Check for updates

An Efficient and Perceptually Motivated Auditory Neural Encoding and Decoding Algorithm for Spiking Neural Networks

Zihan Pan[1], Yansong Chua[2]*, Jibin Wu[1], Malu Zhang[1], Haizhou Li[1] and Eliathamby Ambikairajah[3]

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore, [2] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, Singapore, [3] School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia

The auditory front-end is an integral part of a spiking neural network (SNN) when performing auditory cognitive tasks. It encodes the temporal dynamic stimulus, such as speech and audio, into an efficient, effective and reconstructable spike pattern to facilitate the subsequent processing. However, most of the auditory front-ends in current studies have not made use of recent findings in psychoacoustics and physiology concerning human listening. In this paper, we propose a neural encoding and decoding scheme that is optimized for audio processing. The neural encoding scheme, that we call Biologically plausible Auditory Encoding (BAE), emulates the functions of the perceptual components of the human auditory system, that include the cochlear filter bank, the inner hair cells, auditory masking effects from psychoacoustic models, and the spike neural encoding by the auditory nerve. We evaluate the perceptual quality of the BAE scheme using PESQ; the performance of the BAE based on sound classification and speech recognition experiments. Finally, we also built and published two spike-version of speech datasets: the Spike-TIDIGITS and the Spike-TIMIT, for researchers to use and benchmarking of future SNN research.

Keywords: spiking neural network, neural encoding, auditory perception, spike database, auditory masking effects

INTRODUCTION

The temporal or rate based Spiking Neural Networks (SNN), supported by stronger biological evidence than the conventional artificial neural networks (ANN), represents a promising research direction. Neurons in a SNN communicate using spiking trains that are temporal signals in nature, therefore, making SNN a natural choice for dealing with dynamic signals such as audio, speech, and music.

In the domain of rate-coding, we studied the computational efficiency of SNN (Pan et al., 2019). Recently, further evidence has supported the theory of temporal coding with spike times. To learn a temporal spike pattern, a number of learning rules have been proposed, which include the single-spike Tempotron (Gütig and Sompolinsky, 2006), conductance-based Tempotron (Gütig and Sompolinsky, 2009), the multi-spike learning rule ReSuMe (Ponulak and Kasiński, 2010;

Taherkhani et al., 2015), the multi-layer spike learning rule SpikeProp (Bohte et al., 2002), and the Multi-spike Tempotron (Gütig, 2016), etc. The more recent studies are aggregate-label learning (Gütig, 2016), and a novel probability-based multi-layer SNN learning rule (SLAYER) (Shrestha and Orchard, 2018).

In our research, a question is constantly asked: what are the advantages of SNN over ANN? From the viewpoint of neural encoding, we expect to encode a dynamic stimulus into spike patterns, which was shown to be possible (Maass, 1997; Ghosh-Dastidar and Adeli, 2009). Deep ANNs have benefited from the datasets created in recent years. In the field of image classification, there is ImageNet (Russakovsky et al., 2015); in the field of image detection, there is COCO dataset (Veit et al., 2016); while in the field of Automated Speech Recognition (ASR), there is TIMIT for phonemically and lexically transcribed speech of American English speakers (Garofolo, 1993). With the advent of these datasets, better and faster deep ANNs inevitably follow (Hochreiter and Schmidhuber, 1997; Simonyan and Zisserman, 2014; Redmon et al., 2016). The publicly available datasets become the common platform for technology benchmarking. In the study of neuromorphic computing, there are some datasets such as N-MNIST (Orchard et al., 2015), DVS Gestures (Amir et al., 2017), and N-TIDIGITS (Anumula et al., 2018). They are designed for SNN benchmarking. However, these datasets are relatively small compared with the deep learning datasets.

One may argue that the benchmarking datasets for deep learning may not be suitable for SNN studies. Let us consider image classification as an example. Humans process static images in a similar way as they would process live visual inputs. We note that live visual inputs contain much richer information than 2-D images. When we map (Rueckauer et al., 2017) or quantize (Zhou et al., 2016) static images into spike trains, and compare the performance of an ANN on static images, and a SNN on spike trains, we observe an accuracy drop. One should, however, not hastily conclude that SNNs are inherently poor in image classification as a consequence of event-based activations in SNNs. Rather, the question seems to be: how can one better encode images into spikes that are useful for SNNs, and how can one better use these spikes in an image classification task? For some of the recent image-based neuromorphic datasets, Laxmi et al. (Iyer et al., 2018) has argued that no additional information is encoded in the time domain that is useful for pattern classification. This prompts us to look into the development of event-based datasets that inherently contain spatio-temporal information. On the other hand, a dataset has to be complex enough such that it simulates a real-world problem. There are some datasets that support the learning of temporal patterns (Zhang et al., 2017, 2018, 2019; Wu et al., 2018a), whereby each pattern contains only a single label, such as a sound event or an isolated word. Such datasets are much simpler than those in deep learning studies (Graves et al., 2006), whereby a temporal pattern involves a sequence of labels, such as continuous speech. For SNN study to progress from isolated word recognition toward continuous speech recognition, a continuous speech database is required. In this paper, we would describe how we convert the TIMIT dataset to its event-based equivalent: Spike-TIMIT.

A typical pattern classification task consists of three stages: encoding, feature representation, and classification. The boundaries between each stage are getting less clear in an end-to-end classification neural network. Even then, a good encoding scheme can significantly ease the workload of the subsequent stages in a classification task, for instance, the Mel-Frequency Cepstral Coefficients (MFCC) (Mermelstein, 1976) is still very much in use for automatic speech recognition (ASR). Hence the design of a spiking dataset should consider how the encoding scheme could help reduce the workload of the SNN in a classification task. This cannot be misconstrued as giving the SNN an unfair advantage so long as all SNNs are measured using the same benchmark. The human cochlea performs frequency filtering (Tobias, 2012) while human vision performs orientation discrimination (Appelle, 1972). These all involve encoding schemes to help us better understand our environment. In our earlier work (Pan et al., 2019), on a simple dataset TIDIGITS (Leonard and Doddington, 1993) that contains only single spoken digits, we used a population threshold coding scheme to encode the dataset into events, which we refer to as Spike-TIDIGITS. Using such an encoding scheme, we go on to show that the dataset becomes linearly separable, i.e., the input can be classified based on spike counts alone. This demonstrates that when information is encoded in both the temporal (spike timing) and spatial (which neuron to spike) domain, the encoding scheme is able to project the inputs to a higher dimension, that takes some of the workload off the subsequent feature extraction and classification stages. In the case of Spike-TIDIGITS, the spikes encoded can be directly counted and then classified using a Support Vector Machine (SVM). Using this neural encoding scheme, We further enhance it and then apply it to the TIMIT dataset in this work.

The motivation of this paper is two-fold. Firstly, we believe that we need well-designed spike-encoded datasets that represent the state-of-the-art encoding methodology. With these datasets, one can focus the research on SNN feature representation and classification tasks. Secondly, the datasets should present a challenge in pattern classification, that become the reference benchmark in future SNN studies.

As speech is the most common way of human communication, we are looking into the neural encoding of speech signals in this work. The first question is how best possible to convert speech signals into spikes. There have been many related studies in speech and audio encoding, each of which is optimized for a specific objective, for example, the minimum signal reconstruction error (Loiselle et al., 2005; Dennis et al., 2013; Xiao et al., 2016). However, the speech and audio encoding methods have not taken into consideration the combination of psychoacoustic effects, computational efficiency, and pattern classification performance for neuromorphic implementation. In the SNN applications for speech recognition (Xiao et al., 2016; Darabkh et al., 2018), MFCC (Mermelstein, 1976) are commonly used as the spectral representation in speech recognition. Others have tried to use the biologically plausible cochlear filter bank, but they are either analog filters which are prone to changes in the external environment (Liu and Delbruck, 2010), or yet to be studied in a spike-driven SNN system (Loiselle et al., 2005).

Yang et al. (2016) successfully implements a silicon cochlear for event-driven audio sensing, which has not considered the psychoacoustics of the auditory system.

Considering spectral representation, an important step in neural encoding is to then convert the spectral energy in a perceptual frequency band into a spike train. The most common way is to treat the two-dimensional time-frequency spectrogram as a static image, then converting each "pixel" value into a spike latency time within the framing window size (Wu et al., 2018a), or into the phase of the sub-threshold membrane potential oscillation (Nadasdy, 2009). Such "pixel-conversion" methods do not represent the spatio-temporal dynamics of the auditory signals in the same way as the spike trains in a SNN, therefore, another feature representation step is required, such as the self-organizing map (Wu et al., 2018b), or local spectrogram features (Dennis et al., 2013; Xiao et al., 2016). If the audio encoding is able to capture the spatio-temporal dynamics that are discriminative for classification (Gütig and Sompolinsky, 2009), it is not necessary to encode every speech frame in the front-end, therefore, the spiking rate can be reduced. Finally, it has not been given enough attention as to how to reconstruct a neural encoded speech signal back into its auditory signals for perceptual evaluation. Speech signal reconstruction is a critical task in speech information processing, such as speech synthesis, singing synthesis, and dialogue technology.

To address the need for neuromorphic computing for speech information processing, we propose three criteria for a biologically plausible auditory encoding (BAE) front-end:

(1) Biologically plausible spectral features.
(2) Sparse and energy-efficient spike neural coding scheme.
(3) Friendly for temporal learning algorithms on cognitive tasks.

The fundamental research problem in neural encoding is how to encode the dynamic and continuous speech signals into discrete spike patterns. Spike rate code is thought to be less likely in an auditory system since much evidence suggests otherwise, for example, how bats rely highly on the precise spike timing of their auditory system to locate sound sources by detecting a time difference as short as 5 μs. Latency code and phase code are well supported by neuro-biological observations. However, on its own, they cannot provide an invariant representation of the patterns for a classification task.

To facilitate the processing of an SNN in a cognitive task, neural temporal encoding should not only consider how to encode the stimulus into spikes, but also care about how to represent the invariant features. Just like the auditory and visual sensory representations in the human prefrontal cortex, such representations in the proposed BAE front-end are required in an SNN framework, that can then be implemented with a low-cost neuromorphic solution, that can effectively reduce the processing workload in the subsequent SNN pipeline. A large number of observations in neuroscience support the observation that our auditory sensory neurons encode the input stimulus using threshold crossing events in a population of sensory neurons (Ehret, 1997; Hopfield, 2004). Inspired by these observations, a simple version of threshold coding has been proposed (Gütig and Sompolinsky, 2009), in which a population of encoding neurons with a set of uniformly distributed thresholds encode the spectral energy of different frequency channels into spikes. Such a cross-and-fire mechanism is reminiscent of quantization from the point of view of information coding. In our proposed BAE encoding front-end, such a neural coding scheme is also being incorporated. Further investigation is presented in the "Experiment and Results" section.

