High-Accuracy and Energy-Efficient Acoustic Inference using Hardware-Aware Training and a 0.34nW/Ch Full-Wave Rectifier

Zhou, Sheng ; Chen, Xi ; Kim, Kwantae ; Liu, Shih-Chii

DOI: https://doi.org/10.1109/aicas57966.2023.10168561

This paper has been accepted for presentation at the 2023 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS).

High-Accuracy and Energy-Efficient Acoustic Inference using Hardware-Aware Training and a 0.34 nW/Ch Full-Wave Rectifier

Sheng Zhou*, Xi Chen*, Kwantae Kim, Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
{shengzhou, xichennn, kwantae, shih}@ini.uzh.ch

***Abstract*—A full-wave rectifier (FWR) is a necessary component of many analog acoustic feature extractor (FEx) designs targeted at edge audio applications. However, analog circuits that perform close-to-ideal rectification contribute a significant portion of the total power of the FEx. This work presents an energy-efficient FWR design by using a dynamic comparator and scaling the comparator clock frequency with its input signal bandwidth. Simulated in a 65 nm CMOS process, the rectifier circuit consumes 0.34 nW per channel for a 0.6 V supply. Although the FWR does not perform ideal rectification, an acoustic FEx behavioral model in Python is proposed based on our FWR design, and a neural network trained with the output of the proposed behavioral model recovers high classification accuracy in an audio keyword spotting (KWS) task. The behavioral model also included comparator noise and offset extracted from transistor-level simulation. The whole KWS chain using our behavioral model achieves 89.45% accuracy for 12-class KWS on the Google Speech Commands Dataset.**

I. Introduction

Rapid market growth of Internet-of-Things (IoT) devices has led to an increasing demand for low-power voice interface, which provides an intuitive and hands-free way of interaction for the users. To achieve always-on operation for edge audio devices with a restricted power budget, keyword spotting (KWS) is an indispensable functionality that power-gates more complex and energy-consuming speech processing.

A typical KWS design consists of two components: 1) a feature extractor (FEx) that extracts acoustic features from the output of a microphone, and 2) a deep neural network (DNN)-based classifier that produces classification outputs using the extracted features. The FEx is of crucial importance because it not only affects the KWS accuracy, but also contributes a significant portion of the system power. The digital Mel-Frequency Cepstrum Coefficient (MFCC) FEx of KWS systems can consume more than 50% of the total power budget [1], [2]. As such, several papers have reported designs that improve the energy efficiency of the FEx using analog signal processing approaches [3]–[7].

A typical processing chain of an analog acoustic FEx is shown in Fig. 1. A low noise amplifier (LNA) amplifies the microphone signal and drives a bank of band-pass filters (BPFs) with different central frequencies. The time-varying amplitude information is then extracted from each BPF channel by an amplitude extractor, which is a cascade of a rectifier, a low-pass filter (LPF) and an analog-to-digital converter (ADC). Some FEx designs use an integrate-and-fire (IAF) neuron and a spike counter in place of the LPF and the ADC [3], [4].

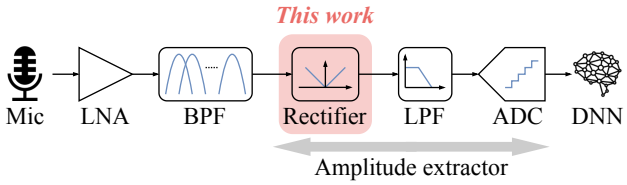


Fig. 1: Building blocks of an analog acoustic feature extractor. The amplitude extractor is a block within each channel.

The rectifier block of a previous state-of-the-art analog FEx design accounted for more than 50% of the total FEx power [3], because an operational transconductance amplifier (OTA) is required for continuous-time (CT) current-mode rectification. [4] reduced the rectifier power by using a single nFET operating in subthreshold as a nonlinear half-wave rectifier (HWR). However, it is sensitive to process, voltage and temperature (PVT) variation, leading to classification accuracy drop if the variation is not compensated for. [5] used a dynamic-comparator-based amplitude extractor [8], but it requires large capacitors to reduce sampling-induced kT/C noise, increasing both chip power and area. [6] proposed a passive-mixer-based amplitude extractor, but I/Q demodulation was not implemented which is essential to recover the

