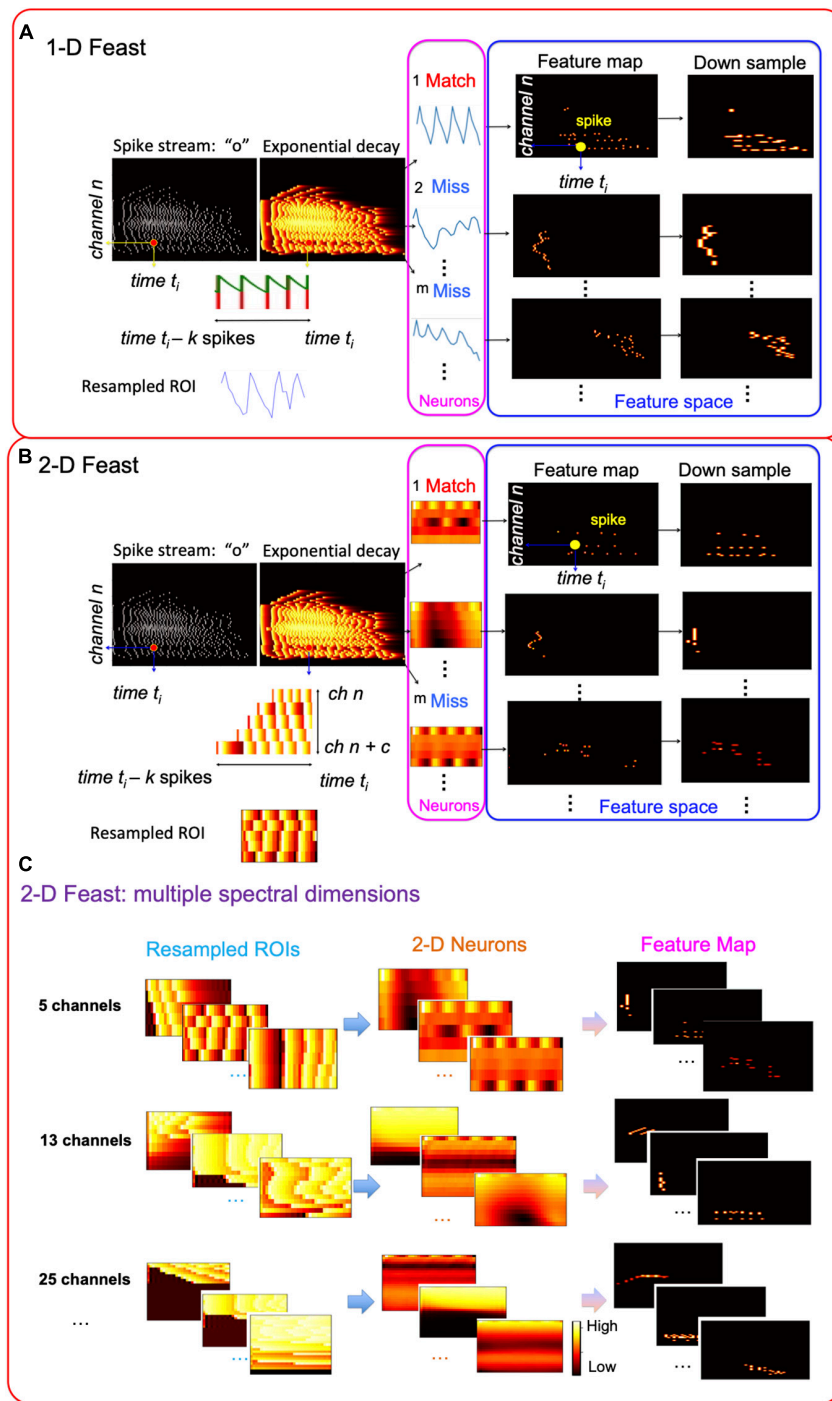
FIGURE 5
Event-based spectrotemporal feature extraction: **(A)** 1-D FEAST; **(B)** 2-D FEAST; **(C)** 2-D FEAST on multiple spectral dimensions.

rates of speech. In speech and music, there are three kinds of temporal modulations (Chi et al., 2005) in the cochlear outputs: Slow modulations that reflect the syllabic rates of speech, which are superimposed upon the intermediate rate modulations due to inter-harmonic interactions occurring at a rate that reflects the fundamental frequency of the input, which in turn are riding upon the fast frequency component driving this channel best, the characteristic frequencies (CF) of each cochlear channel.