Besides effective neural coding representation, an efficient auditory front-end aims to encode acoustic signals into sparse spike patterns, while maintaining sufficient perceptual information. To achieve such a goal, our biological auditory system has provided us a solution best understood as masking effects (Harris and Dallos, 1979; Shinn-Cunningham, 2008). The auditory masking is a complex and yet to be fully understood psychoacoustic phenomenon as some components of the acoustic events are not perceptible in both frequency and time domain (Ambikairajah et al., 1997). From the viewpoint of perceptual coding, these components are regarded as redundancies since they are inaudible. Implementing the masking effects, those inaudible components will be coded with larger quantization noise or not coded at all. Although the mechanism and function of masking are not yet fully understood, its effects have already been successfully exploited in auditory signal compression and coding (Ambikairajah et al., 2001), for efficient information storage, communication, and retrieval. In this paper, we propose a novel idea to apply the auditory masking effects in both frequency and time domain, which we refer to as simultaneous masking and temporal masking, respectively, in our auditory neural encoding front-end so as to reduce the number of encoding spikes. This improves the sparsity and efficiency of our encoding scheme. Given how we address the three optimization criteria of neural encoding, we refer to it as BAE scheme or BAE. Such an auditory encoding front-end also provides an engineering platform to bridge the study of masking effects between psychoacoustics and speech processing.

Our main contributions in this paper are: (1) we emphasize the importance of spike acoustic datasets for SNN research. (2) we propose an integrated auditory neural encoding front-end to further research in SNN-based learning algorithms. With the proposed BAE encoding front-end, the speech or audio datasets can be converted into energy-efficient, information-compact, and well-representative spike patterns for subsequent SNN tasks.

The rest of this paper is organized as follows: in section "Materials and Methods" we discuss the auditory masking effects, and how simultaneous masking in the frequency domain and temporal masking in the time domain for neural encoding of acoustic stimulus is being implemented; the BAE encoding scheme is applied in conjunction with masking to RWCP, TIDIGITS, and TIMIT datasets. In section "Experiment and Results," we describe the details of the resulting spike datasets and evaluate them against their original datasets. We discuss our findings in section "Discussion" and conclude in section "Conclusion."

MATERIALS AND METHODS

Auditory Masking Effects

Most of our speech processing front-ends employ a fixed feature extraction mechanism, such as MFCC, to encode the input signals, whereas the human auditory sensory system ignores some while strongly emphasizes others, commonly referred to as the attention mechanism in psychoacoustics. The auditory masking effects closely emulate this phenomenon (Shinn-Cunningham, 2008).
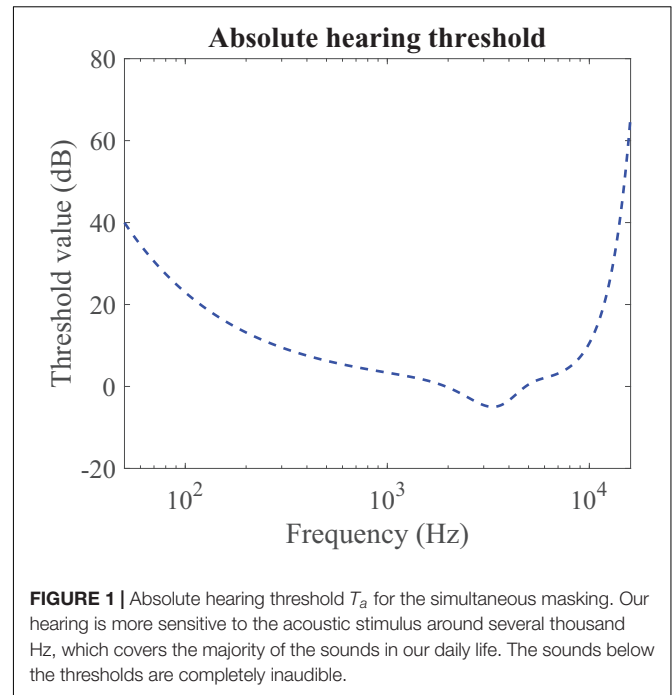
Auditory masking is a known perceptual property of the human auditory system that occurs whenever the presence of a strong audio signal makes its neighborhood of weaker signals inaudible, both in the frequency and time domain. One of the most notable applications of auditory masking is the MPEG/audio international standard for audio signal compression (Ambikairajah et al., 2001; Fogg et al., 2007). It compresses the audio data by removing the acoustically inaudible elements, or by encoding those parts with less number of bits, due to more tolerance to quantization noise (Ambikairajah et al., 1997). To achieve such a goal, the algorithm is supported by two different kinds of auditory maskings according to the psychoacoustic model (Lagerstrom, 2001):

1. In the frequency domain, two kinds of masking effects are used. Firstly, by allocating the quantization noise in the least sensitive regions of the spectrum, the perceptual distortion caused by quantization is minimized. Secondly, an absolute hearing threshold is exploited, below which the spectral components are entirely removed.
2. In the time domain, the masking effect is applied such that the local peaks of the temporal signals in each frequency band will make their ensuing audio signals inaudible.

Motivated by the above signal compression theory, we propose an auditory masking approach to spike neural encoding, which greatly increases the coding efficiency of the spike patterns, by eliminating those perceptually insignificant spike events. The approach is conceptually consistent with the MPEG-1 layer III signal compression standard (Fogg et al., 2007), with modifications according to the characteristics of spiking neurons.

Simultaneous Masking

The masking effect presented in the frequency domain is referred to as simultaneous masking. According to the MPEG-1 standards, there are two sorts of masking strategies in the frequency domain: the absolute hearing threshold and the frequency masking. The simultaneous masking effects are common in our daily life. For instance, the sensible sound levels of our auditory systems vary in different frequencies, therefore, we can be more sensitive to the sounds in our living environment. This is an evolutionary advantage for survival, in both human beings and animals. Besides the absolute hearing threshold, every acoustic event in the spectrum will also influence the perception of the neighboring frequency components, that is, different levels of tones could contribute to masking effects of other frequency tones. For instance, in a symphony show, the sounds from

FIGURE 1 | Absolute hearing threshold T_a for the simultaneous masking. Our hearing is more sensitive to the acoustic stimulus around several thousand Hz, which covers the majority of the sounds in our daily life. The sounds below the thresholds are completely inaudible.

different musical instruments can be fully or partially masked by each other. As a result, we can enjoy the compositions of various frequency components with rich diversities. Such a psychoacoustic phenomenon is called frequency masking.

Figure 1 illustrates the absolute hearing threshold, T_a, as a function of frequency in Hz. The function is derived from psychoacoustic experiments, in which pure tones continuous in the frequency domain are presented to the test subjects and the minimal audible sound pressure levels (SPL) in dB are recorded. The commonly used function to approximate the threshold is (Ambikairajah et al., 1997):

$$T_a(f) = 3.64 \times \left(\frac{f}{1000}\right)^{-0.8} - 6.5 \times e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 0.001 \times \left(\frac{f}{1000}\right)^4 \tag{1}$$

For the frequency masking, in the MPEG-1 standard, some sample pulses under masking thresholds might be partially masked, thus they are encoded by a lower number of bits. However, in the event-based scenario, spike patterns carry no amplitude information, similar to on-off binary values, which means that partial masking can hardly be realized. As such, we have modified the approach such that all components under the frequency masking are fully masked (discarded). Further reconstruction and pattern recognition experiments are necessary to evaluate such an approach. **Figure 2** shows the overall masking thresholds with both masking strategies in the frequency domain. This figure illustrates the simultaneous masking thresholds added to the acoustic events in a spectrogram. The sound signals with different spectral power in different cochlear filter channels will suffer from various masking thresholds.
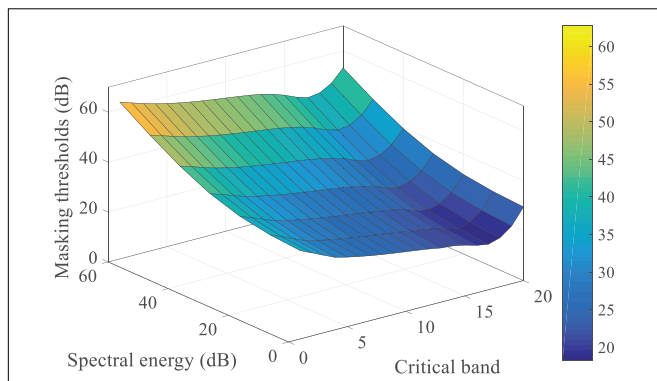
FIGURE 2 | The frequency masking thresholds acting on a maskee (the acoustic events being masked), generated by the acoustic events from the neighboring critical bands, are shown as a surface in a 3-D plot. The acoustic events are referred to as the spectral power of the frames in a spectrogram. The spectral energy axis is the sound level of a maskee; the critical band axis is the frequency bins of the cochlear filter bank, as introduced in section "Spike-TIDIGITS and Spike-TIMIT Databases"; the masking thresholds axis indicates the overall masking levels on the maskees of different sound levels from various critical bands. For example, an acoustic event of 20dB level on the 10th critical band is masked off by the masking threshold of nearly 23dB, which is introduced by the other auditory components of its neighboring frequency channels.
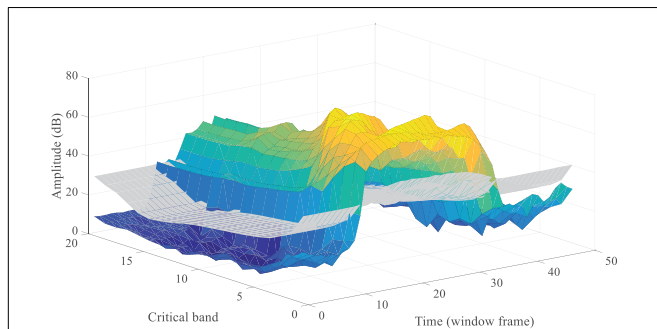

FIGURE 3 | The overall simultaneous masking effects on a speech utterance of "one," in a 3-D spectrogram. Combining the two kinds of masking effects in the frequency domain (refer to Figures 1, 2), the gray surface shows the overall masking thresholds on a speech utterance (the colorful surface). All the spectral energy under the thresholds will be imperceptible.

Figure 3 provides a real-world example of the simultaneous masking. The spectrogram of a speech utterance of "one" from the TIDIGITS dataset is demonstrated in a 3-D plot. The gray surface illustrates the simultaneous masking threshold acting on the spectrogram (colorful surface). By the masking strategy, the acoustic events with spectral energy lower than the threshold surface will be removed. Section "Biologically Plausible Auditory Encoding With Masking Effects" will introduce how to convert the masked spectrogram into a sparse and well-represented spike pattern.

