SPIKING-LEAF: A LEARNABLE AUDITORY FRONT-END FOR SPIKING NEURAL NETWORKS

Zeyang Song[1]*, Jibin Wu*[2,*]*, Malu Zhang*[3]*, Mike Zheng Shou*[1]*, Haizhou Li*[1,4]

[1]National University of Singapore, Singapore
[2]The Hong Kong Polytechnic University, Hong Kong SAR, China
[3]University of Electronic Science and Technology of China, China
[4]Shenzhen Research Institute of Big Data, School of Data Science,
The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Brain-inspired spiking neural networks (SNNs) have demonstrated great potential for temporal signal processing. However, their performance in speech processing remains limited due to the lack of an effective auditory front-end. To address this limitation, we introduce Spiking-LEAF, a learnable auditory front-end meticulously designed for SNN-based speech processing. Spiking-LEAF combines a learnable filter bank with a novel two-compartment spiking neuron model called IHC-LIF. The IHC-LIF neurons draw inspiration from the structure of inner hair cells (IHC) and they leverage segregated dendritic and somatic compartments to effectively capture multi-scale temporal dynamics of speech signals. Additionally, the IHC-LIF neurons incorporate the lateral feedback mechanism along with spike regularization loss to enhance spike encoding efficiency. On keyword spotting and speaker identification tasks, the proposed Spiking-LEAF outperforms both SOTA spiking auditory front-ends and conventional real-valued acoustic features in terms of classification accuracy, noise robustness, and encoding efficiency.

Index Terms— Spiking neural networks, speech recognition, learnable audio front-end, spike encoding

1 Introduction

Recently, the brain-inspired spiking neural networks (SNNs) have demonstrated superior performance in sequential modeling [1, 2]. However, their performance in speech processing tasks still lags behind that of state-of-the-art (SOTA) non-spiking artificial neural networks (ANNs) [3, 4, 5, 6, 7, 8, 9, 10, 11]. This is primarily due to the lack of an effective auditory front-end that can synergistically perform acoustic feature extraction and neural encoding with high efficacy and efficiency.

The existing SNN-based auditory front-ends first extract acoustic features from raw audio signals, followed by encoding these real-valued acoustic features into spike patterns that can be processed by the SNN. For feature extraction, many works directly adopt the frequently used acoustic features based on the Mel-scaled filter-bank [3, 4, 5] or the GammaTone filter-bank [12]. Despite the simplicity of this approach, these handcrafted filter-bank are found to be suboptimal in many tasks when compared to learnable filter-bank [13, 14, 15, 16]. In another vein of research, recent works have also looked into the neurophysiological process happening in the peripheral auditory system and developed more complex biophysical models to enhance the effectiveness of feature extraction [17, 18]. However, these methods not only require fine-tuning a large number of hyperparameters but are also computationally expensive for resource-constrained neuromorphic platforms.

For neural encoding, several methods have been proposed that follow the neurophysiological processes within the cochlea [17, 18]. For instance, Cramer et al. proposed a biologically inspired cochlear model with the model parameters directly taken from biological studies [17]. Additionally, other methods propose to encode the temporal variations of the speech signals that are critical for speech recognition. The Send on Delta (SOD) [19] and threshold coding methods [12, 20, 21], for instance, encode the positive and negative variations of signal amplitude into spike trains. However, these neural encoding methods lack many essential characteristics as seen in the human's peripheral auditory system that are known to be important for speech processing, such as feedback adaptation [22].

To address these limitations, we introduce a Spiking LEarnable Audio front-end model, called Spiking-LEAF. The Spiking-LEAF leverages a learnable auditory filter-bank to extract discriminative acoustic features. Furthermore,
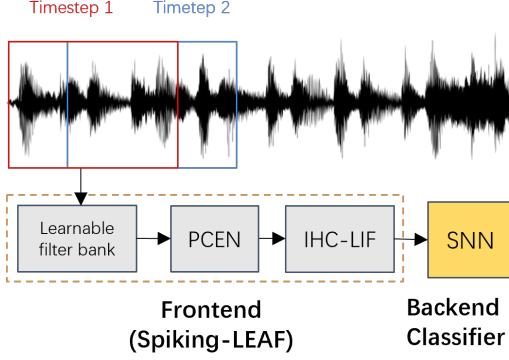
Fig. 1: The overall architecture of the proposed SNN-based speech processing framework.