The first step of Event-based SpectroTemporal Feature Extraction, 1-D FEAST, is to extract such syllabic rates, or slow amplitude changes, from each cochlear channel temporally:

As shown in **Figures 4C, D**, we apply an exponential kernel decaying with a time constant τ on each event across the channels:

$$\tau = f_s \times 10^{-3} \tag{8}$$

Where f_s is the sampling frequency, and τ determines a duration over which the previous event has an impact on the

	Used	Available	Utilization
ALM	13,958	29,080	48%
Memory (bits)	1,278,771	4,567,040	28%
DSPs	123	150	82%

scene, and the current event represents the highest energy, 1.0, as shown in **Figure 4D**. We then define a 1-D E_C for each event that includes a fixed number of spikes. Each E_C should include a sufficient number of spikes such that a change in amplitude can be represented. Since the E_C generated for each spike has a different duration in time, or number of samples, we then resample the E_C into a fixed number of samples. After resampling, all the E_Cs have a same number of samples, while preserving the encoded temporal features. For example, in **Figure 4E**, an onset is shown in five consecutive spikes with gradually increased inter-spike intervals in all the channels.

FEAST is then applied to the 1-D E_Cs to extract 1-D temporal features, as shown in **Figure 5A**, in two phases:

(a) Learning: In the learning phase, the number of the neurons, m, is pre-set, and the initial threshold and weights for each neuron are randomly generated. For each event, we choose k spikes in the past that are the closest to it and resample it to form its E_C. All the extracted E_Cs are presented in random order during training.

For an event at time t_i and *channel n*, the dot product between its E_C and each neuron is calculated. The only neuron with the largest value which is also above its threshold is the winner. The threshold of the winner neuron is then increased by ΔI, and the weights are updated according to (7). If there is no winner, all the neurons' thresholds are then decreased by a fixed amount of ΔE. Multiple epochs of learning are performed empirically until it is converged.

(b) Feature extraction: Once the system is no longer learning, the m neurons are then used to extract features from spike streams. Each neuron generates a feature map in its feature space: For an event at time t_i in channel n, the dot product between its E_C and each neuron is calculated. The only neuron with the largest value is the winner. The winner neuron will emit a spike at time t_i in channel n in its feature space to form a feature map.

The 1-D FEAST extracts channel-wised temporal features, in particular the slow changes in amplitude encoded in the spike streams. It is comparable to the computational spectrotemporal cortical model (Chi et al., 2005) that uses slow rate filters for the temporal analysis to extract syllabic rates in speech.