Temporal Masking

Another auditory masking effect is temporal masking in the time domain. Conceptually similar to the frequency masking, a louder sound will mask the perception of the other acoustic components
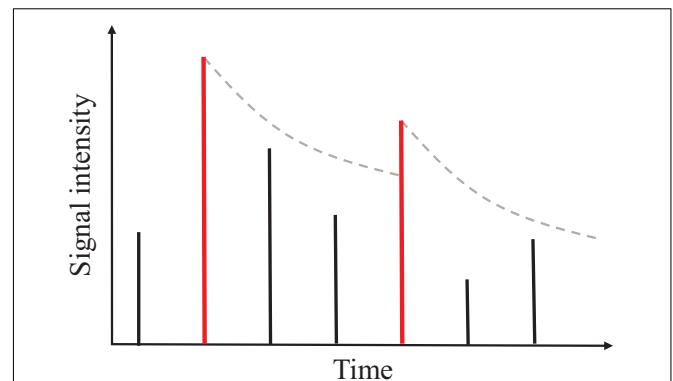

FIGURE 4 | The illustration of temporal masking: each bar represents the acoustic event received by the auditory system. In this paper, acoustic events generally referred to framing spectral power, which are the elements to be parsed to an auditory neural encoding scheme. A local peak event (red bar) forms a masking threshold represented by an exponentially decaying curve. The subsequent events that are weaker than the leading local peak will not be audible until another local peak event exceeds the masking threshold.

in the time domain. As illustrated in **Figure 4**, the vertical bars represent the signal intensity of short-time frames, that is called acoustic events, along the time axis. A local peak (the first red bar) forms a masking threshold that makes the following events inaudible until the next local peak (the second red bar) exceeds the masking threshold. According to the psychoacoustic studies, the temporal masking threshold is modeled as an exponentially decaying curve (Ambikairajah et al., 2001):

$$y(n) = c^n \times p_1 \qquad (2)$$

where $y(n)$ denotes the masking threshold level on the nth following an acoustic event; c is the exponential index and p_1 represents the sound level of the local peak as the beginning of the masking. The decaying parameter c is tuned according to the hearing quality.

Auditory Masking Effects in Both Domains

By applying both the simultaneous masking and temporal masking illustrated above, we can remove those imperceptible acoustic events (frames) from the overall spectrogram. Since our goal is to apply the masking effects in the precise timing neural code, we propose the strategy as follows:

1. The spike pattern $\mathbf{P}_{K \times N}(p_{ij})$ is generated from the raw spectrogram $\mathbf{S}_{K \times N}(s_{ij})$ without masking effects, by some temporal neural coding methods, which will be discussed in section "Neural Spike Encoding" Here the index i, j refers to the time-frequency bin in the spectrogram, with i referring to the frequency bin, and j referring to the time frame index. The spike pattern $\mathbf{P}_{K \times N}$ is defined as a matrix that:

$$p_{ij} = \begin{cases} t_{f,} & \text{if a spike is emitted within the duration of the time-frequency bin } i, \ j. \\ 0, & \text{otherwise} \end{cases}$$

$$(3)$$

where t_f is the encoded precise spike timing. As such the spike pattern $\mathbf{P}_{K \times N}\left(p_{ij}\right)$ is a sparse matrix that records the spike timing.

2. According to the spectrogram $\mathbf{S}_{K \times N}\left(s_{ij}\right)$ and the auditory perceptual model, the simultaneous masking threshold matrix $\mathbf{M}_{\text{simultaneous}}\left(m_{ij}^{\text{simultaneous}}\right)$ and the temporal masking threshold matrix $\mathbf{M}_{\text{temporal}}\left(m_{ij}^{\text{temporal}}\right)$ are obtained. The overall masking threshold $\mathbf{M}_{K \times N}\left(m_{ij}\right)$ is defined as follows. It provides a 2-D masking threshold surface that has the same dimensions as the spectrogram.

$$m_{ij} = \max \left\{ m_{ij}^{\text{simultaneous}}, m_{ij}^{\text{temporal}} \right\} \qquad (4)$$

3. A masking map $\Phi_{K \times N}\left(\phi_{ij}\right)$ is generated, whose dimensions are the same as the spectrogram. The element of the matrix $\Phi_{K \times N}\left(\phi_{ij}\right)$ is defined as:

$$\phi_{ij} = \begin{cases} 1, & \text{if } s_{ij} \geq m_{ij} \\ 0, & \text{if } s_{ij} < m_{ij} \end{cases} \qquad (5)$$

where the time-frequency bin i,j is masked with $\phi_{i,j} = 0$ when the frame energy $s_{i,j}$ is less than the masking threshold m_{ij}, otherwise, $\phi_{i,j} = 1$.

4. Apply the masking map matrix $\Phi_{K \times N}\left(\phi_{ij}\right)$ to the encoded pattern $\mathbf{P}_{K \times N}\left(p_{ij}\right)$ to generate a masked spike pattern $\mathbf{P}^{\text{mask}}(p_{ij}^{\text{mask}})$:

$$\mathbf{P}^{\text{mask}} = \mathbf{P}_{K \times N} \circ \Phi_{K \times N} \qquad (6)$$

where $\circ$ denotes the Hadamard product. By doing so, those perceptually insignificant spikes are eliminated, thus forming a more compact and sparse spike pattern.

Figure 5 demonstrates the auditory masking effects acting in both the frequency and time domains, on a speech utterance of "one" in the TIDIGITS dataset. The colored surface represents the original spectrogram while the gray areas represent the spectral energy values that are being masked. For TIDIGITS datasets, nearly half of the acoustic events (frames) are removed according to our auditory masking strategy, which corresponds to the 55% removal of PCM pulses in speech coding (Ambikairajah et al., 2001).

Cochlear Filters and Spike Coding

The human auditory system is primarily a frequency analyzer (Tobias, 2012). Many studies have confirmed the existence of the perceptual centre frequencies and equivalent bandwidths. To emulate the working of the human cochlea, several artificial cochlear filter banks have been well studied: GammaTone filter bank (Patterson et al., 1987; Hohmann, 2002), Constant Q Transform-based filter bank (CQT) (Brown, 1991; Brown and Puckette, 1992), Bark-scale filter bank (Smith and Abel, 1999), etc. They share the same idea of logarithm distributed centre frequencies and constant Q factors but slightly differ in the exact parameters. To build the auditory encoding system, we adopt an event-based CQT-based filter bank in the time domain, following our previous work (Pan et al., 2018).

Time-Domain Cochlear Filter Bank

Adopting an event-based approach to emulate the human auditory system, we propose a neuronal implementation of the event-driven cochlear filter bank, of which the computation can be parallelized as follows,

- As illustrated in **Figure 6**, a speech waveform (**Figure 6A**) is filtered by K neurons (**Figure 6B**) where each neuron represents one cochlear filter from a particular frequency bin.
- The weights of each neuron in **Figure 6B** are set as the time-domain impulse response of the corresponding cochlear filter. The computing of a neuron with its input is inherently a time-domain convolution process.
- The output of the filter bank neurons is a K-length vector (**Figure 6C**), where K is the number of filters, for each time step. Since the signal (**Figure 6A**) shifts sample by sample, the width of the output matrix is the same as the length of the input signal. As such, the auditory signal is decomposed into multiple channels in parallel, forming a spectrogram.

Suppose a speech signal x with M samples $x = [x_1, x_2, ..., x_M]$ sampled at 16 kHz. For the kth cochlear filter, the impulse response (wavelet) is a M_k-length vector $F_k = [F_k(1), F_k(2), ..., F_k(M_k)]$. We note the impulse response F_k has an infinite window size, however, numerically its amplitude decreases to small values outside an effective window, thus having little influence on the convolution results. As investigated in Pan et al. (2018), we empirically set M_k to an optimal value. So the mth output of the kth cochlear filter neuron is modeled as $y_k(m)$:

$$y_k(m) = \sum_{i=1}^{M_k} \phi_m(i) F_k(i), \; k = 1, 2, ..., K, \; m = 1, 2, ..., M \qquad (7)$$

$$\phi_m = [x_m, x_{m+1}, x_{m+2}, ..., x_{m+M_k-1}], \; m \in 1, ..., M \qquad (8)$$

ϕ_m is a subset of the input samples within the mth window, whose length is the same as that of the M_k-length wavelet, indicated as the samples between the two arrows in **Figures 6A,B**. The window ϕ_m will move sample by sample, naturally along with the flow of the input signal samples. At each time step, a vector of length K, which is the number of filters, is generated as shown in **Figure 6C**. After M such samples, the final output time-frequency map of the filter bank is a $K \times M$ matrix $\mathbf{Y}_{K \times M}$.

After time-domain cochlear filtering, the $K \times M$ time-frequency map $\mathbf{Y}_{K \times M}$ should be framed, which emulates the signal processing of hair cells in the auditory pathway. For the output waveform from each channel, we apply a framing window of length l (samples) with a step size of $l/2$ and calculate the logarithmic frame energy e of one framing window:

$$e = 10 \log \left(\sum_{q=1}^{l} x_q^2 \right) \qquad (9)$$

where x_q denotes the samples within the l-length window; e is the spectral energy of one frame, hence obtaining the
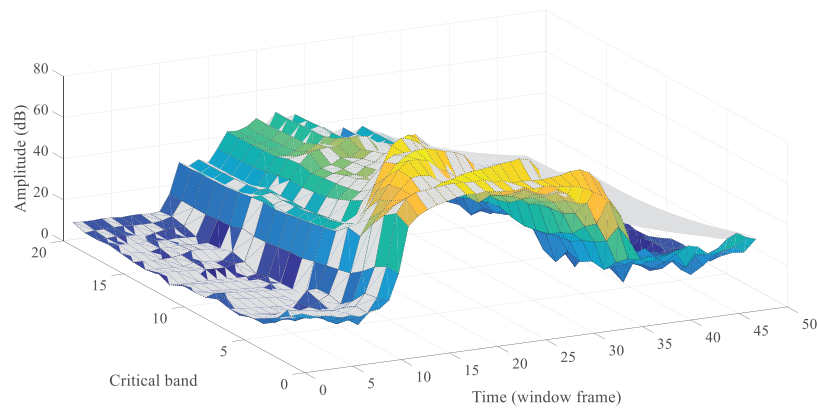
FIGURE 5 | Both the simultaneous and temporal masking effects acting on the 3-D plot spectrogram of a speech utterance of "one." The gray-color shaded parts of the spectrogram are masked.
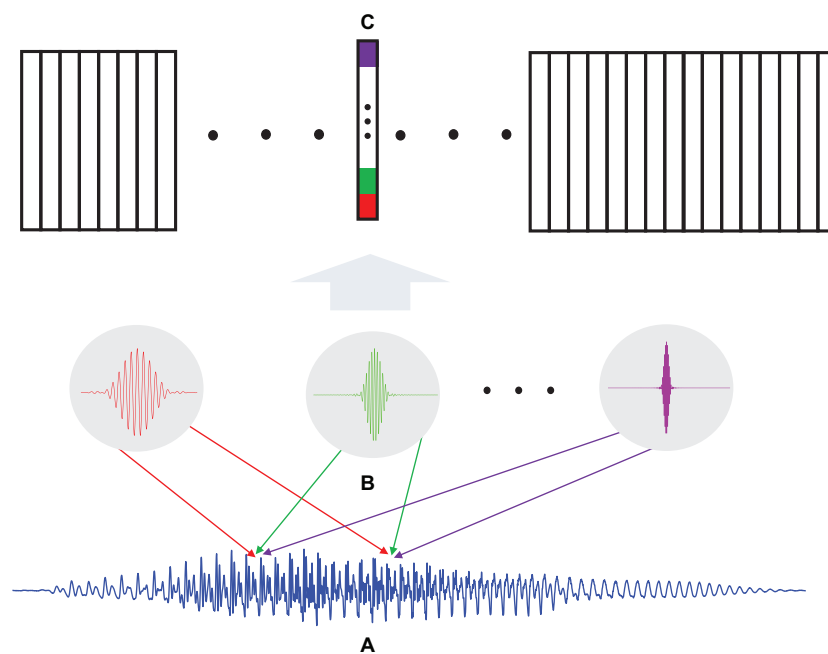

FIGURE 6 | (A) A speech signal of M samples; **(B)** Time-domain filter bank with K neurons that act as filters; **(C)** The output spectrogram that has $K \times M$ dimension.

time-frequency spectrum $\mathbf{S}_{K \times N}$ (s_{ij}) as indicated in section "Auditory Masking Effects in Both Domains" which will be further encoded into spikes.