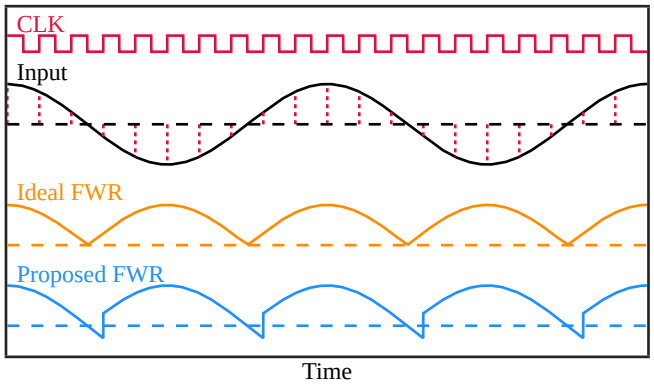


Fig. 2: Nonideal rectification output of the proposed FWR using a dynamic comparator.

This work was partially funded by the Swiss National Science Foundation projects CA-DNNEdge (208227) and SCIDVS (185069).
*Sheng Zhou and Xi Chen are equal contributors to this work.

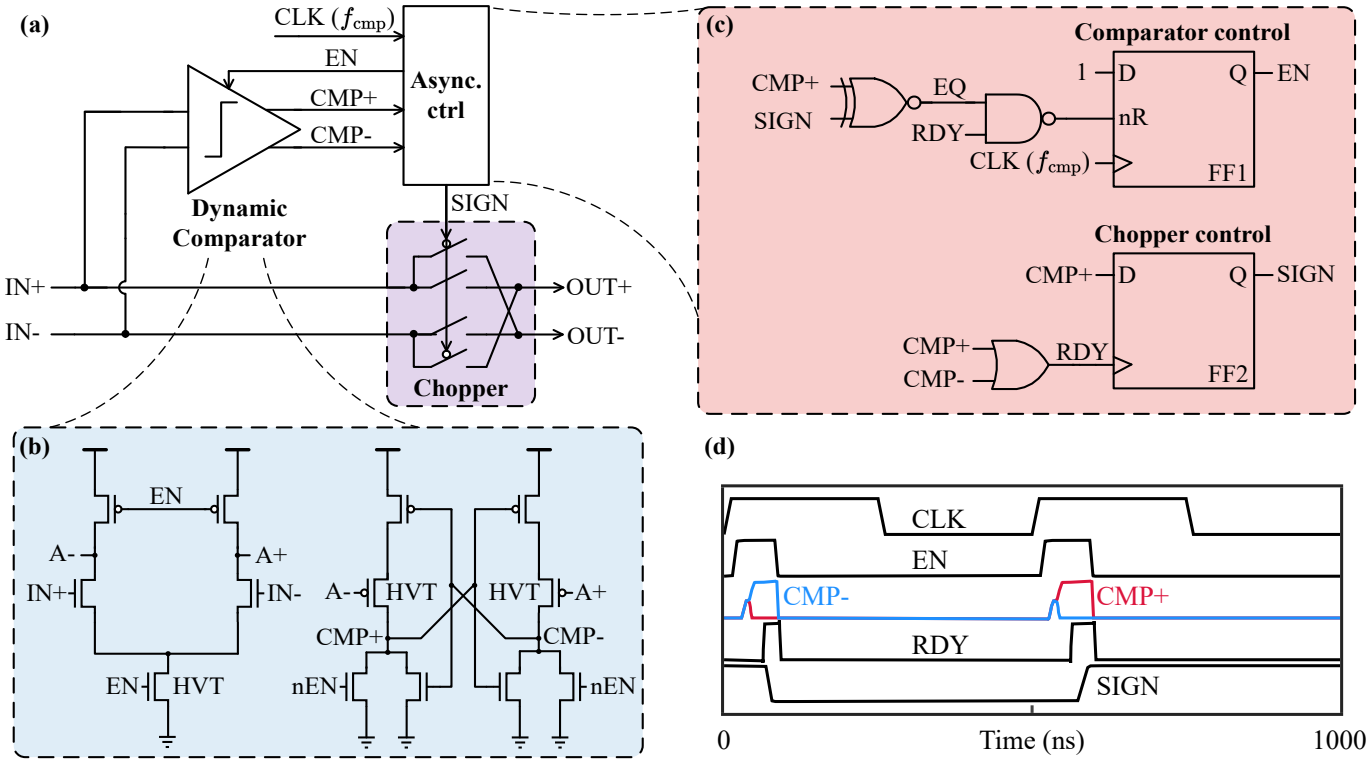


Fig. 3: Proposed FWR design. (a) Block diagram of the FWR. (b) Schematic of the two-stage dynamic comparator. (c) Schematic of the asynchronous controller. (d) Transistor-level transient simulation of the control signals. A short clock period (500 ns) is used for better visualization.

amplitude information.