3.2.2. Spectro-temporal feature extraction − 2-D FEAST

Speech contains spectral modulations created by harmonics and formants, which are also evident in the cochleogram.
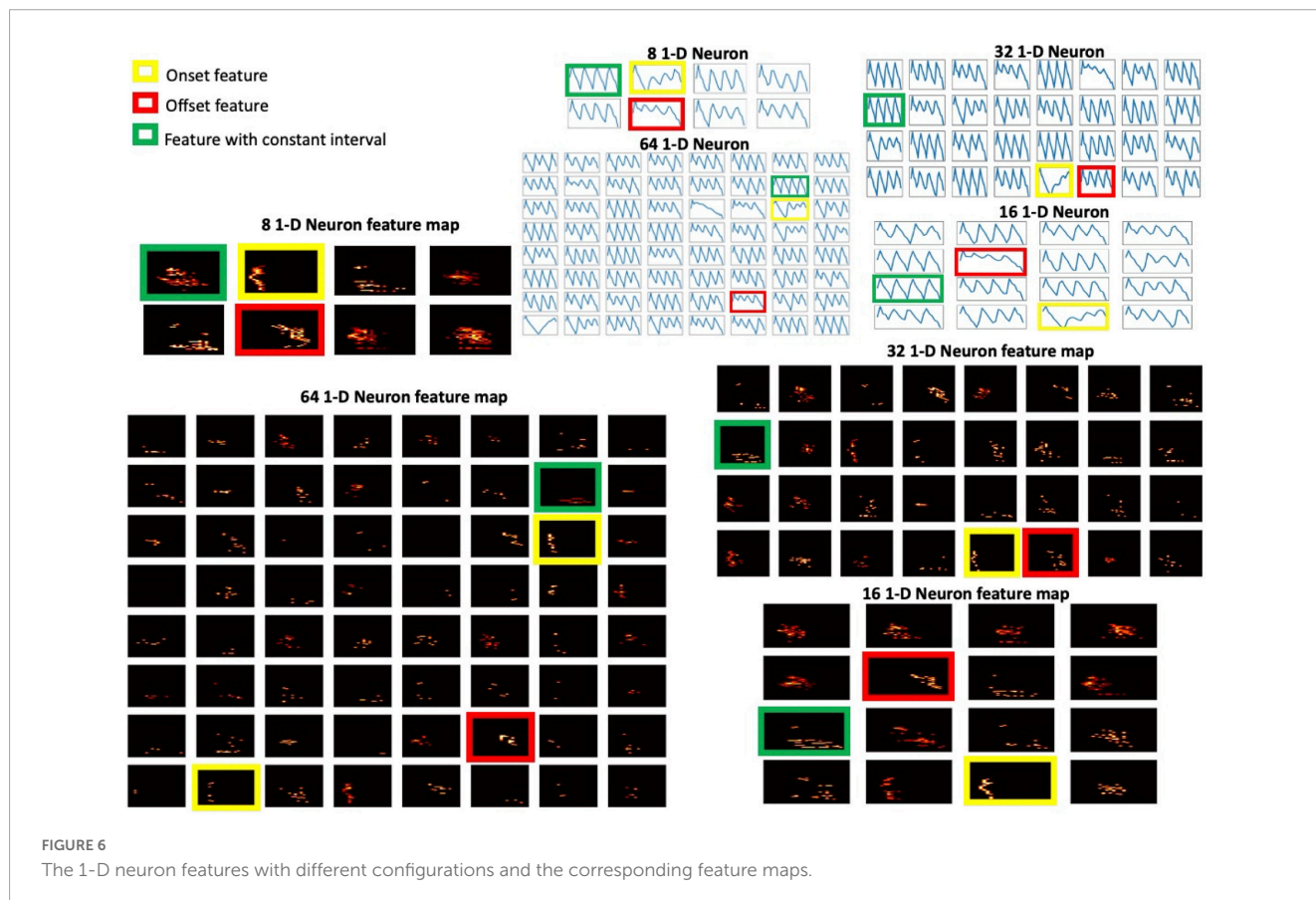

FIGURE 6
The 1-D neuron features with different configurations and the corresponding feature maps.

TABLE 2 Summary of investigated features on the TIDIGITS dataset.

Pre-processing	Method	Neuron number	Accuracy
	Time-binned feature + SVM (baseline)	NA	**88.49%**
	1-D FEAST + SVM	8	**92.96%**
		16	**93.56%**
		32	**93.92%**
		64	**93.64%**
CAR-FAC	2-D FEAST + SVM	16 (5ch)	**93.56%**
		16 (13ch)	**95.78%**
		16 (25ch)	**96.34%**
		16 (37ch)	**95.65%**
		16 × 4 (5+13+25+37ch)	**97.47%**
		32 (5ch)	**95.01%**
		32 (13ch)	**96.62%**
		32 (25ch)	**97.42%**
		32 (37ch)	**96.70%**
		32 × 4 (5+13+25+37ch)	**97.71%**
		64 (5ch)	**94.97%**
		64 (13ch)	**96.13%**
		64 (25ch)	**97.22%**
		64 (37ch)	**96.82%**
		64 × 4 (5+13+25+37ch)	**97.43%**
AMS1b (Anumula et al., 2018)	Constant time bins: exponential features (GRU RNN with Backpropagation)	NA	91.1%
AMS1b	Multilayer Perception (MLP) using Spike Layer Error Reassignment (SLAYER) on Loihi2 (Orchard et al., 2021)	64-256RF-256RF-242LIF	92.14%
MFCC	MFCC-SOM SLAYER with Backpropagation (Shrestha and Orchard, 2018)	484-500-500-11	99.09%

Red means highest accuracy in the experiment. Green mean highest accuracy for each neuron configuration.

Harmonics come from the vocal folds and are considered the source of the sound. Formants come from the vocal tract. Formants filter the harmonic sound source, and thus after harmonics go through the vocal tract, some become louder, and some become softer.

The features of the harmonics/formants are associated with the frequency channels and the next step of the Event-based SpectroTemporal Feature Extraction, 2-D FEAST, is to extract spectral and temporal combined features. As shown in **Figure 5B**, to extend the 1-D E_C, for each event at *time* t_i in *channel n*, we choose c channels in frequency. Within each selected channel, k spikes that are closest to the current event are selected and resampled to form a 2-D E_C. Similar to the 1-D FEAST, in the learning phase, for an event at time t_i and *channel n*, the dot product between its E_C and each 2-D neuron is calculated. The neuron

with the largest value which is also above its threshold is the winner. The threshold of the winner neuron is then increased by ΔI, and the weights are updated according to (7). If there is no winner, all the neurons' thresholds are then decreased by a fixed amount of ΔE. Multiple epochs of learning are performed until the weights have converged. In the feature extraction phase, for each event at time t_i and *channel n*, the dot product between its E_C and each 2-D neuron is calculated. The only neuron with the largest value is the winner. The winner neuron will emit a spike at time t_i in channel n in its feature space to form a feature map.