Neural Spike Encoding

In the inner ear, the motion of the stereocilia in the inner hair cells is converted into a chemical signal that excites adjacent nerve fibers, generating neural impulses that are then transmitted along the auditory pathway. Similarly, we would like to convert the sub-band framing energy into electrical impulses, or so-called spikes, for the purpose of information encoding and transmission. In the prior work, the temporal dynamic sequences are encoded using several different methods: latency coding (Wu et al., 2018a), phase coding (Arnal and Giraud, 2012; Giraud and Poeppel,

2012), latency population coding (Dean et al., 2005), that are adopted for specific applications. These encoding schemes are not optimized for neuromorphic implementation.

We would like to propose a biologically plausible neural encoding scheme by taking into account the three criteria as defined in section "Introduction." In this section, the particular neural temporal coding scheme, which converts perceptual spectral power to precise spike times, is designed to meet the need of synaptic learning rules in SNNs (Gütig and Sompolinsky, 2006; Ponulak and Kasiński, 2010). As such, the resulting temporal spike patterns are supposed to be friendly toward temporal learning rules.

In our previous work (Pan et al., 2019), two mainstream neural encoding schemes, the single neuron temporal codes (latency

coding, phase coding) and the population codes (population latency/phase coding, threshold coding) are compared. It is found that the threshold coding outperforms the other coding schemes in SNN-based pattern recognition tasks. Next are some observations made while comparing threshold coding, and the single neuron temporal coding.

First of all, the single temporal coding scheme, such as latency or phase coding, encodes the spectral power using spike delaying time, or phase-locking time. Suppose a frame of normalized spectral power is e, the nth latency spike timing t_n^f is defined as:

$$t_n^f = (1 - e) * T + (n - 1) * T = (n - e) * T \qquad (10)$$

where T denotes the time duration of the encoding window. For the phase coding, t_n^f is phase-locked to the nearest peak of the sub-threshold membrane oscillation. The spectral power, that represents the amplitude information, e is represented as the relative spike timing $(1 - e) * T$ within each window and the number of spikes embedded are in the order n. Unfortunately, the SNN can hardly decode such an encoding scheme without the knowledge of the encoding window boundaries, implicitly provided by the spike order n and window length T. The spatio-temporal spike patterns could not provide such knowledge explicitly to the SNN. On the other hand, in the population code, such as threshold coding, the multiple encoding neurons naturally represent the amplitudes of the spectral power frames, and we only need to represent the temporal information in the spike timing. For example, the spike timing of the nth onset encoding neuron of the threshold code t_f^n is:

$$t_f^n = t_{\text{crossing}} \qquad (11)$$

t_{crossing} records the time when the spectral tuning curve from one sub-band crosses the onset threshold θ_n of the nth encoding neuron. In this way, both the temporal and amplitude information is encoded and made known to the SNN, which meets the third criterion mentioned above.

Secondly, coding efficiency, which refers to the average encoding spike rates (number of spikes per second), is also studied in Pan et al. (2019). The threshold code has the least average spike rates among all investigated neural codes. As the threshold code encodes only threshold-crossing events, it is supposed to be the most efficient coding method.

Thirdly, the threshold code promises to be more robust against noise, such as spike jitter. As it encodes the trajectory of the dynamics of the sub-band spectral power, the perturbation of precise spike timing will have less impact on the sequence of encoding neurons.

As such, the threshold code is a promising encoding scheme for temporal sequence recognition tasks (Pan et al., 2019). Further evaluation will be provided later in the experiments. While we note that each neural coding scheme has its own advantages, we focus on how the encoding scheme may help subsequent SNN learning algorithms in a cognitive task in this paper. As such, we adopt the threshold code for all experiments in this paper.

Biologically Plausible Auditory Encoding (BAE) With Masking Effects

We propose a BAE front-end with masking effects as illustrated in **Figure 7**.

Firstly the auditory stimuli are sensed and amplified by the microphone and some peripheral circuits, leading to a digital signal (a). This process corresponds to the pathway of the pinna, external auditory meatus, tympanic membrane, and auditory tube. Then the physically sensed stimuli are filtered by the cochlear filter bank (b), that emulates the cochlear function of frequency analysis. The outputs of the cochlear filter bank are parallel streams of time-domain sub-band (or so-called critical band) signals with psychoacoustic centre frequencies and bandwidths. For the purpose of further neural coding and cognitive tasks, the sub-band signals should be framed as the logarithm-scale energy as per Eq. 9. The output of (c), the raw spectrogram, is then converted into a precise spike pattern. The spectrogram is also being used to calculate the simultaneous and temporal masking thresholds, as in (d) and (e), under which the spikes will be omitted. Finally a sparse, perceptually related, and learnable temporal spike pattern for a learning SNN is generated as shown in (g).

Figure 8 gives an example of the intermediate results at different stages in **Figure 7** for a speech data waveform. **Figures 8A,B** show the raw waveform and the spectrogram of a speech utterance "three" spoken by a male speaker. The spectrogram is further encoded into a raw spike pattern by threshold neural coding. **Figure 8D** is the masking thresholds as formulated in section "Auditory Masking Effects," according to which the raw spike pattern **Figure 8C** is masked and results in a masked spike pattern **Figure 8E**. 50.48% of all spikes are discarded, given by the results in the later experiment section. **Figure 9** further demonstrates the comparison between auditory masked/unmasked spike patterns.

EXPERIMENT AND RESULTS

Spike-TIDIGITS and Spike-TIMIT Databases

The TIDIGITS (Leonard and Doddington, 1993) (LDC Catalog No. LDC93S10) is a speech corpus of spoken digits for speaker-independent speech recognition (Cooke et al., 2001; Tamazin et al., 2019). The speakers are from different genders (male and female), age ranges (adults and children), dialect districts (Boston, Richmond, Lubbock, etc.). As such, the corpus provides sufficiently speaker diversity and becomes one of the common benchmarking datasets. The TIDIGITS has a vocabulary of 11 spoken words of digits. The original database contains both isolated digits and digits sequences. In this work, we only use the isolated digits: each utterance contains one individual spoken digit. In this first attempt, we would like to build a spike-version
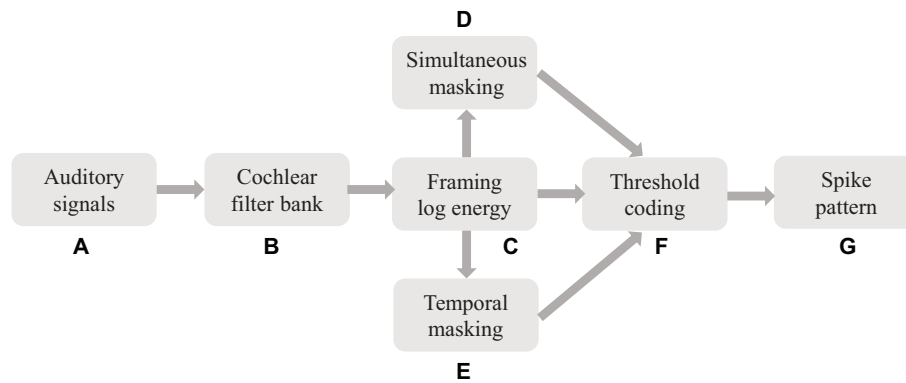
FIGURE 7 | The BAE scheme for temporal learning algorithms in auditory cognitive tasks. The raw auditory signals (a) are filtered by the CQT-based event-driven cochlear filter bank, resulting in a parallel stream of sub-band signals. For each sub-band, the signal is logarithmically framed, which corresponds to the processing in auditory hair cells. The framed spectral signals are then further masked in simultaneous and temporal masking.
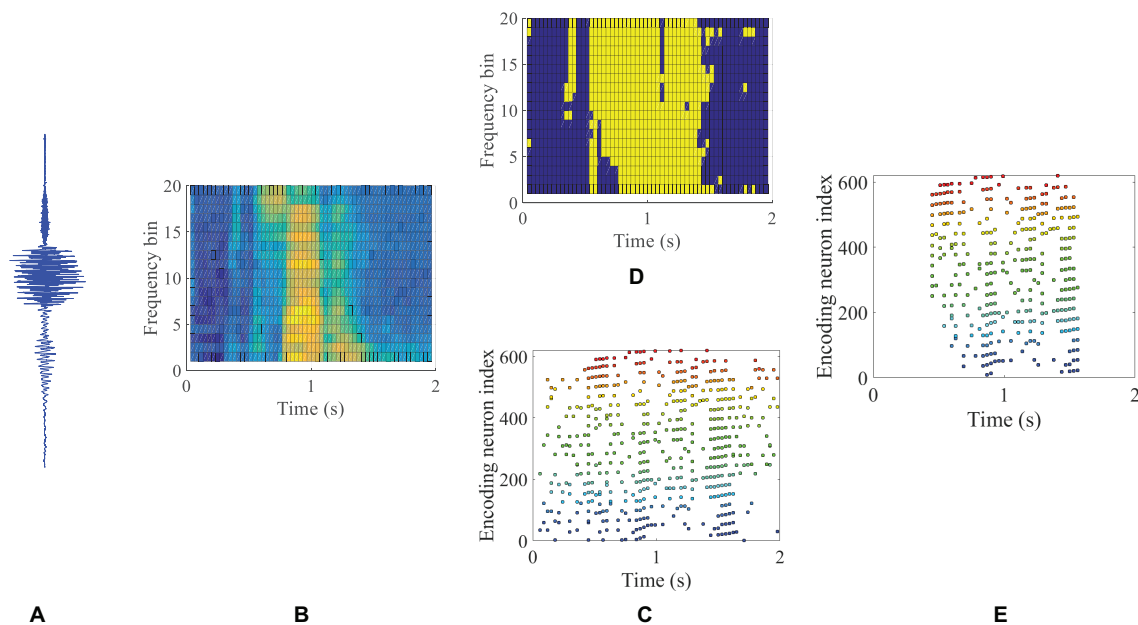

FIGURE 8 | An illustration of the intermediate results in a BAE process. Raw speech signal **(A)** of a speech utterance "three" is filtered and framed into a spectrogram **(B)**, corresponding to the process in **Figure 7** (a) and (c). By applying the neural threshold code, a precise spike pattern **(C)** is generated from the spectrogram. The masking map as described in Eq. 5 is illustrated in **(D)**, where yellow and dark blue color blocks represent the values 1 and 0, respectively. The masking **(D)** is applied to the spike pattern **(C)** and the auditory masked spike pattern is obtained in **(E)**.

speech dataset that contains sufficient diversity and can be immediately used to train an SNN classifier (Pan et al., 2018; Wu et al., 2018a). As each digit is repeated 224 and 226 times, the Spike-TIDIGITS has 224 × 11 = 2464 and 226 × 11 = 2486 isolated digit utterances for the training and testing set, respectively.

