СРАВНЕНИЕ МЕТОДОВ КОДИРОВАНИЯ АУДИОСИГНАЛОВ ДЛЯ МНОГОСЛОЙНЫХ СПАЙКОВЫХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Р. Б. Рыбка[1,2,а], **Д. Г. Крикунов**[1,3,б], **Я. П. Ивина**[1,в]

[1] *Национальный исследовательский центр «Курчатовский институт», г. Москва, Российская Федерация*

[2] *Московский физико-технический институт (Национальный исследовательский университет), г. Москва, Российская Федерация*

[3] *Московский инженерно-физический институт (Национальный исследовательский ядерный университет), г. Москва, Российская Федерация*

[а] *ORCID: http://orcid.org/0000-0002-5595-6398,* ✍ *rybkarb@gmail.com*

[б] *ORCID: http://orcid.org/0009-0004-0770-8842, kdg006@campus.mephi.ru*

[в] *ORCID: http://orcid.org/0009-0008-4307-9376, yana.ivina.00@mail.ru*

Аннотация: рассматривается представление коротких речевых сигналов последовательностями спайков для классификации многослойной спайковой нейронной сетью в потоковом режиме. Сравниваются три метода кодирования: интегральное частотное кодирование, потоковое спектральное кодирование и кодирование локальной динамики сигнала — все с единой архитектурой сети из нейронов Leaky Integrate-and-Fire, обучаемой методом суррогатного градиента. Эксперименты на наборах Google Speech Commands v0.01 (GSC) и Heidelberg Digits (HD) показали: для GSC наивысшая точность (83,1 %) достигается при интегральном кодировании, для HD все методы дают сравнимые точности. Потоковое спектральное кодирование требует лишь 4 шагов симуляции на один шаг спектрограммы, что снижает задержку ответа сети. Кодирование локальной динамики дает на GSC точность 80,1 % (на 3 % ниже интегрального) при более компактном входном представлении. Результаты позволяют дать рекомендации по выбору спайкового кодирования под заданные ограничения на вычислительные ресурсы.

Ключевые слова: спайковые нейронные сети, кодирование аудиосигналов, распознавание речи, многослойный перцептрон, суррогатный градиент.

Благодарности: работа выполнена в рамках государственного задания НИЦ «Курчатовский институт». Работа выполнена с использованием вычислительных ресурсов федерального центра коллективного пользования «Комплекс моделирования и обработки данных для установок мега-класса» Национального исследовательского центра «Курчатовский институт», http://ckp.nrcki.ru/.

Для цитирования: Рыбка Р. Б., Крикунов Д. Г., Ивина Я. П. Сравнение методов кодирования аудиосигналов для многослойных спайковых нейронных сетей в задачах классификации. *Успехи кибернетики.* 2026;7(3):126–132.

Поступила в редакцию: 22.07.2026. *В окончательном варианте: 12.08.2026.*

COMPARISON OF AUDIO SIGNAL ENCODING METHODS FOR CLASSIFICATION WITH MULTILAYER SPIKING NEURAL NETWORKS

R. B. Rybka[1,2,a], **D. G. Krikunov**[1,3,b], **Ya. P. Ivina**[1,c]

[1] *National Research Centre "Kurchatov Institute", Moscow, Russian Federation*

[2] *Moscow Institute of Physics and Technology (National Research University), Moscow, Russian Federation*

[3] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation*

[a] *ORCID: http://orcid.org/0000-0002-5595-6398,* ✍ *rybkarb@gmail.com*

[b] *ORCID: http://orcid.org/0009-0004-0770-8842, kdg006@campus.mephi.ru*

[c] *ORCID: http://orcid.org/0009-0008-4307-9376, yana.ivina.00@mail.ru*

Abstract: we compared three computationally simple methods for encoding short speech recordings as spike sequences for real-time classification with a multilayer spiking neural network. The methods included integral coding of spectral features, stream spectral encoding, and direct encoding of

local signal dynamics. All three methods used the same network of Leaky Integrate-and-Fire neurons trained by surrogate-gradient backpropagation. Experiments on the Google Speech Commands v0.01 (GSC) and Heidelberg Digits (HD) datasets showed that integral rate coding achieved the highest accuracy on GSC (83.1%), whereas all three methods performed comparably on HD. Stream spectral encoding required only four simulation time steps per spectrogram frame, substantially reducing response latency. Speech2Spikes achieved 80.1% accuracy on GSC, only 3% below integral rate coding, while using a more compact input representation. The results provide a basis for selecting spike encoding methods under different computational and latency constraints.