Fig. 2: Computational graphs of LIF and IHC-LIF neurons.

inspired by the structure and dynamics of the inner hair cells (IHCs) within the cochlea, we further proposed a two-compartment neuron model for neural encoding, namely IHC-LIF neuron. Its two neuronal compartments work synergistically to capture the multi-scale temporal dynamics of speech signals. Additionally, the lateral inhibition mechanism along with spike regularization loss is incorporated to enhance the encoding efficiency. The main contributions of this paper can be summarized as follows:

- We propose a learnable auditory front-end for SNNs, enabling the joint optimization of feature extraction and neural encoding processes to achieve optimal performance in the given task.
- We propose a two-compartment spiking neuron model for neural encoding, called IHC-LIF, which can effectively extract multi-scale temporal information with high efficiency and noise robustness.
- Our proposed Spiking-LEAF shows high classification accuracy, noise robustness, and encoding efficiency on both keyword spotting and speaker identification tasks.

2 Methods

As shown in Fig. 1, similar to other existing auditory front-ends, the proposed Spiking-LEAF model consists of two parts responsible for feature extraction and neural encoding, respectively. For feature extraction, we apply the Gabor 1d-convolution filter bank along with the Per-Channel Energy Normalization (PCEN) to perform frequency analysis. Subsequently, the extracted acoustic feature is processed by the IHC-LIF neurons for neural encoding. Given that both the feature extraction and neural encoding parts are parameterized, they can be optimized jointly with the backend SNN classifier.

2.1 Parameterized acoustic feature extraction

In Spiking-LEAF, the feature extraction is performed with a 1d-convolution Gabor filter bank along with the PCEN that is tailored for dynamic range compression [23]. The Gabor 1d-convolution filters have been widely used in speech pro-
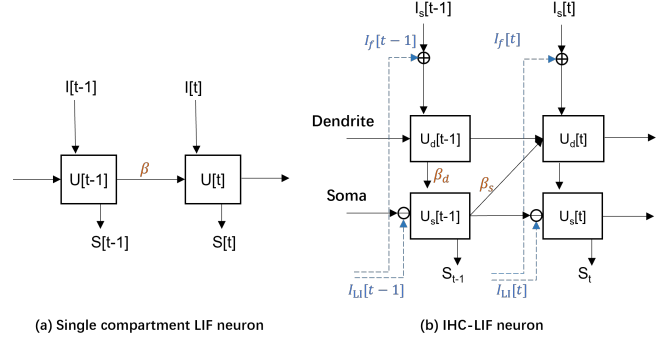
cessing [24, 16], and its formulation can be expressed as per:

$$\phi_n(t) = e^{i2\pi\eta_n t}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{t^2}{2\sigma_n^2}} \quad (1)$$

where η_n and σ_n denote learnable parameters that characterize the center frequency and bandwidth of filter n, respectively. In particular, for input audio with a sampling rate of 16 kHz, there are a total of 40 convolution filters, with a window length of 25ms ranging over $t = -L/2, ..., L/2$ ($L = 401$ samples), have been employed in Spiking-LEAF. These 1d-convolution filters are applied directly to the audio waveform x to get the time-frequency representation F.

Following the neurophysiological process in the peripheral auditory system, the PCEN [16, 23] has been applied subsequently to further compress the dynamic range of the obtained acoustic features:

$$PCEN(F(t,n)) = \left(\frac{F(t,n)}{(\varepsilon + M(t,n))^{\alpha_n} + \delta_n}\right)^{r_n} - \delta_n^{r_n} \quad (2)$$

$$M(t,n) = (1-s)M(t-1,n) + sF(t,n) \quad (3)$$

In Eqs. 2 and 3, $F(t,n)$ represents the time-frequency representation for channel n at time step t. r_n and α_n are coefficients that control the compression rate. The term $M(t,n)$ is the moving average of the time-frequency feature with a smoothing rate of s. Meanwhile, ε and δ_n stands for a positive offset introduced specifically to prevent the occurrence of imaginary numbers in PCEN.