The BAE encoder proposed in section "Biologically Plausible Auditory Encoding With Masking Effects" and **Figure 7** is applied as the standard encoding scheme to generate this spike dataset. **Tables 1**, **2** describe the parameters in the encoding process of Spike-TIDIGITS.

Next, we encode one of the most popular speech dataset TIMIT (Garofolo, 1993) into a spike-version, Spike-TIMIT. TIMIT dataset consists of richer acoustic-phonetic content than TIDIGITS (Messaoud and Hamida, 2011). It consists of continuous speech utterances, that are useful for the evaluation of speech coding schemes (Besacier et al., 2000), speech enhancement El-Solh et al. (2007) or ASR systems (Mohamed et al., 2011; Graves et al., 2013). Similar to TIDIGITS, the speakers of TIMIT corpus are from eight different dialect regions in the United States, 438 males and 192 females. There are 4621 and 1679 speech sequences in the training and testing sets. This
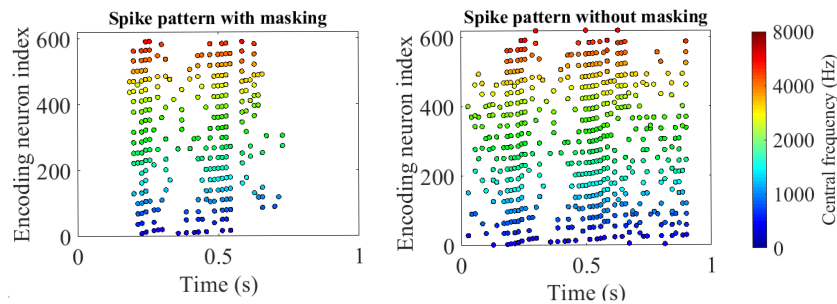
FIGURE 9 | Encoded spike patterns by threshold coding with/without masking. The two spike patterns are encoded from a speech utterance of "five" in the TIDIGITS dataset. The *x*-axis and *y*-axis represent the time and encoding neuron index. The positions of the colorful dots indicate the spike timings of the corresponding encoding neurons. The colors distinguish the centre frequencies of the cochlear filter bank. With auditory masking, the number of spikes reduces by nearly 50%, which are close to the 55% reducing rate of coding pulses as reported in Ambikairajah et al. (1997).

TABLE 1 | Parameters of neural threshold encoding for the speech and audio databases.

Parameter	Value
Window size	30 ms
Stride size	15 ms
Frequency range	[200 Hz, 8000 Hz]
Sampling rate	20 kHz

TABLE 2 | Cochlear filter parameters: we use a total of 20 cochlear filters in the BAE front-end.

Cochlear filter index	Centre frequency (Hz)	Bandwidth (Hz)
1	200.2	69.3
2	238.3	83.0
3	283.2	98.6
4	336.4	117.2
5	400.4	139.6
6	476.1	166.0
7	565.9	197.3
8	672.3	234.4
9	800.8	278.3
10	952.1	331.1
11	1131.3	394.5
12	1345.2	468.8
13	1600.6	557.6
14	1903.3	663.1
15	2263.7	788.1
16	2690.9	937.5
17	3200.2	1114.3
18	3805.7	1325.2
19	4525.9	1576.2
20	8000.5	6949.2

The center frequency and bandwidth of each filter are listed.

corpus has a vocabulary of 6224 words, which is larger than that of TIDIGITS.

Our proposed BAE scheme is next evaluated in the following sections, using both reconstruction and speech pattern recognition experiments.

Audio Reconstruction From Masked Patterns

According to Eq. 5, we adopt the binary auditory mask $\Phi_{K \times N}(\phi_{ij})$ which either fully encodes or ignores an acoustic event. It is suggested in auditory theory (Ambikairajah et al., 1997) that partial masking may exist in the frequency domain, especially in the presence of rich frequency tones. We would like to evaluate the masking effect in the BAE front-end both objectively and subjectively.

We begin by reconstructing the spike trains into speech signals, and then evaluate the speech quality using several objective speech quality measures: Perceptual Evaluation of Speech Quality (PESQ), Root Mean Square Error (RMSE), and Signal to Distortion Ratio (SDR). The PESQ, defined in Beerends et al. (2002) and Rix et al. (2002), is standardized as ITU-T recommendation P.862 for speech quality test methodology (Recommendation P.862, 2001). The core principle of PESQ is the use of the human auditory perception model (Rix et al., 2001) for speech quality assessment. For speech coding, especially the perceptual masking proposed in this paper, the PESQ measure could correctly distinguish between audible and inaudible distortions and thus assess the impact of perceptually masked coding noise. Besides, the PESQ is also used in the assessment of MPEG audio coding where auditory masking is involved. In this paper, the PESQ scores are further converted to MOS-LQO (Mean Opinion Score-Listening Quality Objective) scores ranging from 1 to 5, which are more intuitive for assessing speech quality. The mapping function is obtained from ITU-T Recommendation P.862.1 (ITU-T Recommendation, 2003). **Table 3** defines the MOS scales and their corresponding speech quality subjective descriptions (Recommendation P.800, 1996).

TABLE 3 | Perceptual evaluation of speech quality (MOS) scores and their corresponding perceptual speech quality subjective assessments.

PESQ (MOS) scores	5	4	3	2	1
Speech quality	Excellent	Good	Fair	Poor	Bad

FIGURE 10 | The reconstruction from a spike pattern into a speech signal. Parallel streams of threshold- encoded spike trains that represent the dynamics of multiple frequency channels are first decoded into sub-band digital signals. The sub-band signals are further fed into a series of synthesis filters, which are built inversely from the corresponding analysis cochlear filters as in **Figure 6**. The synthesis filters compensate for the gains from the analysis filters for each frequency bin. Finally, the outputs from the synthesis filter banks sum up to generate the reconstructed speech signal.
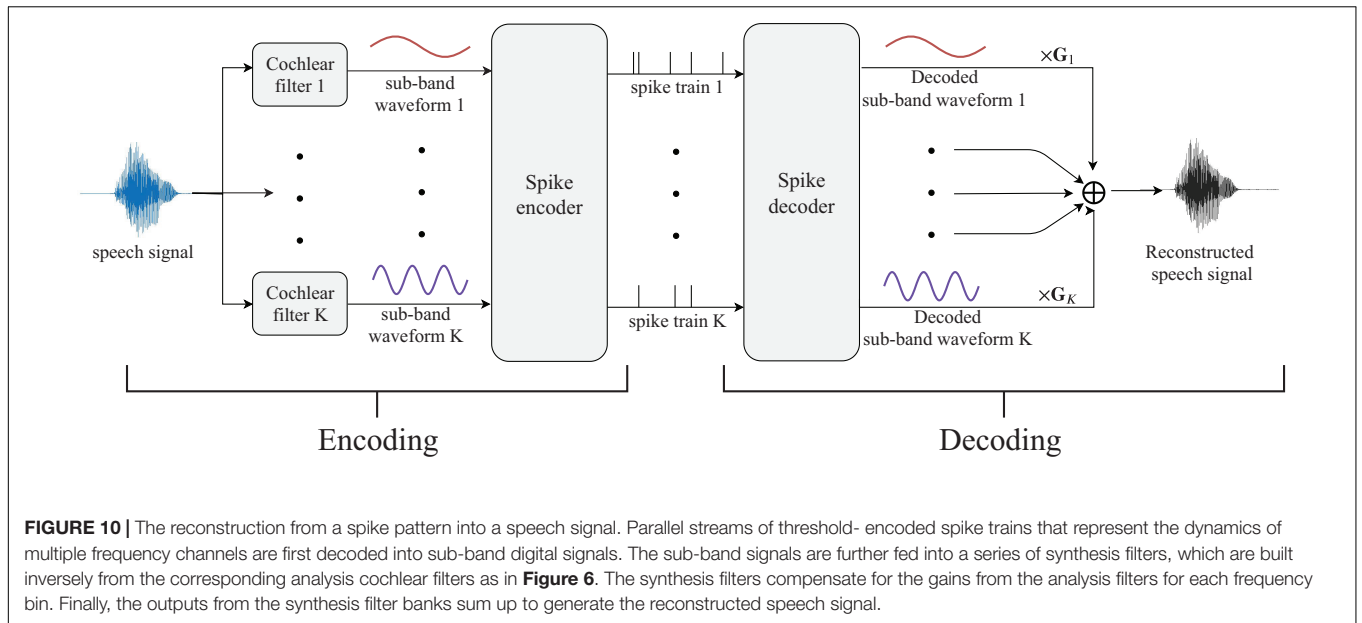
TABLE 4 | The objective audio quality measures of the reconstructed audio signals for environmental sounds dataset RWCP.

Reconstructed signal	PESQ	RMSE	SDR (dB)	Reduced rate (%)
$\hat{s}_{raw}$	4.53	2.79×10^{-4}	38.96	0
$\hat{s}_{mask}$	4.15	5.67×10^{-4}	33.13	39.38
$\hat{s}_{random}$	2.81	1.85×10^{-2}	4.95	40.07

The reduced rates refer to the ratio of masked spikes.

TABLE 5 | The objective speech quality measures of the reconstructed speech signals for the spoken digits dataset TIDIGITS.

Reconstructed signal	PESQ	RMSE	SDR (dB)	Reduced rate (%)
$\hat{s}_{raw}$	4.54	4.78×10^{-4}	34.60	0
$\hat{s}_{mask}$	4.43	7.49×10^{-4}	29.94	50.48
$\hat{s}_{2 \times threshold}$	3.80	8.50×10^{-3}	5.65	50.80
$\hat{s}_{random}$	2.92	1.05×10^{-2}	4.76	49.91

The reduced rates refer to the ratio of masked spikes.

Besides PESQ, the RMSE (Eq. 12) and Expand SDR (Eq. 13) measures are also reported, where x_i and $\hat{x}_i$ denote the ith time-domain sample of the original and reconstructed speech signals $\boldsymbol{x}_{1 \times M}$ and $\hat{\boldsymbol{x}}_{1 \times M}$, respectively.