Keywords: spiking neural networks, audio signal encoding, speech recognition, multilayer perceptron, surrogate gradient.

Acknowledgements: this study is a part of a government contract awarded to the National Research Centre “Kurchatov Institute”. The authors utilized the computing resources of the Simulation and Data Processing for Megascience shared research facility at the “Kurchatov Institute”, http://ckp.nrcki.ru/.

Cite this article: Rybka R. B., Krikunov D. G., Ivina Ya. P. Comparison of Audio Signal Encoding Methods for Classification with Multilayer Spiking Neural Networks. *Russian Journal of Cybernetics.* 2026;7(3):126–132.

Original article submitted: 22.07.2026. *Revision submitted: 12.08.2026.*

Введение

Импульсные (спайковые) нейронные сети перспективны для энергоэффективного искусственного интеллекта благодаря событийной обработке и совместимости с нейроморфными устройствами [1–6]. Для распознавания речевых команд в реальном времени важны малая задержка и компактность модели.

Частотное кодирование отображает амплитуду признака спектрограммы в среднюю частоту спайков. Оно обеспечивает высокую точность [7–10], но требует обработки всей спектрограммы. Потоковое спектральное кодирование работает с текущим кадром, а кодирование локальной динамики передает направление изменения признака относительно порога [5, 11]. Частотное кодирование имеет три ограничения: для представления признака без потери информации нужно несколько временных шагов (что увеличивает задержку ответа), размерность входа растет с числом временных отсчетов спектрограммы, а само формирование спектрограммы требует предварительного накопления фрагмента сигнала. Потоковые методы кодирования частично компенсируют сокращение временного контекста за счёт внутренней динамики сети — нелинейности мембранного потенциала, утечки и порогового механизма.

Цель работы — сравнить эти методы при фиксированной архитектуре сети Leaky Integrate-and-Fire, обучаемой суррогатным градиентом [1, 8].

Основные результаты:

- на GSC интегральное кодирование даёт 83,1 %, а на HD — 77,1 %;
- на HD точности методов близки: 78,5 %, 77,1 % и 76,5 %;
- на GSC динамическое кодирование превосходит потоковое: 80,1 % против 70,9 %;
- даны рекомендации по выбору метода с учётом точности, задержки и ресурсов.

Наборы данных

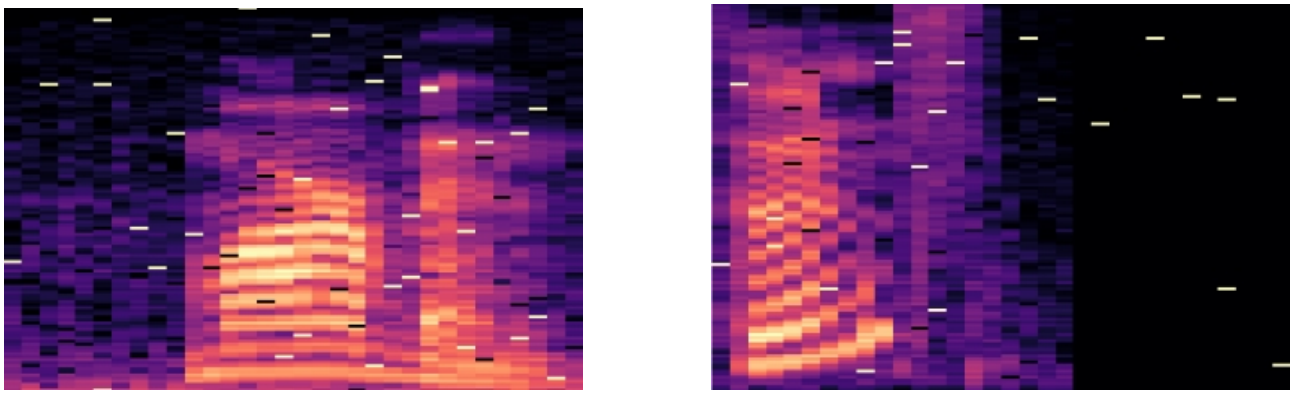
Используются Google Speech Commands (GSC) и Heidelberg Digits (HD); основные характеристики приведены в таблице 1. GSC содержит 30 голосовых команд от 1881 диктора [7], HD — 20 классов цифр на английском и немецком языках от 12 дикторов [4].

