A 23-μW Keyword Spotting IC With Ring-Oscillator-Based Time-Domain Feature Extraction

Kim, Kwantae; Gao, Chang; Graca, Rui; Kiselev, Ilya; Yoo, Hoi Jun; Delbruck, Tobi; Liu, Shih Chii

DOI
10.1109/JSSC.2022.3195610

Publication date
2022

Document Version
Final published version

Published in
IEEE Journal of Solid-State Circuits

A 23-μW Keyword Spotting IC With Ring-Oscillator-Based Time-Domain Feature Extraction

Kwantae Kim, *Member, IEEE*, Chang Gao, *Member, IEEE*, Rui Graça, Ilya Kiselev, *Member, IEEE*, Hoi-Jun Yoo, *Fellow, IEEE*, Tobi Delbruck, *Fellow, IEEE*, and Shih-Chii Liu, *Fellow, IEEE*

***Abstract*— This article presents the first keyword spotting (KWS) IC that uses a ring-oscillator-based time-domain processing technique for its analog feature extractor (FEx). Its extensive usage of time-encoding schemes allows the analog audio signal to be processed in a fully time-domain manner except for the voltage-to-time conversion stage of the analog front end. Benefiting from fundamental building blocks based on digital logic gates, it offers better technology scalability compared to conventional voltage-domain designs. Fabricated in a 65-nm CMOS process, the prototyped KWS IC occupies 2.03 mm^2 and dissipates 23-μW power consumption, including analog FEx and digital neural network classifier. The 16-channel time-domain FEx achieves a 54.89-dB dynamic range for 16-ms frame shift size while consuming 9.3 μW. The measurement result verifies that the proposed IC performs a 12-class KWS task on the Google Speech Command dataset (GSCD) with >86% accuracy and 12.4-ms latency.**

***Index Terms*— Analog, bandpass filter (BPF), classifier, feature extractor (FEx), Google Speech Command dataset (GSCD), keyword spotting (KWS), rectifier, recurrent neural network (RNN), ring oscillator, time domain.**

I. Introduction

With incredible advances in artificial intelligence (AI) fields, there is an increasing demand for low-power audio Internet of Things (IoT) devices that process human speech on the device without data transmission to the cloud. These smart devices are required to ensure always-on operation, real-time response, small form factor, and longer battery lifetime. As such, an ultra-low-power wake-up functionality is being highlighted with rapidly growing popularity because it allows hierarchical power gating of increasingly complex tasks for audio IoT nodes. The keyword spotting (KWS) and voice activity detection (VAD) are widely used user-interactive methods to wake up smart devices. The KWS is used to detect predefined keywords in an audio stream, while VAD detects when a human voice is present.

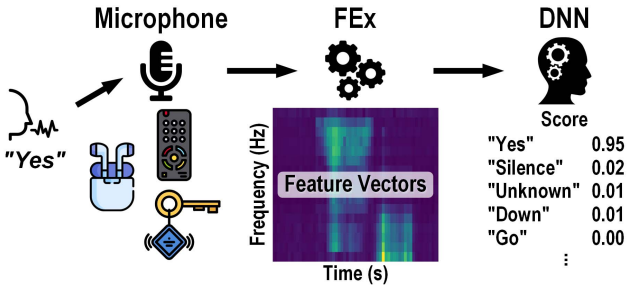


Fig. 1. Processing stages for KWS in an audio IoT device.

Fig. 1 shows the typical processing stages for KWS. The user says a keyword into the microphone of an edge device, such as a remote control or wireless earbud. The microphone output is further processed by a feature extractor (FEx), which generates frequency-selective feature vectors (FVs) that are continuously streamed to a deep neural network (DNN)-based classifier. The classifier outputs the probability scores of different keywords. The IoT devices benefit from a tiny form factor and the use of a small battery, such as a coin cell, e.g., for smart tags. Generally, a <100-μW system-level power is desirable, including not only the KWS IC itself but also the microphone and other system components. Moreover, a low-latency response is desired considering a KWS-driven hierarchical processing system used in an interactive environment. For example, a study on the perception of the self-generated speech showed that a delay exceeding 20 ms becomes disturbing for users [1].

A 12-class KWS IC that includes the whole processing chain starting from the analog-to-digital converter (ADC) to the DNN classifier [2] reported that the FEx is the most power-hungry stage accounting for 40% of the power dissipation in the entire IC. To reduce the power budget of edge

Manuscript received 16 April 2022; revised 22 June 2022 and 26 July 2022; accepted 28 July 2022. Date of publication 17 August 2022; date of current version 24 October 2022. This article was approved by Associate Editor Fabio Sebastiano. This work was supported in part by the Swiss National Science Foundation, HEAR-EAR, under Grant 200021-172553; and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2020-0-01847. *(Corresponding author: Shih-Chii Liu.)*

Kwantae Kim was with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea. He is now with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, 8057 Zürich, Switzerland (e-mail: kwantae@ini.uzh.ch).