Furthermore, for each event, we generate multiple sets of 2-D E_Cs, as shown in **Figure 5C**. Each set includes a different number of channels so that it covers multiple scales in frequency. For example, as shown in **Figure 5C**, a 5-channel dimension only includes one harmonic, whereas a 13-channel can cover two harmonics, and so on. The choice of channel numbers is based on the Greenwood function used in the CAR-FAC model (Greenwood, 1990).

$$f = 165.4 \left(10^{2.1x} - 1\right) \tag{9}$$

Where coefficient x is the normalised position along the cochlea, varying from 0 at the apex of the BM, to 1 at the basal end, and coefficient f is the CF of each frequency channel.

We then apply the 2-D FEAST described previously on each set of the E_C in parallel. The same learning and feature extraction phases described above are applied to each event. For each event at time t_i and *channel n*, there is one winner neuron in each dimension. The winner neuron of each dimension will emit a spike at time t_i in channel n in its feature space.

2-D FEAST is comparable to a spectrotemporal cortical model that uses different "seed functions" as scale filters for spectrotemporal analysis (Chi et al., 2005) to extract harmonics and formants.

4. Results

4.1. CAR-FAC on FPGA

The CAR-FAC FPGA implementation has been investigated and measured by Xu et al. (2016) and Xu et al. (2018b). In this work, we use the proposed CAR-FAC cochlear system on FPGA to generate spike streams from the TIDIGITS database. We first simulated the CAR-FAC and the LIF neuron in Python with floating-point numbers. Here τ_{LIF} of the LIF in (2) is set as 10 ms, V_{reset} is set as 0, and *threshold* is only set as a medium value, 0.0004. Next, we verified the model using the fixed-point numbers to determine the required word length for the hardware implementation. We use 20-bit BM variables, 20-bit DOHC variables, 16-bit DIHC variables, and 16-bit LIF neuron variables. Then the system is designed using Verilog HDL with the same word length as the fixed-point computer simulation. There is no difference between the hardware measurements and the fixed-point computer simulation results, and insignificant difference with the floating-point computer version. The device utilisation of the binaural 64 channel CAR-FAC ($2 \times 64 \times 9$) system is shown in **Table 1**.
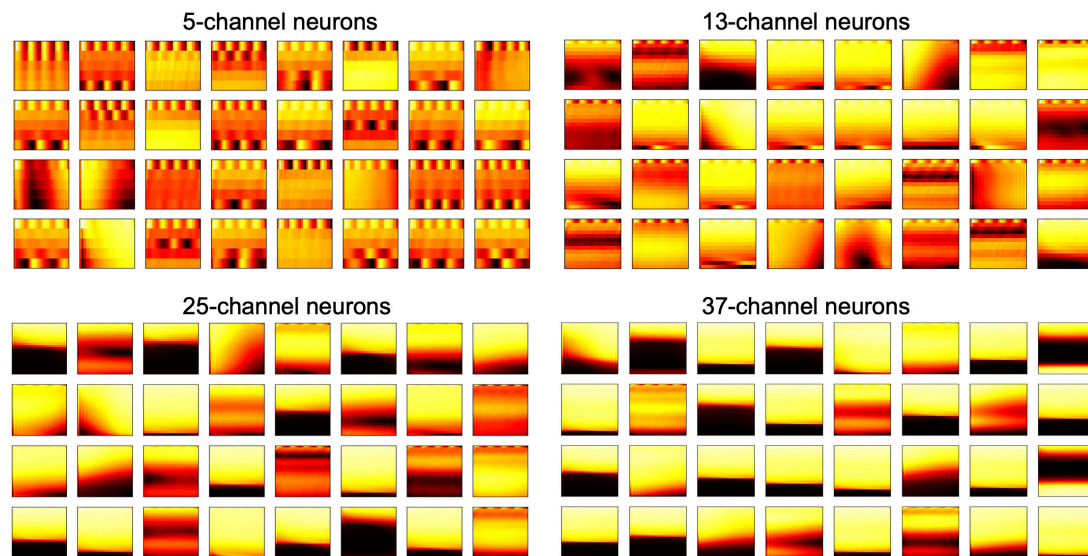
FIGURE 7
32-neuron features for all the different 2-D neuron with different dimensions.