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (x_i - \hat{x}_i)^2} \qquad (12)$$

$$SDR = 10 \log_{10} \left(\frac{\sum_{i=1}^{M} (x_i)^2}{\sum_{i=1}^{M} (x_i - \hat{x}_i)^2} \right) \qquad (13)$$

For comparison, we compare three groups of reconstructed speech signals: (1) the reconstructed signal $\hat{s}_{mask}$ from spike trains

TABLE 6 | The objective speech quality measures of the reconstructed speech signals for the continuous and large-vocabulary speech dataset TIMIT.

Reconstructed signal	PESQ	RMSE	SDR (dB)	Reduced rate (%)
$\hat{s}_{raw}$	4.54	1.23×10^{-4}	42.28	0
$\hat{s}_{mask}$	4.44	3.10×10^{-4}	34.02	29.33
$\hat{s}_{random}$	2.35	9.20×10^{-3}	4.83	30.8

The reduced rates refer to the ratio of masked spikes.

with auditory masking; (2) the reconstructed signal $\hat{s}_{raw}$ from raw spike trains without auditory masking; and (3) the reconstructed signal $\hat{s}_{random}$ from randomly masked spike trains.

Figure 10 depicts the flowchart of the reconstruction process. The left and right panels represent the spike encoding and decoding processes. The raw speech signals are first decomposed by a series of cochlear analysis filters, generating parallel streams of sub-band signals as in **Figure 7** (b). The 20 sub-band waveforms are encoded into spike times with masking strategies and then decoded back to sub-band speech signals. The reproduced sub-band waveforms 1 to K (20 in this work) are gain-weighted and summed to form the reconstructed speech signal for perceptual quality evaluation. Since the cochlear filters decompose the input signal by various weighting gains in different frequency bands, the weighting gains in the decoding part represent the inverse processing of the cochlear filters.

The audio quality of the three groups of reconstructed signals is compared, as reported in **Tables 4–6**. For a fair comparison, we first simulate a random masking effect by randomly dropping the same amount of spikes as that of the auditory masking. We further simulate a masking effect by doubling the firing thresholds that are used in **Figure 7** (f) purely according to energy level. The reconstructed signals from such elevated thresholds

TABLE 7 | Spiking neural network architectures for RWCP sound classification and TIDIGITS speech recognition tasks.

Dataset and task	Input layer	Output layer	Learning rule
RWCP sound classification	1 × 620 encoding neurons	1 × 10 LIF neurons	Tempotron
TIDIGITS speech recognition	1 × 620 encoding neurons	1 × 11 LIF neurons	MPD-AL

LIF refers to the Leaky-Integrate and Fire spiking neuron.

TABLE 8 | The RWCP classification accuracy for different neural encoding schemes and the average spike rates.

Neural encodings-classifier model	Classification accuracy (%)	# Spikes/ second
BAE-tempotron	99.5	245
Latency coding-tempotron	10.1	1, 598
Phase coding-tempotron	10.1	1, 598
Population coding-tempotron	99.0	4, 627
Threshold coding-tempotron	99.5	557
MFCC-HMM Dennis et al., 2013	90.0	–
LSF-SNN Dennis et al., 2013	98.5	–
LTF-SNN Xiao et al., 2016	97.5	–
DKP-SNN Yao et al., 2019	99.1	–

The average spike rates refer to the average number of spikes per second when the same audio samples are encoded.

are denoted as $\hat{s}_{2 \times \text{threshold}}$. The raw spike patterns without any masking are used as a reference.

The perceptual quality scores of the $\hat{s}_{mask}$ and $\hat{s}_{raw}$ are rather close at a high level of approximately 4.5, which suggests a satisfying subjective quality between "Excellent" and "Good" according to **Table 3**. It is noted that the speech signals with random masking are perceived as "Poor" in quality. Besides the PESQ, the other two measures also lead to the same conclusion. The RMSE of $\hat{s}_{raw}$ and $\hat{s}_{mask}$ are approximately two orders of magnitude larger than that of the $\hat{s}_{random}$; the SDRs also show a great gap.

Sound Classification by SNN for RWCP Dataset

Firstly, the BAE encoding is evaluated by a sound classification task on the RWCP sound scene database (Nakamura et al., 2000), which records the non-speech environmental sounds with rich and diverse frequency components. There are 10 sound categories in this database: bells, bottle, buzzer, cymbals, horn, kara, metal, phone, ring, and whistle. The sound recordings, each clip lasting for several seconds, are encoded into sparse spatio-temporal spike patterns and classified by a supervised learning SNN. We adopt a network structure that is shown in **Table 7** with Tempotron synaptic learning rule (Gütig and Sompolinsky, 2006).

Besides the BAE scheme, we also implement several other neural encoding schemes for the threshold coding (**Figure 7**) (f), such as latency coding (Gollisch and Meister, 2008), phase coding (Giraud and Poeppel, 2012), population coding (Dean et al., 2005), etc. The detailed implementation can be found

at (Pan et al., 2019). The classification accuracy, as well as the average spike rates are summarized in **Table 8**.

The results in **Table 8** show that our BAE scheme achieves the best classification accuracy (99.5%) with a spiking rate of 245 spikes per second, across the other commonly used neural encoding schemes. The results suggest that the proposed BAE encoding scheme is both effective in pattern classification, and energy-efficient.

Speech Recognition by SNN for TIDIGITS Dataset

In this section, we evaluate the BAE scheme in a speech recognition task, which also aims to evaluate the coding fidelity of our proposed methodology. The spike patterns encoded from TIDIGITS speech dataset are fed into an SNN, and the outputs correspond to the labels of which spoken digits the patterns are encoded from. The spike learning rule is Membrane Potential Driven Aggregate-Label Learning (MPD-AL) (Zhang et al., 2019). The network structure is given in **Table 7**.

To evaluate the effectiveness and robustness of the BAE front-end, we compare the classification performances between spike patterns with and without auditory masking. Gaussian noise, measured by Signal-to-Noise Ratio (SNR) in dB, is added to the original speech waveforms before the encoding process. **Figure 11** shows the classification accuracy under noisy conditions and in the clean condition. Besides, we also compare our scheme with the other benchmarking results in **Table 9**.

The results show that the pattern classification accuracy for masked patterns is slightly higher than those of the unmasked patterns, under different test conditions. Above all, referring to **Table 5**, our proposed BAE scheme helps to reduce nearly half of the spikes, which is a dramatic improvement in coding efficiency.

Large Vocabulary Speech Recognition for TIMIT Dataset

In section "Biologically Plausible Auditory Encoding With Masking Effects," we present how the TIMIT dataset has been encoded into spike trains, which we henceforth refer to as Spike-TIMIT. We next train a recurrent neural network, the LSTM (Hochreiter and Schmidhuber, 1997) on both the original TIMIT and Spike-TIMIT datasets, with the CTC loss function (Graves et al., 2006). The LSTM architectures and the system performances on various TIMIT datasets are summarized in **Table 10**. The LSTM networks are adopted from Tensorpack[1]. We obtained a PER of 27% and 30%, respectively, for the TIMIT and Spike-TIMIT datasets. For comparison, we also report the accuracy for Spike-TIMIT without masking that shows 28% PER.

---

[1]Wu Y. et al. (2016). *Tensorpack*. Available at: https://github.com/tensorpack/.

FIGURE 11 | The classification accuracy for the Spike-TIDIGITS dataset under different signal-to-noise ratios, with or without masking effects. The accuracy for the Spike-TIDIGITS with masking effects is slightly higher than that for the Spike-TIDIGITS without masking effects.
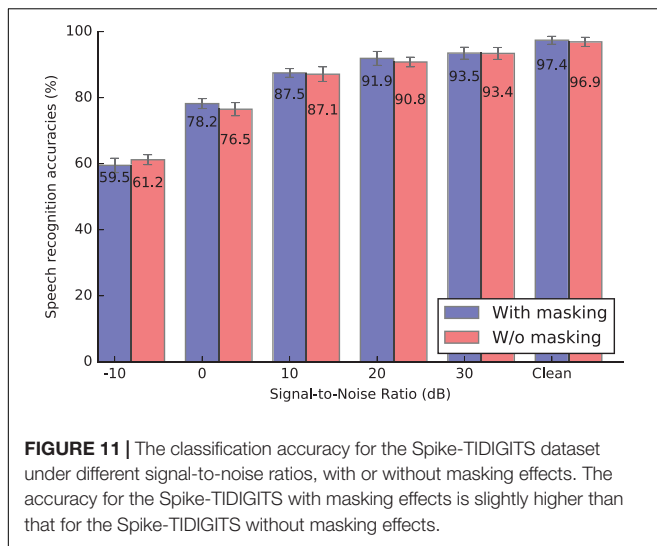
TABLE 9 | The TIDIGITS speech recognition accuracy for different neural encoding schemes.

Models of TIDIGITS speech recognition	Classification accuracy (%)
Biologically plausible auditory encoding (BAE)-MPDAL	97.4
Liquid state machine Zhang et al., 2015	92.3
SNN-SVM Tavanaei and Maida, 2017b	91.0
Spiking CNN-HMM Tavanaei and Maida, 2017a	96.0
AER silicon cochlea-SVM Abdollahi and Liu, 2011	95.6
Auditory spectrogram-SVM Abdollahi and Liu, 2011	78.7

We also notice some improvement in accuracy when dropout is introduced for Spike-TIMIT but not for TIMIT. Although the phone error rates are quite close across these datasets, we observe that the Spike-TIMIT derived from our proposed BAE scheme shows the highest spike efficiency (30% spike reduction in **Table 6**), which further improves the energy efficiency.

DISCUSSION

In this paper, we propose a BAE scheme, especially for speech signals. The encoding scheme is inspired by the modeling of the human auditory sensory system, which is composed of spectral analysis, neural spike coding, as well as the psychoacoustic perception model. We adopt three criteria for formulating the auditory encoding scheme.

For the spectral analysis part, a time-domain event-based cochlear filter bank is applied, with the perceptual scale of centre frequencies and bandwidths. The key feature of the spectral analysis is the parallel implementation of time-domain convolution. One of the most important properties of SNN is its asynchronous processing. The parallel implementation makes the neural encoding scheme a friendly frontend for any SNN processing. The neural encoding scheme, the threshold code in our case, helps to generate a sparse and representative spike patterns for efficient computing in the SNN classifier. The

threshold code helps in two aspects: firstly it tracks the trajectory of the spectral power tuning curves, which represents the features in the acoustic dynamics; secondly, the threshold code, as a form of population neural code, is able to project the dynamics in the time domain onto the spatial domain, which facilitates the parallel processing of spiking neurons on cognitive tasks (Pan et al., 2019). Another key component of the BAE front-end is the implementation of auditory masking that benefits from findings in human psychoacoustic experiments. The integrated auditory encoding scheme fulfills the three proposed design criteria. We have evaluated our BAE scheme through signal reconstruction and speech recognition experiments giving very promising results. To share our study with the research community, the spike-version of TIDIGITS and TIMIT speech corpus, namely, Spike-TIDIGITS, and Spike-TIMIT, will be made available as benchmarking datasets.