This work presents an energy-efficient FWR for an analog acoustic FEx using a dynamic comparator. The proposed FWR operates on CT signals, thereby avoiding sampling-induced kT/C noise. Although its rectification output is not ideal due to input zero-crossings between adjacent comparisons as shown in Fig. 2, we propose a Python-based behavioral model incorporating such nonidealities and demonstrate that high classification accuracy can be recovered in a 12-class KWS task on Google Speech Commands Dataset (GSCD) [9].

The rest of this paper is organized as follows. Sec. II details the proposed low-power FWR circuit design. Sec. III describes the proposed behavioral model and the setup of the KWS experiments. The experimental results are presented in Sec. IV and the conclusion is provided in Sec. V.

II. Low-power Full-Wave Rectifier Circuit Design

The proposed FWR (Fig. 3 (a)) contains a dynamic comparator, a chopper and an asynchronous controller. Both input and output to the FWR are differential. The asynchronous controller triggers the dynamic comparator based on an external clock CLK with frequency f_{cmp} to determine the input polarity, and controls the chopper based on the comparator output such that the input is passed through for one polarity and inverted for the other.

The comparator (Fig. 3 (b)) is an energy-efficient two-stage dynamic comparator [10]. Each current path contains at least one high-V_t device to reduce leakage power [11]. Other devices have low V_t for sufficient comparison speed. The comparator outputs, CMP+ and CMP-, are reset to 0 when EN = 0. After a comparison is made by setting EN = 1, either CMP+ or CMP- becomes 1.

The asynchronous controller (Fig. 3 (c)) is triggered by the external clock CLK and provides the internal control signals. The controller contains two flip-flops, one for the comparator (FF1) and the other for the chopper (FF2). The timing of the control signals from transistor-level simulation is shown in Fig. 3 (d), and the operation is explained as follows. At the rising edge of CLK, EN is set to 1 by FF1 to activate the dynamic comparator. After the comparison is finished (RDY = 1), FF2 stores CMP+ as SIGN. The chopper, implemented as 4 transmission gates, passes the input unchanged when SIGN = 1 and inverts the input otherwise. When the comparison is done (RDY = 1) and FF2 has settled (CMP+ = SIGN), FF1 resets the comparator by EN = 0. The controller and the comparator will stay idle until the next rising edge of CLK.

Since the power consumption of the FWR design is dynamic-only, it can be further reduced by scaling f_{cmp} with the central frequency of the BPF preceding it. We employ a simple f_{cmp} scaling scheme by dividing the maximum FWR clock frequency, f_{max}, by a power of two using a cascade of 1/2 frequency dividers. Each 1/2 frequency division stage can be implemented by a flip-flop with its input connected to its negated state output. The frequency division factors f_{max}/f_{cmp} for different channels are described in Table I.

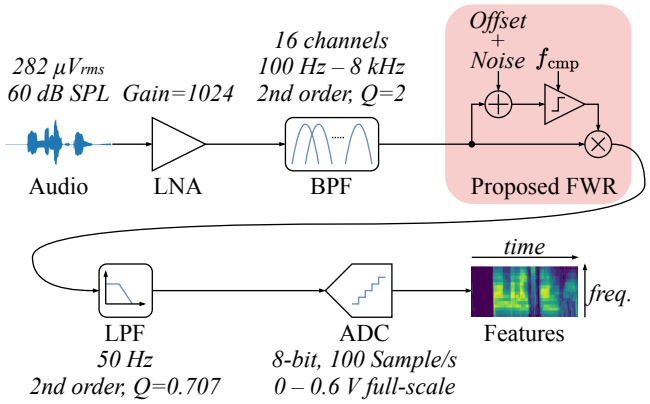


Fig. 4: Behavioral model of the FEx used for KWS task.

III. Behavioral Model and Experimental Setup

The FWR determines the input polarity only at the rising edges of the external clock CLK. Therefore, if the input changes polarity right after a clock rising edge, the chopper cannot be updated until the next rising edge comes. Our FWR produces a nonideal rectification output as illustrated in Fig. 2. The nonideality can be reduced by increasing f_{cmp} but at the cost of higher FWR power. Instead, we show in this work that its impact on KWS accuracy can also be mitigated by *rectifier-aware training*, i.e. training the DNN using the features generated by a behavioral model of a FEx that incorporates the FWR nonideality. This section describes the proposed behavioral model of the FWR (Sec. III-A), the FEx (Sec. III-B), and the KWS experiments setup (Sec. III-C).