4.2. FEAST and linear classifier

The 1-D and 2-D FEAST are tested, respectively, on an isolated spoken digit recognition task using the TIDIGIT dataset. Here we use the isolated spoken digits (zero to ten) from 225 speakers (female and male) as the training and testing data, of which 4950 samples are included in the total (2,464 for training and 2,486 for testing). The Support Vector Machine (SVM) with linear kernels and optimal regularisation is used as the back-end classifier to investigate the performance of the FEAST.

4.2.1. 1-D feature for temporal feature extraction

In the 1-D FEAST, we chose $k = 4$ spikes for an E_C so that a change in amplitude can be represented sufficiently. The E_C is then resampled into 32 samples. In this experiment, the algorithm had converged after ten epochs of training. The parameters were configured as $\Delta I = 0.001$, $\Delta E = 0.003$, and $\eta = 0.001$ in (7), which were derived empirically. The optimal number of neurons depend greatly on the nature of the data. In this experiment, 8, 16, 32, and 64 neurons are tested. The generated feature map for each neuron is down-sampled via fixed time binning (Anumula et al., 2018), as shown in **Figure 6**. By observing the features of the neurons, we can see the neuron with gradually decreasing intervals often represents an onset, whereas the neuron with gradually increasing intervals represents an offset of an utterance. The evenly distributed intervals represent an unchanged amplitude of the utterance. The generated 1-D features are then used as input for the SVM. Additionally, according to Acharya et al. (2018) and Anumula et al. (2018), the time-binned spikes show the highest accuracy compared to other statistical features in the isolated spoken digit recognition, so in this experiment, the time-binned spikes generated from the proposed cochlear system are investigated as a baseline. The classification results are shown in **Table 2**. For all the configurations, the 1-D FEAST shows better accuracy than the time-binned spikes, and the 32-neuron configuration shows the best accuracy, 93.92%.

4.2.2. 2-D feature for temporal feature extraction

In the 2-D FEAST, we chose $k = 4$ spikes and resample them to 32 samples temporally, and 5, 13, 25, and 37 channels for the 2-D E_Cs. In the training phase, we train each set of the 2-D E_Cs in parallel, using 16, 32 and 64 neurons, respectively. **Figure 7** shows all the features of the 32 neurons and **Figure 8** shows the corresponding feature maps. The small-sized neurons tend to show fine spectral features, whereas the large-sized neurons only show coarse intensity information in **Figure 8**.

We then use the 2-D features as input for the SVM. Firstly, we test each set of neurons separately. As shown in **Table 2**, the 32 neuron system shows the best accuracy, and we found for the same number of neurons, the 25-channel size tends to provide better accuracy. Next, we combine all the sizes together for each neuron configuration, and get an improved accuracy, 97.71%. As comparisons, the results of the same experiment by Anumula et al. (2018) are also shown in **Table 2**, in which a Gated Recurrent Unit (GRU) Recurrent Neural Network (RNN) is used for a constant time binning of the exponential features. In Orchard et al. (2021), a Multilayer Perception (MLP) using Spike Layer Error Reassignment (SLAYER) on Loihi2 is built for the same dataset, and it shows 92.14% accuracy. Currently, the highest accuracy of 99.09% on the same task is achieved by Shrestha and Orchard (2018) using a 484-500-500-11 neuron spiking neural network with backpropagation, whereas in our approach, we only use one-layer of 128 neurons (32 neurons × 4 sizes) and a simple linear classifier.

5. Discussion

This paper presents a reconfigurable digital implementation of an event-based binaural cochlear system and an event-driven spectrotemporal receptive field feature extraction approach. The algorithm is tested on an isolated spoken digit