2.2 Two-compartment spiking neuron model

The Leaky Integrate-and-Fire (LIF) neuron model [25], with a single neuronal compartment, has been widely used in brain simulation and neuromorphic computing [3, 4, 5, 7, 8]. The internal operations of a LIF neuron, as illustrated in Fig. 2 (a), can be expressed by the following discrete-time formulation:

$$I[t] = \Sigma_i w_i S[t-1] + b \quad (4)$$
$$U[t] = \beta * U[t-1] + I[t] - V_{th}S[t-1] \quad (5)$$
$$S[t] = \mathbb{H}(U[t] - V_{th}) \quad (6)$$

where $S[t-1]$ represents the input spike at time step t. $I[t]$ and $U[t]$ denote the transduced synaptic current and membrane potential, respectively. β is the membrane decaying constant that governs the information decaying rate within the LIF neuron. As the Heaviside step function indicated in Eq. 6, once the membrane potential exceeds the firing threshold V_{th}, an output spike will be emitted.

Despite its ubiquity and simplicity, the LIF model possesses inherent limitations when it comes to long-term information storage. These limitations arise from two main factors: the exponential leakage of its membrane potential and the resetting mechanism. These factors significantly affect the model's efficacy in sequential modeling. Motivated by the intricate structure of biological neurons, recent work has developed a two-compartment spiking neuron model, called TC-LIF, to address the limitations of the LIF neuron [26]. The neuronal dynamics of TC-LIF neurons are given as follows:

$$I[t] = \Sigma_i w_i S[t-1] + b \tag{7}$$

$$\begin{aligned} U_d[t] = U_d[t-1] + \beta_d * U_s[t-1] + I[t] \\ - \gamma * S[t-1] \end{aligned} \tag{8}$$

$$U_s[t] = U_s[t-1] + \beta_s * U_d[t-1] - V_{th}S[t-1] \tag{9}$$

$$S[t] = \mathbb{H}(U[t] - V_{th}) \tag{10}$$

where $U_d[t]$ and $U_s[t]$ represent the membrane potential of the dendritic and somatic compartments. The β_d and β_s are two learnable parameters that govern the interaction between dendritic and somatic compartments. Facilitated by the synergistic interaction between these two neuronal compartments, TC-LIF can retain both short-term and long-term information which is crucial for effective speech processing [26].

2.3 IHC-LIF neurons with lateral feedback

Neuroscience studies reveal that lateral feedback connections are pervasive in the peripheral auditory system, and they play an essential role in adjusting frequency sensitivity of auditory neurons [27]. Inspired by this finding, as depicted in Figure 2 (b), we further incorporate lateral feedback components into the dendritic compartment and somatic compartment of the TC-LIF neuron, represented by $I_f[t]$ and $I_{LI}[t]$ respectively. Specifically, each output spike will modulate the neighboring frequency bands with learnable weight matrices $ZeroDiag(W_f)$ and $ZeroDiag(W_{LI})$, whose diagonal entries are all zeros.

The lateral inhibition feedback of hair cells within the cochlea is found to detect sounds below the thermal noise level and in the presence of noise or masking sounds [28, 29]. Motivated by this finding, we further constrain the weight matrix $W_{LI} \geq 0$ to enforce lateral inhibitory feedback at the somatic compartment, which is responsible for spike generation. This will suppress the activity of neighboring neurons after the spike generation, amplifying the signal of the most activated neuron while suppressing other neurons. This results in a sparse yet informative spike representation of in-

Tasks	Front-end	Classifier Structure	Classifier Type	Test Accuracy (%)
KWS	Fbank [3]	512-512	Feedforward	83.03
	Fbank+LIF	512-512	Feedforward	85.24
	Heidelberg[17]	512-512	Feedforward	68.14
	Spiking-LEAF	512-512	Feedforward	**92.24**
	Speech2spike [30]	256-256-256	Feedforward	88.5
	Spiking-LEAF	256-256-256	Feedforward	**90.47**
	Fbank [3]	512-512	Recurrent	93.58
	Fbank+LIF	512-512	Recurrent	92.04
	Spiking-LEAF	512-512	Recurrent	**93.95**
SI	Fbank	512-512	Feedforward	29.42
	Fbank+LIF	512-512	Feedforward	27.23
	Spiking-LEAF	512-512	Feedforward	**30.17**
	Fbank	512-512	Recurrent	31.76
	Fbank+LIF	512-512	Recurrent	29.74
	Spiking-LEAF	512-512	Recurrent	**32.45**

Table 1: Comparison of different auditory front-ends on KWS and SI tasks. The bold color denotes the best model for each network configuration.