Figure 12 illustrates some interesting findings in our proposed auditory masking strategy. The upper, middle and lower panels of **Figure 12** represent three speech utterances from the TIDIGITS dataset. The first and second column illustrates the encoded spike patterns with and without auditory masking effects. It is apparent that a large number of spikes are removed. The graphs in the third column demonstrate the membrane potential of the output neuron in the trained SNN classifier after being fed with both patterns during the testing phase. For example, the LIF neuron in (**Figure 12C**) responds to the speech utterance of "six." As such, the encoded pattern of spoken "six," as in (**Figures 12A,B**) will trigger the corresponding neuron to fire a spike in the testing phase. The sub-figure (c) demonstrates that though the sub-threshold membrane potentials of masking/unmasking patterns have different trajectories, the two membrane potential curves will exceed the firing threshold (which is 1 in this example) at close timing. Similar results are observed in **Figures 12F,I**. The spike patterns with or without auditory masking lead to similar neuronal responses, either in spiking activities (firing or not) or in membrane potential dynamics, as observed in **Figures 12C,F,I**. It is interesting to observe that auditory masking has little impact on the neuronal dynamics. As a psychoacoustic phenomenon, the auditory masking is always studied using listening tests. It remains unclear how the human auditory system responds to auditory masking. **Figure 12** provides an answer to the same question from an SNN perspective.

The parameters of auditory masking effects in this work, such as the exponential decaying parameter c in Eq. 2, or the cross-channel simultaneous masking thresholds in **Figure 2**, are all derived in the acoustic model of MPEG-1 Layer III standard (Fogg et al., 2007) and tuned according to the particular tasks. However, from a neuroscience point of view, our brain is adaptive to different environments. This suggests that the parameters could be optimized by machine learning methodology, for different tasks and datasets. Also, the threshold neural code, which encodes the dynamics of the spectrum using threshold-crossing events, relies heavily on the choice of thresholds. We use 15 uniformly distributed thresholds for simplicity. We note that the recording of threshold-crossing events is analogous to quantization in

TABLE 10 | Phone error rates (PER) for TIMIT and Spike-TIMIT datasets, using different LSTM structures.

Dataset	Network architecture	PER (%)
TIMIT	$1 \times 39 - 1 \times 512$ LSTM $- 1 \times 512$ LSTM $- 1 \times 62$	27
TIMIT	$1 \times 39 - 1 \times 1024$ LSTM $- 1 \times 1024$ LSTM $- 1 \times 62$	27
Spike-TIMIT	$1 \times 620 -$ Dropout $(0.2) - 1 \times 512$ LSTM $- 1 \times 512$ LSTM $- 1 \times 62$	30
Spike-TIMIT	$1 \times 620 -$ Dropout $(0.2) - 1 \times 1024$ LSTM $- 1 \times 1024$ LSTM $- 1 \times 62$	35
Spike-TIMIT-w/o mask	$1 \times 620 -$ Dropout $(0.2) - 1 \times 512$ LSTM $- 1 \times 512$ LSTM $- 1 \times 62$	28
Spike-TIMIT-w/o mask	$1 \times 620 -$ Dropout $(0.2) - 1 \times 1024$ LSTM $- 1 \times 1024$ LSTM $- 1 \times 62$	30
TIMIT[1]	$1 \times 39 - 1 \times 512$ LSTM $- 1 \times 512$ LSTM $- 1 \times 62$	28

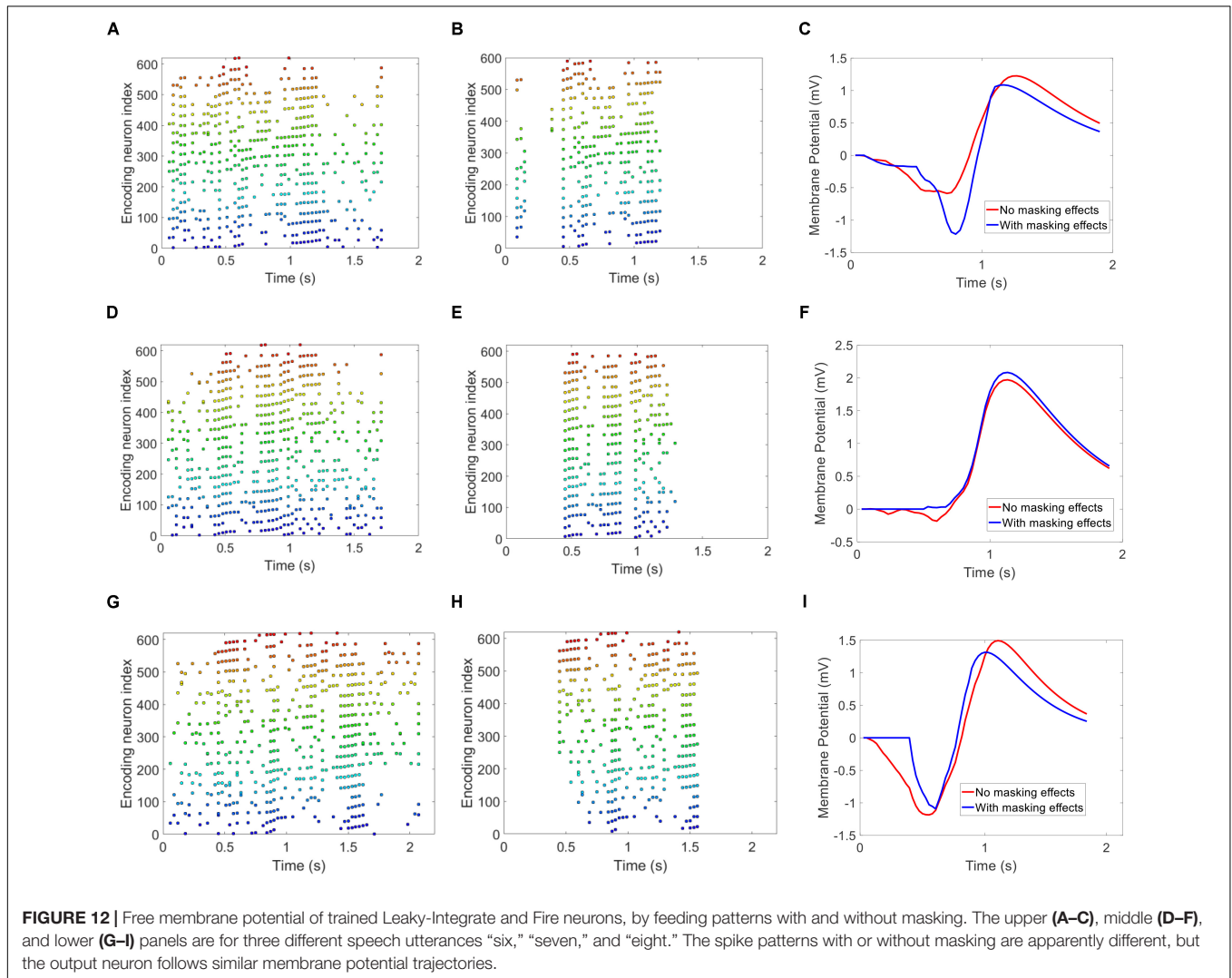[1]Wu Y. et al. (2016). Tensorpack. Available at: https://github.com/tensorpack/.

FIGURE 12 | Free membrane potential of trained Leaky-Integrate and Fire neurons, by feeding patterns with and without masking. The upper **(A–C)**, middle **(D–F)**, and lower **(G–I)** panels are for three different speech utterances "six," "seven," and "eight." The spike patterns with or without masking are apparently different, but the output neuron follows similar membrane potential trajectories.

digital coding, that the maximal coding efficiency (maximal information being conveyed constrained by the numbers of neurons or spikes) may be derived using an information-theoretic approach. The Efficient Coding Hypothesis (ECH) (Barlow, 1961; Srinivasan et al., 1982) that describes the link between neural encoding and information theory could provide us the theoretical framework to determine the optimal threshold distribution in the neural threshold

code. It may also otherwise be learned using machine learning techniques.

CONCLUSION

Our proposed BAE scheme, motivated by the human auditory sensory system, could encode temporal audio data into spike

patterns that are sparse, efficient, and friendly to SNN learning rules. It is both efficient and effective. We use the BAE scheme to encode popular speech datasets, namely, TIDIGITS and TIMIT into their spike versions: Spike-TIDIGITS and Spike-TIMIT. The two spike datasets are to be published as benchmarking datasets, in the hope of improving SNN-based classifiers.

DATA AVAILABILITY STATEMENT

The datasets Spike-TIDIGITS and Spike-TIMIT for this study can be found at https://github.com/pandarialTJU/Biologically-plausible-Auditory-Encoding/tree/master.

REFERENCES

Abdollahi, M. and Liu, S.-C. (2011). Speaker-independent isolated digit recognition using an aer silicon cochlea. *Proceeding of the 2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (Piscataway, NJ: IEEE), 269–272

Ambikairajah, E., Davis, A., and Wong, W. (1997). Auditory masking and mpeg-1 audio compression. *Electron. Commun. Eng. J.* 9, 165–175 doi: 10.1049/ecej: 19970403

Ambikairajah, E., Epps, J., and Lin, L. (2001). Wideband speech and audio coding using gammatone filter banks. *Proceeding of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing.* (Piscataway, NJ: IEEE) 2, 773–776

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., et al. (2017). A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (Piscataway, NJ: IEEE) 7243–7252

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

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. *Psychol. Bull.* 78, 266–278 doi: 10.1037/h0033117

Arnal, L. H. and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398 doi: 10.1016/j.tics.2012.05.003

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sens. Commun.* 1, 217–234

Beerends, J. G., Hekstra, A. P., Rix, A. W., and Hollier, M. P. (2002). Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model. *J. Audio Eng. Soc.* 50, 765–778

Besacier, L., Grassi, S., Dufaux, A., Ansorge, M., and Pellandini, F. (2000). Gsm speech coding and speaker recognition. In *Proceeding of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing,* (Piscataway, NJ: IEEE) 2, II1085-II1088

Bohte, S. M., Kok, J. N., and La Poutre, H. (2002). Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing* 48, 17–37 doi: 10.1016/s0925-2312(01)00658-0

Brown, J. C. (1991). Calculation of a constant q spectral transform. *J. Acoust. Soc. Am.* 89, 425–434 doi: 10.1121/1.400476

Brown, J. C. and Puckette, M. S. (1992). An efficient algorithm for the calculation of a constant q transform. *J. Acoust. Soc. Am.* 92, 2698–2701 doi: 10.1121/1.404385

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Commun.* 34, 267–285 doi: 10.1016/s0167-6393(00)00034-0

Darabkh, K. A., Haddad, L., Sweidan, S. Z., Hawa, M., Saifan, R., and Alnabelsi, S. H. (2018). An efficient speech recognition system for arm-disabled students based on isolated words. *Comput. Appl. Eng. Educ.* 26, 285–301 doi: 10.1002/cae.21884

Dean, I., Harper, N. S., and McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neurosci.* 8, 1684–1689 doi: 10.1038/nn1541

Dennis, J., Yu, Q., Tang, H., Tran, H. D., and Li, H. (2013). Temporal coding of local spectrogram features for robust sound recognition. In *Proceeding of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE),* (Piscataway, NJ: IEEE) 803–807

Ehret, G. (1997). The auditory cortex. *J. Comp. Physiol. A* 181, 547–557

El-Solh, A., Cuhadar, A., and Goubran, R. A. (2007). Evaluation of speech enhancement techniques for speaker identification in noisy environments. In *Proceeding of the 9th IEEE International Symposium on Multimedia Workshops (ISMW 2007) (IEEE),* (Piscataway, NJ: IEEE) 235–239

Fogg, C., LeGall, D. J., Mitchell, J. L., and Pennebaker, W. B. (2007). *MPEG Video Compression Standard* (Berlin: Springer Science & Business Media)

Garofolo, J. S. (1993). *Timit Acoustic Phonetic Continuous Speech Corpus.* Philadelphia, PA: Linguistic Data Consortium.