FIGURE 8
The 32 2-D neurons and down-sampled feature maps.
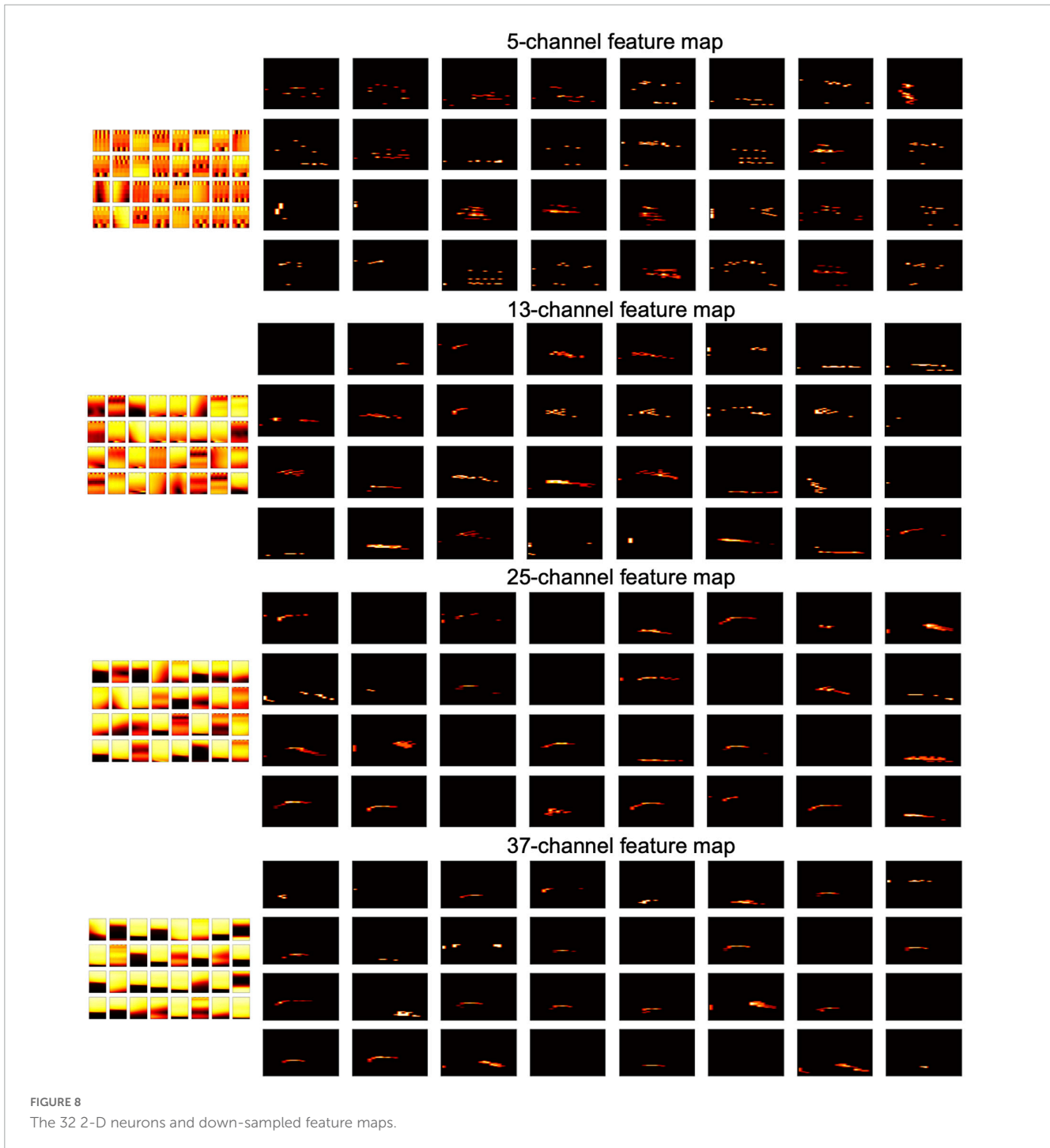
recognition task. The features extracted from FEAST provide better multi-resolution representations of the event-based data than statistical approaches that have been classically used for decoding spike streams.

Like any other data modalities, noise in event-data poses challenges to effective processing and FEAST helps in learning noise-robust features. The CAR-FAC model has been shown to provide noise-robust features in audio to perform speaker identification (Islam et al., 2022). Audio features from the CAR-FAC cochlea model have also been used to perform noise-robust binaural sound localisation (Xu et al., 2018a, 2019, 2021).

Since the FEAST is an unsupervised method, it cannot perform classification and requires a backend classifier. In follow-up work, we will use a generalised model of the FEAST method that performs feature extraction and classification in a single architecture (Bethi et al., 2022).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YX and AS proposed the idea and designed the FPGA. YX and SP recorded the data. YX, SA, YB, and AS evaluated and discussed the results. YX and YB wrote the manuscript. All authors discussed the results, commented on the manuscript, and approved it for publication.