put signals. The neuronal dynamics of the resulting IHC-LIF model can be described as follows:

$$I_s[t] = \Sigma_i w_i S[t-1] + b \tag{11}$$

$$I_f[t] = ZeroDiag(W_f) * S[t-1] \tag{12}$$

$$I_{LI}[t] = ZeroDiag(W_{LI}) * S[t-1] \tag{13}$$

$$\begin{aligned} U_d[t] = U_d[t-1] + \beta_d * U_s[t-1] + I_s[t] \\ - \gamma * S[t-1] + I_f[t] \end{aligned} \tag{14}$$

$$\begin{aligned} U_s[t] = U_s[t-1] + \beta_s * U_d[t-1] - V_{th}S[t-1]) \\ - I_{LI}[t] \end{aligned} \tag{15}$$

$$S[t] = \mathbb{H}(U_s[t] - V_{th}) \tag{16}$$

To further enhance the encoding efficiency, we incorporate a spike rate regularization term L_{SR} into the loss function L. It has been applied alongside the classification loss L_{cls}: $L = L_{cls} + \lambda L_{SR}$ where $L_{SR} = ReLU(R - SR)$. Here, R represents the average spike rate per neuron per timestep and SR denotes the expected spike rate. Any spike rate higher than SR will incur a penalty, and λ is the penalty coefficient.

3 Experimental Results

In this section, we evaluate our model on keyword spotting (KWS) and speaker identification task. For keyword spotting, we use Google Speech Command Dataset V2 [31], which contains 105,829 one-second utterances of 35 commands. For speaker identification (SI), we use the Voxceleb1 dataset [32] with 153,516 utterances from 1,251 speakers, resulting in a classification task with 1,251 classes. We focus our evaluations on the auditory front-end by keeping model architecture and hyperparameters of the backend SNN classifier fixed.

3.1 Superior feature representation

Table 1 compares our proposed Spiking-LEAF model with other existing auditory front-ends on both KWS and SI tasks. Our results reveal that the Spiking-LEAF consistently outperforms the SOTA spike encoding methods as well as the fbank

Acoustic features	Neuron type	I_f	I_{LI}	L_{SR}	Firing rate	Accuracy
Fbank	LIF	-	-	-	17.94%	85.24%
Learnable	LIF	-	-	-	18.25%	90.73%
Learnable	TC-LIF	-	-	-	34.21%	91.89%
Learnable	TC-LIF	✓	-	-	40.35%	92.24%
Learnable	TC-LIF	✓	✓	-	34.54%	**92.43%**
Learnable	TC-LIF	✓	-	✓	15.03%	90.82%
Learnable	TC-LIF	✓	✓	✓	**11.96%**	92.04%

Table 2: Ablation studies of various components of the proposed Spiking-LEAF model on the KWS task.
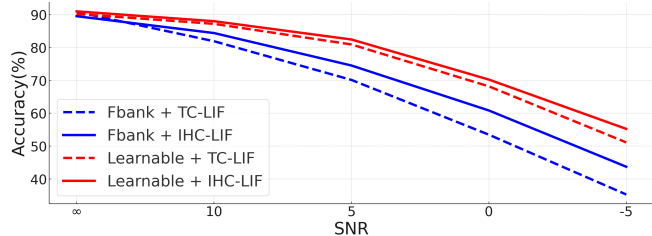

Fig. 3: Test accuracy on the KWS task with varying SNRs.

features [3], demonstrating a superior feature representation power. In the following section, we validate the effectiveness of key components of Spiking-LEAF: learnable acoustic feature extraction, two-compartment LIF (TC-LIF) neuron model, lateral feedback I_f, lateral inhibition I_{LI}, and firing rate regulation loss L_{SR}.

3.2 Ablation studies

Learnable filter bank and two-compartment neuron. As illustrated in row 1 and row 2 of Table 2, the proposed learnable filter bank achieves substantial enhancement in feature representation when compared to the widely adopted Fbank feature. Notably, further improvements in classification accuracy are observed (see row 3) when replacing LIF neurons with TC-LIF neurons that offer richer neuronal dynamics. However, it is important to acknowledge that this improvement comes at the expense of an elevated firing rate, which has a detrimental effect on the encoding efficiency.