A. Behavioral model of the proposed full-wave rectifier

The behavioral model of the FWR is illustrated as the red box in Fig. 4. The dynamic comparator of the FWR runs at a channel-dependent frequency f_{cmp} (Table I). For behavioral simulation, the input audio signal is sampled at f_{sim}, and therefore, there are $S = f_{\text{sim}}/f_{\text{cmp}}$ samples between two adjacent comparisons. For every S input samples, the model decides on the input polarity using the first sample and changes the output polarity for all S samples accordingly as implemented by the chopper in Fig. 3 (a). S is set sufficiently high for accurate modeling of the FWR (see Sec. III-B).

The FWR behavioral model also adds the input-referred noise and random offset to the comparator input, using statistics extracted from transistor-level simulation (Sec. IV-A), similar to [3]. The noise is sampled independently for each comparison, while the offset is sampled only once for each input audio sample.

B. Behavioral model of the feature extractor

The behavioral model of the analog acoustic FEx, shown in Fig. 4, contains 16 second-order BPF channels with Q = 2 and central frequencies geometrically scaled from 100 Hz to 8 kHz. The FWRs are as described in Sec. III-A. The second-order Butterworth LPFs have 50 Hz cutoff frequency. The ADCs have 8-bit resolution, 0 to 0.6 V full-scale and a 100 Hz sampling rate. $S \geqslant 6$ (Sec. III-A) is used so that the FEx behavioral model generates features with less than 1 LSB RMS error on GSCD. The entire analog signal processing chain has a gain of 1024 lumped in the LNA, similar to the gain implemented in [4].

TABLE I: FWR frequency division factors $f_{\text{max}}/f_{\text{cmp}}$ and BPF central frequencies f_c of different channels. See also Fig. 4.

Channel	1	2-4	5-6	7-8	9-11	12-16
$f_{\text{max}}/f_{\text{cmp}}$	32	16	8	4	2	1
f_c (kHz)	0.1	0.13-0.24	0.32-0.43	0.58-0.77	1.04-1.86	2.49-8

C. Setup of keyword spotting experiments

A network using the output of the FEx behavioral model is trained and evaluated on a 12-class KWS task[1] using GSCD [9]. Each recording is normalized to have 282 μV_{rms} amplitude, which corresponds to normal conversational speech at 60 dB sound pressure level (SPL) using a low-power microphone [12]. The audio is upsampled to $f_{\text{sim}} = 480$ kHz, so that $S \geqslant 6$ for all experimented f_{cmp}. Note that the upsampling is not done to increase the spectral information but rather to have an accurate simulation of the FWR.

For all experiments, we use a two-layer gated recurrent unit (GRU) network with 48 hidden units per layer and 24.2 k parameters in total. The input features to the network (i.e. ADC output codes) are first compressed logarithmically and then normalized using per-channel mean and standard deviation calculated from the entire GSCD training set. The cross-entropy loss is used, and a dropout probability of p = 0.5 is applied after each GRU layer. Other training parameters include the use of the AdamW optimizer [13], an initial learning rate of 1e-3, a weight decay of 1e-3 and a batch size of 64. Each GRU model is trained for 50 epochs with the cosine annealing learning rate [14].

IV. Experimental Results

A. Full-wave rectifier circuit simulations

We first determine how the FWR power scales with f_{cmp} by simulating the transistor-level circuit using a 0.6 V supply and a sinusoidal input with 289 mV_{rms} differential amplitude[2]. The FWR power is 2.5 nW when $f_{\text{cmp}} = 80$ kHz, and it scales almost linearly with f_{cmp} (Fig. 5a). The dynamic comparator and the asynchronous controller contribute respectively to 56% and 44% of the total FWR power. The comparator has an input-referred noise of 150 μV_{rms}. Monte-Carlo simulation shows that the input-referred offset follows a normal distribution with 7.52 mV standard deviation (Fig. 5b).

B. Keyword spotting behavioral simulations

The 12-class KWS accuracy of the network on GSCD using input features generated by the proposed behavioral model is shown in Fig. 6. The reported results are the average of 15 runs with different random network initialization.