Conflict of interest

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

Publisher's note

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

References

Acharya, J., Patil, A., Li, X., Chen, Y., Liu, S.-C., and Basu, A. (2018). A comparison of low-complexity real-time feature extraction for neuromorphic speech recognition. *Front. Neurosci.* 12:160. doi: 10.3389/fnins.2018.00160

Afshar, S., Nicholson, A. P., Van Schaik, A., and Cohen, G. (2020b). Event-based object detection and tracking for space situational awareness. *IEEE Sens. J.* 20, 15117–15132. doi: 10.1109/JSEN.2020.3009687

Afshar, S., Hamilton, T. J., Davis, L., van Schaik, A., and Delic, D. (2020a). Event-based processing of single photon avalanche diode sensors. *IEEE Sens. J.* 20, 7677–7691. doi: 10.1109/JSEN.2020.2979761

Afshar, S., Ralph, N., Xu, Y., Tapson, J., van Schaik, A., and Cohen, G. (2020c). Event-based feature extraction using adaptive selection thresholds. *Sensors* 20:1600. doi: 10.3390/s20061600

Anumula, J., Neil, D., Delbruck, T., and Liu, S. C. (2018). Feature representations for neuromorphic audio spike streams. *Front. Neurosci.* 12:23. doi: 10.3389/fnins.2018.00023

Baldwin, R., Liu, R., Almatrafi, M. M., Asari, V. K., and Hirakawa, K. (2022). Time-ordered recent event (TORE) volumes for event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 1–14. doi: 10.1109/TPAMI.2022.3172212

Bethi, Y., Xu, Y., Cohen, G., van Schaik, A., and Afshar, S. (2022). An optimized deep spiking neural network architecture without gradients. *IEEE Access* 10, 97912–97929. doi: 10.1109/ACCESS.2022.3200699

Chakrabartty, S., and Liu, S. C. (2010). "Exploiting spike-based dynamics in a silicon cochlea for speaker identification," in *Proceedings of the ISCAS 2010 – 2010 IEEE international symposium on circuits and systems: Nano-bio circuit fabrics and systems*, Paris, 513–516. doi: 10.1109/ISCAS.2010.5537578

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.1945807

Cohen, G., Afshar, S., Morreale, B., Bessell, T., Wabnitz, A., Rutten, M., et al. (2019). Event-based sensing for space situational awareness. *J. Astronaut. Sci.* 66, 125–141. doi: 10.1007/s40295-018-00140-5

Gerstner, W., and Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge: Cambridge University Press. doi: 10.2277/0511075065

Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J. Acoust. Soc. Am.* 87, 2592–2605. doi: 10.1121/1.399052

Haessig, G., Milde, M. B., Aceituno, P. V., Oubari, O., Knight, J. C., van Schaik, A., et al. (2020). Event-based computation for touch localization based on precise spike timing. *Front. Neurosci.* 14:420. doi: 10.3389/fnins.2020.00420

Islam, M. A., Xu, Y., Monk, T., Afshar, S., and van Schaik, A. (2022). Noise-robust text-dependent speaker identification using cochlear models. *J. Acoust. Soc. Am.* 151, 500–516. doi: 10.1121/10.0009314

Lagorce, X., Ieng, S. H., Clady, X., Pfeiffer, M., and Benosman, R. B. (2015). Spatiotemporal features for asynchronous event-based data. *Front. Neurosci.* 9:46. doi: 10.3389/fnins.2015.00046

Li, C. H., Delbruck, T., and Liu, S. C. (2012). "Real-time speaker identification using the AEREAR2 event-based silicon cochlea," in *Proceedings of the ISCAS 2012 – 2012 IEEE international symposium on circuits and systems*, Seoul, 1159–1162. doi: 10.1109/ISCAS.2012.6271438

Liu, S. C., van Schaik, A., Minch, B. A., and Delbruck, T. (2010). "Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms," in *Proceedings of the ISCAS 2010 – 2010 IEEE international symposium on circuits and systems: Nano-bio circuit fabrics and systems*, Paris. doi: 10.1109/ISCAS.2010.5537164

Liu, S. C., van Schaik, A., Minch, B. A., and Delbruck, T. (2014). "Asynchronous binaural spatial audition sensor with 2×64×4 channel output," in *Proceedings of the IEEE transactions on biomedical circuits and systems*. IEEE Transactions on Biomedical Circuits and Systems. doi: 10.1109/TBCAS.2013.2281834

Lyon, R. F. (2017). *Human and machine hearing -extracting meaning from sound*. Cambridge: Cambridge University Press.