Таблица 1

Сводные характеристики используемых наборов данных

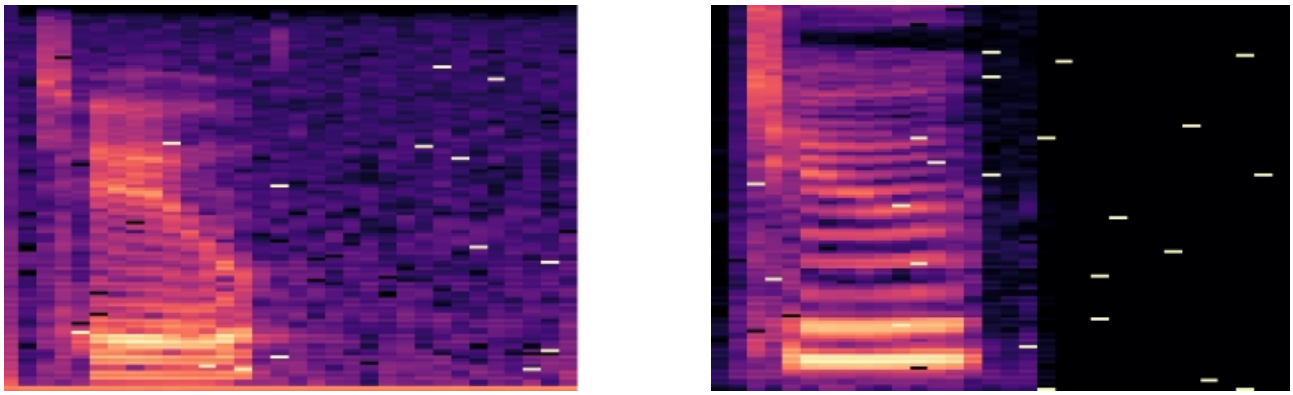
Характеристики	Google Speech Commands	Heidelberg Digits
Длина аудио: мин/макс/ср., с	0.37/1.00/0.98	0.23/1.37/0.72
Классов	30	20
Аудиозаписей в обучающей выборке	≈ 51000	≈ 8300
Аудиозаписей в валидационной выборке	≈ 6800	–
Аудиозаписей в тестировочной выборке	≈ 6800	≈ 2100

Примеры мел-спектрограмм представлены на рисунках 1 и 2.



GSC: команда **marvin** *HD: цифра 3 (нем.)*

Рис. 1. *Примеры мел-спектрограмм различных классов из наборов данных GSC и HD*



GSC: цифра 2 (англ.) *HD: цифра 2 (англ.)*

Рис. 2. *Мел-спектрограммы записей произнесения английской цифры 2 в наборах данных GSC (слева) и HD (справа)*

Все сигналы приводятся к длительности 1 с и частоте дискретизации 16 кГц: короткие записи дополняются нулями, длинные обрезаются. На их основе вычисляются мел-спектрограммы с $N = 128$ частотными каналами:

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

Спектрограмма имеет размерность $N \times T_{\text{frames}}$, где T_{frames} определяется шагом окна. Значения признаков нормируются по всему набору данных:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{2}$$

Методы кодирования в спайковые последовательности

Интегральное кодирование (rate coding)

Нормированное значение признака x_{norm} задает вероятность спайка на каждом из T временных шагов [3]:

$$s_t \sim \text{Bernoulli}(p(x_{\text{norm}})), \quad t = 1, \ldots, T. \tag{3}$$

Средняя частота спайков кодирует амплитуду признака, а размер входа равен $N_{\text{in}} = N \times T_{\text{frames}}$.