Ghosh-Dastidar, S. and Adeli, H. (2009). Spiking neural networks. *Intl. J. Neural Sys.* 19, 295–308

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–5187 doi: 10.1038/nn.3063

Gollisch, T. and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *science* 319, 1108–1111 doi: 10.1126/science.1149639

Graves, A., Fernàndez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (New York,NY: ACM), 369–376

Graves, A., Mohamed, A.-R., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceeding of the 2013 IEEE international conference on acoustics, speech and signal processing* (Piscataway, NJ: IEEE), 6645–6649

Gütig, R. (2016). Spiking neurons can discover predictive features by aggregate-label learning. *Science* 351:aab4113 doi: 10.1126/science.aab4113

Gütig, R. and Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing–based decisions. *Nat. Neurosci.* 9, 420–428 doi: 10.1038/nn1643

Gütig, R. and Sompolinsky, H. (2009). Time-warp–invariant neuronal processing. *PLoS Biol.* 7:e1000141 doi: 10.1371/journal.pbio.1000141

Harris, D. M. and Dallos, P. (1979). Forward masking of auditory nerve fiber responses. *J. Neurophysiol.* 42, 1083–1107 doi: 10.1152/jn.1979.42.4.1083

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural computation 9, 1735–1780

Hohmann, V. (2002). Frequency analysis and synthesis using a gammatone filterbank. *Acta Acust. United Acust.* 88, 433–442

Hopfield, J. (2004). Encoding for computation: recognizing brief dynamical patterns by exploiting effects of weak rhythms on action-potential timing. *Proc. Natl. Acad Sci.* 101, 6255–6260 doi: 10.1073/pnas.0401125101

AUTHOR CONTRIBUTIONS

ZP performed the experiments and wrote the manuscript. All authors contributed to the experiments design, result interpretation, and writing.

FUNDING

ITU-T Recommendation, (2003). *862.1: Mapping Function for Transforming p. 862 Raw Result Scores to Mos-Lqo.* Geneva: International Telecommunication Union.

Iyer, L. R., Chua, Y., and Li, H. (2018). Is neuromorphic mnist neuromorphic? analyzing the discriminative power of neuromorphic datasets in the time domain. *arXiv* [Preprint] 1807.01013

Lagerstrom, K. (2001). *Design and Implementation of An Mpeg-1 Layer iii Audio Decoder.* Gothenburg: Chalmers University of Technology.

Leonard, R. G. and Doddington, G. (1993). *Tidigits ldc93s10.* Philadelphia, PA: Linguistic Data Consortium.

Liu, S.-C. and Delbruck, T. (2010). Neuromorphic sensory systems. *Curr. Opin. Neurobiol.* 20, 288–295 doi: 10.1016/j.conb.2010.03.007

Loiselle, S., Rouat, J., Pressnitzer, D., and Thorpe, S. (2005). Exploration of rank order coding with spiking neural networks for speech recognition. In *Proceeding of the Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (IEEE)*, (Piscataway, NJ: IEEE) 4, 2076–2080

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671 doi: 10.1016/s0893-6080(97)00011-7

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* 116, 374–388

Messaoud, Z. B. and Hamida, A. B. (2011). Combining formant frequency based on variable order lpc coding with acoustic features for timit phone recognition. *Int. J. Speech Technol.* 14:393 doi: 10.1007/s10772-011-9119-z

Mohamed, A.-R., Dahl, G. E., and Hinton, G. (2011). Acoustic modeling using deep belief networks. *Proceeding of the IEEE transactions on audio, speech, and language processing* (Piscataway, NJ: IEEE) 20, 14–22 doi: 10.1109/tasl.2011.2109382

Nadasdy, Z. (2009). Information encoding and reconstruction from the phase of action potentials. *Front. Sys. Neurosci.* 3:6 doi: 10.3389/neuro.06.006.2009

Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., and Yamada, T. (2000). Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of the Second International Conference on Language Resources and Evaluation* France: LREC

Orchard, G., Jayawant, A., Cohen, G. K., and Thakor, N. (2015). Converting static image datasets to spiking neuromorphic datasets using saccades. *Front. Neurosci.* 9:437 doi: 10.3389/fnins.2015.00437

Pan, Z., Li, H., Wu, J., and Chua, Y. (2018). An event-based cochlear filter temporal encoding scheme for speech signals. In *Proceeding of the 2018 International Joint Conference on Neural Networks (IJCNN)* (Piscataway, NJ: IEEE), 1–8

Pan, Z., Wu, J., Chua, Y., Zhang, M., and Li, H. (2019). Neural population coding for effective temporal classification. In *Proceeding of the 2019 International Joint Conference on Neural Networks (IJCNN)* (Piscataway, NJ: IEEE), 1–8

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. In *Proceeding of the a meeting of the IOC Speech Group on Auditory Modelling at RSRE.* (Malvern: RSRE).

Ponulak, F. and Kasiński, A. (2010). Supervised learning in spiking neural networks with resume: sequence learning, classification, and spike shifting. *Neural Comput.* 22, 467–510 doi: 10.1162/neco.2009.11-08-901

Recommendation P.800 (1996). *P. 800: Methods for subjective determination of transmission quality.* Geneva: International Telecommunication Union 22.

Recommendation P.862 (2001). Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Geneva: International Telecommunication Union.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* (Piscataway, NJ: IEEE). 779–788

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proceeding of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)* (Piscataway, NJ: IEEE) 2, 749–752

Rix, A. W., Hollier, M. P., Hekstra, A. P., and Beerends, J. G. (2002). Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech

quality assessment part i–time-delay compensation. *J. Audio Eng. Soc.* 50, 755–764

Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., and Liu, S.-C. (2017). Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Front. Neurosci.* 11:682 doi: 10.3389/fnins.2017.00682

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186 doi: 10.1016/j.tics.2008.02.003

Shrestha, S. B. and Orchard, G. (2018). Slayer: spike layer error reassignment in time. In *Proceeding of the Advances in Neural Information Processing Systems*, Montreal, 1412–1421

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint] *1409.1556*

Smith, J. O. and Abel, J. S. (1999). Bark and erb bilinear transforms. *IEEE Trans. Speech Audio Process.* 7, 697–708 doi: 10.1109/89.799695

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. Series B. Biol. Sci.* 216, 427–459 doi: 10.1098/rspb.1982.0085

Taherkhani, A., Belatreche, A., Li, Y., and Maguire, L. P. (2015). Dl-resume: A delay learning-based remote supervised method for spiking neurons. *IEEE Trans. Neural Netw. and Learn. Syst.* 26, 3137–3149 doi: 10.1109/TNNLS.2015.2404938

Tamazin, M., Gouda, A., and Khedr, M. (2019). Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients. *Appl. Sci.* 9, 2166 doi: 10.3390/app9102166

Tavanaei, A. and Maida, A. (2017a). Bio-inspired multi-layer spiking neural network extracts discriminative features from speech signals. In *International Conference on Neural Information Processing* (Berlin: Springer), 899–908 doi: 10.1007/978-3-319-70136-3_95

Tavanaei, A. and Maida, A. S. (2017b). A spiking network that learns to extract spike signatures from speech signals. *Neurocomputing* 240, 191–199 doi: 10.1016/j.neucom.2017.01.088

Tobias, J. (2012). *Foundations of Modern Auditory Theory*, Amsterdam: Elsevier.

Veit, A., Matera, T., Neumann, L., Matas, J., and Belongie, S. (2016). Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv* [Preprint] 1601.07140

Wu, J., Chua, Y., and Li, H. (2018a). A biologically plausible speech recognition framework based on spiking neural networks. In *Proceeding of the 2018 International Joint Conference on Neural Networks (IJCNN)*, (Piscataway, NJ: IEEE), 1–8

Wu, J., Chua, Y., Zhang, M., Li, H., and Tan, K. C. (2018b). A spiking neural network framework for robust sound classification. *Front. Neurosci.* 12:836. doi: 10.3389/fnins.2018.00836

Xiao, R., Yan, R., Tang, H., and Tan, K. C. (2016). A spiking neural network model for sound recognition. In *Proceeding of the International Conference on Cognitive Systems and Signal Processing* (Berlin: Springer), 584–594 doi: 10.1007/978-981-10-5230-9_57

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

Yao, Y., Yu, Q., Wang, L., and Dang, J. (2019). A spiking neural network with distributed keypoint encoding for robust sound recognition. In *Proceeding of the 2019 International Joint Conference on Neural Networks (IJCNN)* (Piscataway, NJ: IEEE), 1–8

Zhang, M., Qu, H., Belatreche, A., Chen, Y., and Yi, Z. (2018). A highly effective and robust membrane potential-driven supervised learning method for spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems* (Piscataway, NJ: IEEE) 1–15

Zhang, M., Qu, H., Belatreche, A., and Xie, X. (2017). Empd: An efficient membrane potential driven supervised learning algorithm for spiking neurons. *IEEE Trans. Cogn. Dev. Syst.* 10, 151–162 doi: 10.1109/tcds.2017.2651943

Zhang, M., Wu, J., Chua, Y., Luo, X., Pan, Z., Liu, D., et al. (2019). Mpd-al: an efficient membrane potential driven aggregate-label learning algorithm for spiking neurons. *Proc. AAAI Conf. Artif. Intell.* 33, 1327–1334 doi: 10.1609/aaai.v33i01.33011327

Zhang, Y., Li, P., Jin, Y., and Choe, Y. (2015). A digital liquid state machine with biologically inspired learning and its application to speech recognition. *IEEE Trans. Neural Netw. Learn. Sys.* 26, 2635–2649 doi: 10.1109/TNNLS.2015.2388544

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., and Zou, Y. (2016). Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv*[Preprint] 1606.06160