Maqueda, A. I., Loquercio, A., Gallego, G., Garcia, N., and Scaramuzza, D. (2018). "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, (Dl)*, Salt Lake City, UT, 5419–5427. doi: 10.1109/CVPR.2018.00568

Neil, D., and Liu, S. C. (2016). "Effective sensor fusion with event-based sensors and deep network architectures," in *Proceedings of the IEEE international symposium on circuits and systems, 2016-July*, Montreal, QC, 2282–2285. doi: 10.1109/ISCAS.2016.7539039

Orchard, G., Frady, E. P., Rubin, D. B. D., Sanborn, S., Shrestha, S. B., Sommer, F. T., et al. (2021). "Efficient neuromorphic signal processing with Loihi 2," in *Proceedings of the IEEE workshop on signal processing systems, SiPS: Design and implementation, 2021-October*, Coimbra, 254–259. doi: 10.1109/SiPS52927.2021.00053

Ralph, N., Joubert, D., Jolley, A., Afshar, S., Tothill, N., van Schaik, A., et al. (2022). Real-time event-based unsupervised feature consolidation and tracking for space situational awareness. *Front. Neurosci.* 16:821157. doi: 10.3389/fnins.2022.821157

Rasetto, M., Dominguez-Morales, J. P., Jimenez-Fernandez, A., and Benosman, R. (2021). *Event Based Time-Vectors for auditory features extraction: A neuromorphic approach for low power audio recognition*. Available online at: http://arxiv.org/abs/2112.07011 (accessed December 13, 2021).

Shrestha, S. B., and Orchard, G. (2018). Slayer: Spike layer error reassignment in time. *Adv. Neural Inform. Process. Syst.* 2018, 1412–1421.

Singh, R. K., Xu, Y., Wang, R., Hamilton, T. J., van Schaik, A., and Denham, S. L. (2018). "CAR-lite: A multi-rate cochlear model on FPGA for spike-based sound encoding," in *Proceedings of the IEEE transactions on circuits and systems I: Regular papers*. doi: 10.1109/ISCAS.2018.8351394

Tapson, J., Cohen, G., and van Schaik, A. (2015). ELM solutions for event-based systems. *Neurocomputing* 149(Pt A), 435–442. doi: 10.1016/j.neucom.2014.01.074

Uysal, I., Sathyendra, H., and Harris, J. G. (2006). A biologically plausible system approach for noise robust vowel recognition. *Midwest Symp. Circuits Syst.* 1, 245–249. doi: 10.1109/MWSCAS.2006.382043

Xu, Y. (2019). *A digital neuromorphic auditory pathway*. Penrith, NSW: Western Sydney University.

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., and van Schaik, A. (2018b). A FPGA implementation of the CAR-FAC cochlear model. *Front. Neurosci.* 12:198. doi: 10.3389/fnins.2018.00198

Xu, Y., Afshar, S., Singh, R. K., Hamilton, T. J., Wang, R., and van Schaik, A. (2018a). "A machine hearing system for binaural sound localization based on instantaneous correlation," in *Proceedings of the 2018 IEEE international symposium on circuits and systems (ISCAS)*, Florence. doi: 10.1109/ISCAS.2018.8351367

Xu, Y., Afshar, S., Singh, R. K., Wang, R., van Schaik, A., and Hamilton, T. J. (2019). "A binaural sound localization system using deep convolutional neural networks,"

in *Proceedings of the international symposium on circuits and systems*, Sapporo. doi: 10.1109/ISCAS.2019.8702345

Xu, Y., Afshar, S., Wang, R., Cohen, G., Thakur, C. S., Hamilton, T. J., et al. (2021). A biologically inspired sound localisation system using a silicon cochlea pair. *Appl. Sci.* 11, 1–21. doi: 10.3390/app11041519

Xu, Y., Thakur, C. S., Singh, R. K., Wang, R., and van Schaik, A. (2016). "Electronic cochlea: CAR-FAC model on FPGA," in *Proceedings of the IEEE biomedical circuits and systems conference*, Shanghai.

Yang, M., Chien, C.-H., Delbruck, T., and Liu, S.-C. (2016). A 0.5 V 55 µW 64 × 2 channel binaural silicon cochlea for event-driven stereo-audio sensing. *IEEE J. Solid State Circ.* 51, 2554–2569. doi: 10.1109/JSSC.2016.2604285

Zappa, F., Villa, F., Lussana, R., Delic, D., Mau, M. C. J., Redouté, J. M., et al. (2020). "Microelectronic 3D imaging and neuromorphic recognition for autonomous UAVs," in *Advanced technologies for security applications*, ed. C. Palestini (Dordrecht: Springer), 185–194. doi: 10.1007/978-94-024-2021-0_17