Потоковое спектральное кодирование (frequency stream coding)

Метод обрабатывает кадры мел-спектрограммы последовательно [5]. Каждый кадр разбивается на $K = 4$ подшага, а нормированная амплитуда определяет число равномерно распределённых спайков m_{k,t_f}. Поэтому $T = T_{\text{frames}} \times K$, а $N_{\text{in}} = N$.

Кодирование локальной динамики сигнала (Speech2Spikes)

Метод кодирует увеличение и уменьшение признака относительно порога θ [11]. Для каждого признака формируются положительный и отрицательный потоки:

$$s^{+}_{i,t} = \mathbb{I}\left(x_{i,t} - x_{i,t-1} > \theta\right), \tag{4}$$

$$s^{-}_{i,t} = \mathbb{I}\left(x_{i,t-1} - x_{i,t} > \theta\right), \tag{5}$$

где θ — порог чувствительности, а $\mathbb{I}(\cdot)$ — индикаторная функция. Первый поток соответствует увеличению признака, второй — его уменьшению.

Размер входа равен $N_{\text{in}} = 2N$. Для сопоставимости с потоковым методом используется шаг спектрограммы hop/K, где $K = 4$.

Параметры приведены в таблице 2.

Таблица 2

Параметры извлечения мел-спектрограммы и размерность входа спайковой сети при различных методах кодирования

Метод	Шаг сдвига, точек	Шаг сдвига, мс	N_{in}	Длина окна T_{frames}	Длит. симуляции T
Интегральное кодирование	512	32.0	4096	1	10
Потоковое спектральное кодирование	512	32.0	128	32	4
Кодирование локальной динамики сигнала	122	7.6	256	128	32

Для всех методов используется частота дискретизации 16000 Гц, размер окна БПФ 400 и число частотных каналов $N = 128$.

Архитектура спайковой нейронной сети

Используется полносвязная многослойная сеть MLP-LIF, состоящая из входного слоя, двух скрытых слоев и выходного классификатора (рисунок 3). Размерность входа N_{in} приведена в таблице 2.

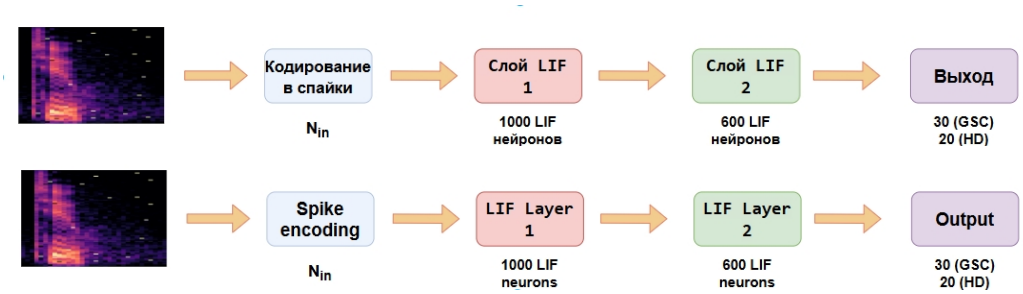


Рис. 3. *Схема архитектуры MLP-LIF: спектрограмма, кодирование, два скрытых слоя LIF и выходной слой классификатора*

Скрытые слои содержат 1000 и 600 нейронов LIF. Динамика мембранного потенциала описывается уравнениями (6) и (7).

$$v_{t+1} = \alpha v_t + W x_t - v_{\text{th}} s_t, \tag{6}$$

где v_t — мембранный потенциал в момент времени t, x_t — вектор входных спайков, W — матрица весов, $\alpha \in (0{,}1)$ — коэффициент утечки, v_{th} — порог генерации спайка, $s_t \in \{0{,}1\}$ — бинарный выход нейрона. Спайк генерируется при превышении порога:

$$s_t = H(v_t - v_{\text{th}}), \tag{7}$$

где $H(\cdot)$ — функция Хевисайда.

Выходной слой содержит 30 нейронов для GSC и 20 для HD. Спайки суммируются по времени, после чего применяются линейное преобразование и *softmax*. Для всех методов изменяется только размер входного слоя. Архитектура MLP-LIF намеренно выбрана простой: в ней отсутствуют рекуррентные связи и специальные блоки памяти, а вся временная информация обрабатывается за счет внутренней динамики мембранного потенциала LIF-нейронов. Такой выбор упрощает интерпретацию результатов: изменение точности при переходе от одного метода кодирования к другому можно напрямую связывать с различиями во входном представлении, поскольку механизмы обработки внутри сети остаются неизменными.