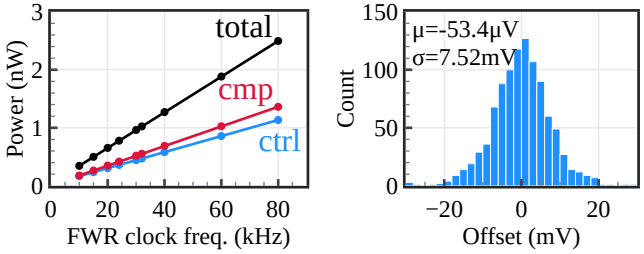
[1]The 12 classes are: "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", "Unknown", "Silence".

[2]Same amplitude as in the behavioral simulation after the LNA and BPF.

TABLE II: Comparison of rectifier designs for analog acoustic FEx.

	Shi, TCAS 2021 [5]	Yang, JSSC 2019 [3]	Yang, JSSC 2021 [4]	This work
Data obtained by	Measurement	Measurement	Measurement	Simulation
Process (nm)	180	180	65	65
Supply (V)	0.65	0.6	0.6	0.6
Rectifier power/Ch. (nW)	-	12.8[A]	0.15	0.34
Amp. extractor power/Ch. (nW)	7.6	13.6	0.79	0.71[B]
Amp. extractor building blocks	FWR, LPF, ADC	FWR, IAF, spike counter	HWR, IAF, spike counter	FWR, LPF, ADC
Task	VAD	VAD	2-class KWS[C]	12-class KWS
Accuracy	90%/86%[D]	84%/85%[D]	94.18%	89.45%[E]

[A] The same FWR design adapted to the 65 nm process we used consumes 10.6 nW in simulation.
[B] Including transistor-level simulated LPF power of 0.07 nW using flipped voltage follower [15], and estimated 8-bit 100 S/s ADC power of 0.3 nW from [16].
[C] Single keyword 'four'. [D] Speech/non-speech hit rate at 10 dB SNR. [E] Behavioral simulation result.



(a) (b)

Fig. 5: Circuit simulation results in a 65 nm CMOS process. (a) FWR power consumption versus f_{cmp}. (b) Distribution of dynamic comparator offset, 1000-run Monte-Carlo simulation.

A baseline accuracy of 90.85% is achieved when an ideal FWR model (Fig. 2) is used within each channel of the FEx behavioral model for both training and testing (red dashed line), close to the 91.35% accuracy reported in [7]. However, the accuracy of the baseline model drops when tested using features from a FEx behavioral model in which the proposed FWR is used in place of the ideal FWR model (orange curve). The behavioral model included the scaling of f_{cmp} (Table I), as well as the comparator noise and offset obtained from circuit simulations (Sec. IV-A). The test accuracy degradation is more significant as the maximum FWR clock frequency f_{max} decreases, because the rectification output of the proposed FWR becomes less similar to that of the ideal FWR (Fig. 2).

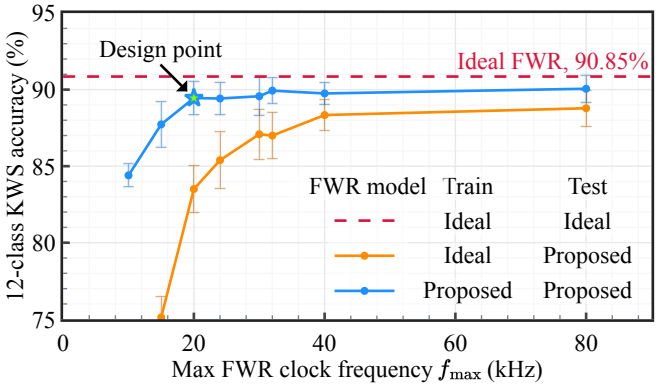


Fig. 6: 12-class KWS accuracy on GSCD using input features simulated by the proposed behavioral model, which includes the comparator noise and offset.

When rectifier-aware training (Sec. III) is applied, high KWS accuracy can be recovered (blue curve). For the chosen design point of $f_{max} = 20$ kHz, the KWS accuracy is 89.45%. Without rectifier-aware training, the model is not able to achieve the same accuracy even with $f_{max} = 80$ kHz. Transistor-level simulation shows that at the design point, the total FWR power for 16 channels including the clock divider is 5.43 nW. This corresponds to 0.34 nW FWR power per channel.