Lateral feedback. Row 4 and row 5 of Table 2 highlight the potential of lateral feedback mechanisms in enhancing classification accuracy, which can be explained by the enhanced frequency sensitivity facilitated by the lateral feedback. Furthermore, the incorporation of lateral feedback is also anticipated to enhance the neuron's robustness in noisy environments. To substantiate this claim, our model is trained on clean samples and subsequently tested on noisy test samples contaminated with noise from the NOISEX-92 [33] and CHiME-3 [34] datasets. Fig. 3 illustrates the results of this evaluation, demonstrating that both the learnable filter bank and lateral feedback mechanisms contribute to enhanced noise robustness. This observation aligns with prior studies that have elucidated the role of the PCEN in fostering noise robustness [16]. Simultaneously, Fig. 4 showcases how the lateral feedback aids in filtering out unwanted spikes.

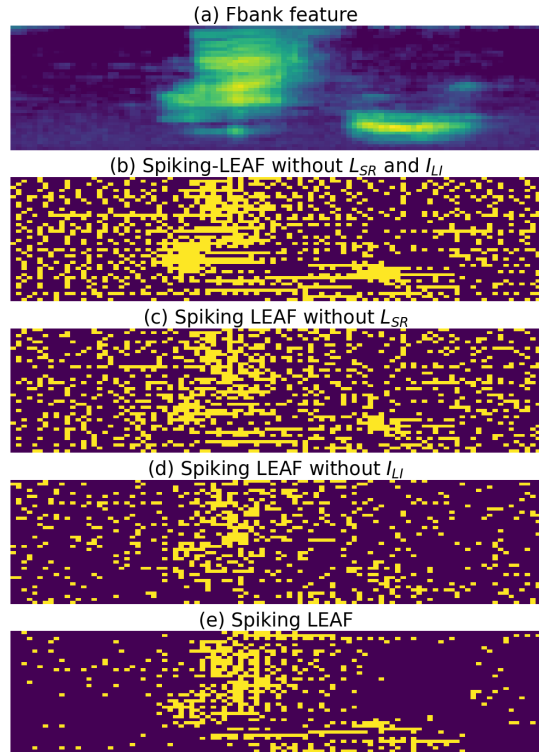**Lateral inhibition and spike rate regularization loss.**

Fig. 4: This figure illustrates the Fbank feature and spike representation generated by Spiking-LEAF without and with lateral inhibition and spike rate regularization loss.

As seen in Fig. 4 (b), when the spike regularization loss and lateral inhibition are not applied, the output spike representation involves a substantial amount of noise during non-speech periods. Introducing lateral inhibition or spike regularization loss alone can not suppress the noise that appeared during such periods (Figs. (b) and (c)). Particularly, introducing the spike regularization loss alone results in a uniform reduction in the output spikes (Fig. 4 (d)). However, this comes along with a notable reduction in accuracy as highlighted in Table 2 row 6. Notably, the combination of lateral inhibition and spike rate regularization (Fig. 4 (e)) can effectively suppress the unwanted spike during non-speech periods, yielding a sparse and yet informative spike representation.

4 Conclusion

In this paper, we presented a fully learnable audio front-end for SNN-based speech processing, dubbed Spiking-LEAF. The Spiking-LEAF integrated a learnable filter bank with a novel IHC-LIF neuron model to achieve effective feature extraction and neural encoding. Our experimental evaluation on KWS and SI tasks demonstrated enhanced feature representation power, noise robustness, and encoding efficiency over SOTA auditory front-ends. It, therefore, opens up a myriad of opportunities for ultra-low-power speech processing at the edge with neuromorphic solutions.

5 References

[1] Bojian Yin, Federico Corradi, and Sander M Bohté, "Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 905–913, 2021.

[2] Yuchen Wang, Kexin Shi, Chengzhuo Lu, Yuguo Liu, Malu Zhang, and Hong Qu, "Spatial-temporal self-attention for asynchronous spiking neural networks," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Edith Elkind, Ed. 8 2023, pp. 3085–3093, International Joint Conferences on Artificial Intelligence Organization, Main Track.

[3] Alexandre Bittar and Philip N Garner, "A surrogate gradient spiking baseline for speech command recognition," *Frontiers in Neuroscience*, vol. 16, pp. 865897, 2022.

[4] Jibin Wu, Emre Yılmaz, Malu Zhang, Haizhou Li, and Kay Chen Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," *Frontiers in neuroscience*, vol. 14, pp. 199, 2020.

[5] Jibin Wu, Yansong Chua, Malu Zhang, Haizhou Li, and Kay Chen Tan, "A spiking neural network framework for robust sound classification," *Frontiers in neuroscience*, vol. 12, pp. 836, 2018.