Обучение и суррогатная функция активации

Для обучения используется адаптация метода обратного распространения ошибки на основе суррогатных градиентов: в прямом проходе применяется ступенчатая функция (7), а в обратном проходе ее производная заменяется гладкой аппроксимацией в окрестности порога.

В качестве суррогатной функции для мембранного потенциала $\sigma(U)$ используется трёхлепестковая функция (Multi-Gaussian), предложенная в работе [12]:

$$\frac{\partial\sigma}{\partial v} \approx (1-h)\,g(v-v_{\text{th}}) - h\,g(v-v_{\text{th}}-l_s) - h\,g(v-v_{\text{th}}+l_s), \tag{8}$$

где $g(x) = \exp\bigl(-x^2/(2s^2)\bigr)$ — нормированная гауссова функция, v_{th} — порог срабатывания нейрона, а параметры $h \in (0{,}1)$, $l_s > 0$ и $s > 0$ задают параметры боковых лепестков. В численных экспериментах параметры суррогатного градиента фиксировались: $h = 0.15$, $l_s = 1.0$, $s = 6.0$.

Суррогатная функция передает градиент вблизи порога и ограничивает его норму при BPTT. В процессе обучения нейросетевой модели настраиваются веса сети и коэффициент утечки нейронов скрытых слоев α .

Эксперименты и результаты

Сравниваются три метода кодирования при фиксированной архитектуре MLP-LIF и одинаковых гиперпараметрах обучения (см. таблицу 3). Каждый эксперимент повторяется четыре раза с различными начальными весами. Гиперпараметры подбирались по валидационной выборке набора GSC (как более сложного среди выбранных датасетов) и далее не изменялись при смене метода кодирования. Тестовые точности приведены в таблице 4.

Таблица 3

Гиперпараметры обучения модели MLP-LIF

Параметр	**Значение**
Оптимизатор	Adam
Функция потерь	NLLLoss
Размер батча	256
Число эпох	50
Длина временного окна для rate coding	10 шагов
Начальная скорость обучения	$2 \cdot 10^{-3}$
Порог генерации спайка v_{th}	1.0
Инициализация весов	Xavier uniform

Таблица 4

Точность и характеристики итоговой модели MLP-LIF при различных методах кодирования

Метод	**Пар.**	$N_{\text{in}} \times T$	**Режим**	**GSC, %**	**HD, %**
Интегральное кодирование	≈4.7M	4096×10	офлайн	$\mathbf{83.1 \pm 1.2}$	77.1 ± 2.5
Кодирование локальной динамики	≈0.9M	256×128	онлайн	80.1 ± 1.7	76.5 ± 1.3
Потоковое частотное кодирование	≈0.7M	128×128	онлайн	70.9 ± 1.5	$\mathbf{78.5 \pm 2.1}$

На HD методы дают близкие результаты (76,5–78,5 %). На GSC интегральный метод дает 83,1 %, динамический — 80,1 %, а потоковый — 70,9 %. Для базовых нерекуррентных спайковых

сетей без обучаемых задержек на близком по составу наборе Spiking Heidelberg Digits (SHD) [4] сообщаются точности порядка 75–80 %, тогда как добавление адаптивной динамики и совместно обучаемых синаптических задержек позволяет повысить точность до ≈ 90 % и выше [12, 13]. Таким образом, наблюдаемое ограничение точности на HD связано не только со способом кодирования, но и с простотой архитектуры MLP-LIF.

Обсуждение результатов

Выбор кодирования определяется компромиссом между точностью, размером модели и задержкой. Интегральное кодирование использует вектор $N_{\text{in}} = 4096$ и дает наивысшую точность на GSC, но требует около 4,7 млн параметров и предварительного накопления записи. В отличие от интегрального кодирования, Speech2Spikes передает не абсолютное значение спектрального признака в кадре, а лишь величину его изменения относительно предыдущего кадра. Это делает метод чувствительным к разрешению спектрограммы по времени и частоте: при крупном шаге дискретизации локальные изменения сигнала между соседними кадрами могут оказаться меньше порога θ и не порождать спайков, что приводит к потере информации о сигнале. Для повышения точности целесообразно использовать более плотные спектральные признаки.