C. Comparison with the prior art

Table II compares the proposed rectifier design with other designs used in state-of-the-art analog FEx. The proposed FWR design reduces the rectifier power by $\times 31.2$ compared to the FWR design in [3] that performs close-to-ideal rectification. When the proposed FWR is combined with a low-power ADC [16] and a LPF based on the flipped voltage follower [15] to build an amplitude extractor, the estimated power per channel is only 0.71 nW, lower than the amplitude extractor designs in [3]–[5].

V. Conclusion

In this work, we present a low-power FWR design for analog acoustic FEx targeting always-on edge audio devices. It achieves high energy-efficiency by using a dynamic comparator and scaling the comparator frequency with the FWR input bandwidth. Simulated in a 65 nm CMOS process, the proposed FWR consumes 0.34 nW per channel. A Python-based behavioral model of the proposed FWR is developed, taking into account of the dynamic nature of the comparator as well as circuit-level nonidealities such as comparator noise and offset. We show that the behavioral model can help recover the KWS accuracy drop due to nonideal rectification. Using features generated by the proposed behavioral model, a two-layer GRU with 24.2 k parameters achieves 89.45% accuracy for 12-class KWS on GSCD. The hardware-software co-design solution presented in this work can help reduce the FWR power for analog acoustic FEx while retaining high KWS accuracy.

REFERENCES

[1] J. S. P. Giraldo, S. Lauwereins, K. Badami, and M. Verhelst, "Vocell: A 65-nm speech-triggered wake-up SoC for 10-μW keyword spotting and speaker verification," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 868–878, 2020.

[2] W. Shan, M. Yang, T. Wang, Y. Lu, H. Cai, L. Zhu, J. Xu, C. Wu, L. Shi, and J. Yang, "A 510-nW wake-up keyword-spotting chip using serial-FFT-based MFCC and binarized depthwise separable CNN in 28-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 151–164, 2021.

[3] M. Yang, C.-H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "Design of an always-on deep neural network-based 1 μW voice activity detector aided with a customized software model for analog feature extraction," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1764–1777, 2019.

[4] M. Yang, H. Liu, W. Shan, J. Zhang, I. Kiselev, S. J. Kim, C. Enz, and M. Seok, "Nanowatt acoustic inference sensing exploiting nonlinear analog feature extraction," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 10, pp. 3123–3133, 2021.

[5] E. Shi, X. Tang, and K. P. Pun, "A 270 nW switched-capacitor acoustic feature extractor for always-on voice activity detection," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 3, pp. 1045–1054, 2021.

[6] D. A. Villamizar, D. G. Muratore, J. B. Wieser, and B. Murmann, "An 800 nW switched-capacitor feature extraction filterbank for sound classification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1578–1588, 2021.

[7] K. Kim, C. Gao, R. Graça, I. Kiselev, H.-J. Yoo, T. Delbruck, and S.-C. Liu, "A 23 μW keyword spotting IC with ring-oscillator-based time-domain feature extraction," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 11, pp. 3298–3311, 2022.

[8] E. Shi, D. de Godoy, P. R. Kinget, and K.-P. Pun, "A 9.6 nW, 8-bit, 100 S/s envelope-to-digital converter for respiratory monitoring," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 3, pp. 445–449, 2020.

[9] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018. [Online]. Available: https://arxiv.org/abs/1804.03209

[10] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. A. M. Klumperink, and B. Nauta, "A 10-bit charge-redistribution ADC consuming 1.9 μW at 1 MS/s," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 5, pp. 1007–1015, 2010.

[11] P. J. A. Harpe, C. Zhou, Y. Bi, N. P. van der Meijs, X. Wang, K. Philips, G. Dolmans, and H. de Groot, "A 26 μW 8 bit 10 MS/s asynchronous SAR ADC for low energy radios," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1585–1595, 2011.

[12] InvenSense, "ICS-40310: Ultra-low current, low-noise microphone with analog output," 2014. [Online]. Available: https://invensense.tdk.com/wp-content/uploads/2015/02/ICS-40310-datasheet-v1.2.pdf

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[14] ——, "SGDR: stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[15] M. De Matteis and A. Baschirotto, "A biquadratic cell based on the flipped-source-follower circuit," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 8, pp. 867–871, 2017.

[16] P. Harpe, H. Gao, R. v. Dommele, E. Cantatore, and A. H. M. van Roermund, "A 0.20 mm^2 3 nW signal acquisition IC for miniature sensor nodes in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 240–248, 2016.