[6] Jibin Wu, Yansong Chua, and Haizhou Li, "A biologically plausible speech recognition framework based on spiking neural networks," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[7] Zihan Pan, Jibin Wu, Malu Zhang, Haizhou Li, and Yansong Chua, "Neural population coding for effective temporal classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[8] Malu Zhang, Xiaoling Luo, Yi Chen, Jibin Wu, Ammar Belatreche, Zihan Pan, Hong Qu, and Haizhou Li, "An efficient threshold-driven aggregate-label learning algorithm for multimodal information processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 592–602, 2020.

[9] Jibin Wu, Chenglin Xu, Xiao Han, Daquan Zhou, Malu Zhang, Haizhou Li, and Kay Chen Tan, "Progressive tandem learning for pattern recognition with deep spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7824–7840, 2021.

[10] Xingyi Yang, Jingwen Ye, and Xinchao Wang, "Factorizing Knowledge in Neural Networks," in *European Conference on Computer Vision*, 2022.

[11] Xinyin Ma, Gongfan Fang, and Xinchao Wang, "LLM-Pruner: On the Structural Pruning of Large Language Models," in *Advances in neural information processing systems*, 2023.

[12] Zihan Pan, Yansong Chua, Jibin Wu, Malu Zhang, Haizhou Li, and Eliathamby Ambikairajah, "An efficient and perceptually motivated auditory neural encoding and decoding algorithm for spiking neural networks," *Frontiers in neuroscience*, vol. 13, pp. 1420, 2020.

[13] Tara Sainath, Ron J Weiss, Kevin Wilson, Andrew W Senior, and Oriol Vinyals, "Learning the speech front-end with raw waveform cldnns," 2015.

[14] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4624–4628.

[15] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[16] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[17] Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke, "The heidelberg spiking data sets for the systematic evaluation of spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2744–2757, 2020.

[18] Robert Legenstein, Dejan Pecevski, and Wolfgang Maass, "A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback," *PLoS computational biology*, vol. 4, no. 10, pp. e1000180, 2008.

[19] Marek Miskowicz, "Send-on-delta concept: An event-based data reporting strategy," *sensors*, vol. 6, no. 1, pp. 49–63, 2006.

[20] Malu Zhang, Jibin Wu, Yansong Chua, Xiaoling Luo, Zihan Pan, Dan Liu, and Haizhou Li, "Mpd-al: an efficient membrane potential driven aggregate-label learning algorithm for spiking neurons," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 1327–1334.

[21] Benjamin Schrauwen and Jan Van Campenhout, "Bsa, a fast and accurate spike train encoding scheme," in *Proceedings of the International Joint Conference on Neural Networks, 2003*. IEEE, 2003, vol. 4, pp. 2825–2830.

[22] Mark Bear, Barry Connors, and Michael A Paradiso, *Neuroscience: exploring the brain, enhanced edition: exploring the brain*, Jones & Bartlett Learning, 2020.

[23] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F Lyon, and Rif A Saurous, "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.

[24] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux, "Learning filterbanks from raw speech for phone recognition," in *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2018, pp. 5509–5513.

[25] Wulfram Gerstner and Werner M Kistler, *Spiking neuron models: Single neurons, populations, plasticity*, Cambridge university press, 2002.

[26] Shimin Zhang, Qu Yang, Chenxiang Ma, Jibin Wu, Haizhou Li, and Kay Chen Tan, "Long short-term memory with two-compartment spiking neuron," *arXiv preprint arXiv:2307.07231*, 2023.

[27] AJ Hudspeth, "Integrating the active process of hair cells with cochlear function," *Nature Reviews Neuroscience*, vol. 15, no. 9, pp. 600–614, 2014.

[28] John J Guinan Jr, "Olivocochlear efferents: Their action, effects, measurement and uses, and the impact of the new conception of cochlear mechanical responses," *Hearing research*, vol. 362, pp. 38–47, 2018.

[29] Aritra Sasmal and Karl Grosh, "The competition between the noise and shear motion sensitivity of cochlear inner hair cell stereocilia," *Biophysical Journal*, vol. 114, no. 2, pp. 474–483, 2018.

[30] Kenneth Michael Stewart, Timothy Shea, Noah Pacik-Nelson, Eric Gallo, and Andreea Danielescu, "Speech2spikes: Efficient audio encoding pipeline for real-time neuromorphic systems," in *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*, 2023, pp. 71–78.

[31] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *ArXiv e-prints*, Apr. 2018.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

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

[34] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.