Потоковое кодирование использует $N_{\text{in}} = 128$ и около 0,7 млн параметров, обеспечивая минимальную задержку. Кодирование локальной динамики требует $N_{\text{in}} = 256$ и около 0,9 млн параметров; на GSC оно уступает интегральному методу примерно на 3 процентных пункта. Интегральное кодирование предпочтительно при возможности накопления фрагментов длительностью более 1 с и поддержке входного слоя размером ≈4 тыс. нейронов. При более жестких ограничениях (компактность модели, снижение задержек и затрат на память) предпочтительны потоковые методы.

Заключение

Сравнены три метода кодирования аудиосигналов в сети MLP-LIF на наборах GSC и HD. На GSC интегральное кодирование обеспечивает 83,1 %, а динамическое — 80,1 % при сокращении числа параметров примерно с 4,7 до 0,9 млн. На HD точности методов составляют 76,5–78,5 %.

Дальнейшая работа включает совместную настройку кодирования и архитектуры сети, в том числе сочетание потоковых методов с рекуррентными связями, адаптивной динамикой и обучаемыми задержками [2, 8, 12, 13].

ЛИТЕРАТУРА

1. Neftci E. O., Mostafa H., Zenke F. Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks. *IEEE Signal Processing Magazine*. 2019;36(6):51–63.
2. Bellec G., Salaj D., Subramoney A., Legenstein R., Maass W. Spike-Frequency Adaptation Provides a Long Short-Term Memory to Networks of Spiking Neurons. *Nature Communications*. 2020;11:3690.
3. Auge D., Hille J., Mueller E., Knoll A. A Survey of Encoding Techniques for Signal Processing in Spiking Neural Networks. *Neural Networks*. 2021;142:273–301.
4. Cramer B., Stradmann Y., Schemmel J., Zenke F. The Heidelberg Spiking Data Sets for the Systematic Evaluation of Spiking Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2022;33(7):2744–2757.
5. Rybka R. B., Dyakova E. O., Serenko A. V., Sboev A. G. Spiking Neural Network with Quantized Weights for Audio Stream Classification. *Studies in Computational Intelligence*. 2025;1179:89–97.
6. Matsukatova A. N., Prudnikov N. V., Kulagin V. A. et al. Combination of Organic-Based Reservoir Computing and Spiking Neuromorphic Systems for a Robust and Efficient Pattern Classification. *Adv. Intell. Syst.* 2023;5:2200407.
7. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. 2018. arXiv:1804.03209.
8. Bittar A., Garner P. N. A Surrogate Gradient Spiking Baseline for Speech Command Recognition. *Frontiers in Neuroscience*. 2022;16:865897.

9. Wang J., Yu L., Huang L., Zhou C., Zhang H., Song Z., Zhang M., Ma Z., Zhang Z. Efficient Speech Command Recognition Leveraging Spiking Neural Network and Curriculum Learning-Based Knowledge Distillation. 2024. arXiv:2412.12858.
10. Zhang S., Yang Q., Ma C., Wu J., Li H., Tan K. C. TC-LIF: A Two-Compartment Spiking Neuron Model for Long-Term Sequential Modelling. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.
11. Stewart K. M., Shea T., Pacik-Nelson N., Gallo E., Danielescu A. Speech2Spikes: Efficient Audio Encoding Pipeline for Real-Time Neuromorphic Systems. *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference (NICE)*. 2023:1–7. DOI: 10.1145/3584954.3584995.
12. Yin B., Corradi F., Bohté S. M. Accurate and Efficient Time-Domain Classification With Adaptive Spiking Recurrent Neural Networks. *Nature Machine Intelligence*. 2021;3:905–913.
13. Hammouamri I., Khalfaoui-Hassani I., Masquelier T. Learning Delays in Spiking Neural Networks Using Dilated Convolutions with Learnable Spacings. *The Twelfth International Conference on Learning Representations (ICLR)*. 2024.