Chang Gao was with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, 8057 Zürich, Switzerland. He is now with the Delft University of Technology, 2628 CD Delft, The Netherlands.

Rui Graça, Ilya Kiselev, Tobi Delbruck, and Shih-Chii Liu are with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, 8057 Zürich, Switzerland (e-mail: shih@ini.uzh.ch).

Hoi-Jun Yoo is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2022.3195610.

Digital Object Identifier 10.1109/JSSC.2022.3195610

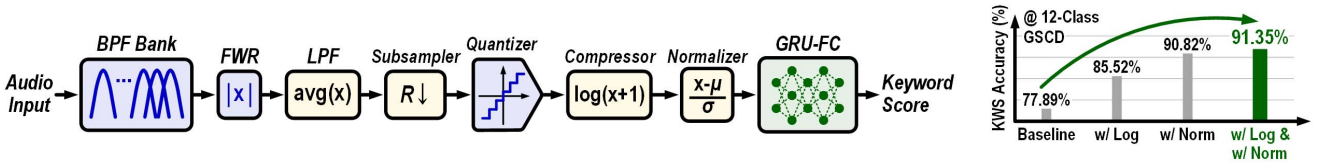


Fig. 2. Architecture of the KWS software model (left) and simulated KWS accuracy (right).

devices, thereby facilitating longer battery lifetime or smaller battery size, various circuit design techniques have been proposed for both KWS and VAD ICs. However, most of them traded off between power and latency. In [3], a 142-nW VAD IC using sequential mixer-based FEx was proposed where the operational principle is similar to that used for bio-impedance sensors [4], [5]. However, this sequential frequency scanning is too undersampled for the KWS and results in a 512-ms latency for VAD. In [6], a serialized digital FEx was used in the KWS IC where the processing stages are pipelined. Although this IC consumed only 510 nW, its latency was limited to 64 ms, and it needed an off-chip 16-bit ADC and had only 2-KB memory for binary convolutional neural network (CNN); thereby, its accuracy was only reported for five keywords.

Another approach is the use of an analog voltage-domain FEx, which exploits low-power analog circuits to achieve both low-power and low-latency responses. The processing chain of the analog FEx typically consists of a multi-channel bandpass filter (BPF), a half-wave rectifier (HWR) or a full-wave rectifier (FWR), and an ADC. Here, the speed requirement of ADC is highly relaxed to 10–100 ms (10–100 Hz), which corresponds to the size of frame shift in audio signal processing. This is possible because the output of the rectifier represents the magnitude response of the input speech, and thus, it is a low-frequency signal. Previous works that used a voltage-domain analog FEx and a back-end classifier to implement VAD [7], [8] and KWS [9] tasks reported 205-nW-to-1-μW power dissipation and 10–100-ms latency. However, voltage-domain analog FEx is unfriendly for CMOS technology scaling; thereby, the power efficiency of analog approaches is predicted to be degraded in advanced nanometer-scale processes. This is because V_{DD} is scaling down faster than V_{TH}; thus, voltage-domain signals have less headroom. Reduced headroom results in reduced maximal signal swing, which, in turn, reduces the dynamic range (DR) that is critical for keeping KWS accuracy high across a range of audio amplitude levels. Furthermore, the intrinsic gain ($g_{\mathrm{m}}r_{\mathrm{o}}$) of the transistors is also degraded, leading to the dc gain reduction in analog feedback loops. This issue can be mitigated with a larger transistor length, gain boosting, or multistage amplifiers; however, these approaches come with costs in the area, power, and bandwidth.

To this end, we propose a time-domain analog FEx that exploits the scaling-friendly nature of the ring oscillator. It is the first silicon-verified ring-oscillator-based audio FEx reported to date. When integrated with an on-chip recurrent neural network (RNN) classifier, the resulting IC demonstrates power-efficient KWS capability. The FEx circuits extensively use time-domain signal representation techniques, including pulsewidth modulation (PWM) and pulse-frequency modulation (PFM); therefore, it does not suffer from headroom degradation and its associated signal swing loss issue. In other words, it is more suitable for low-supply implementation than voltage-domain designs. The ring-oscillator-based circuit utilizes its infinite dc gain characteristic when configured as a time-domain integrator [10]. As such, the transfer function of time-domain FEx circuits, such as BPF, is not affected by the degradation of the intrinsic gain of transistors. Overall, the proposed KWS IC consumes 23 μW and has only 12.4-ms inference latency on a 12-class Google Speech Command dataset (GSCD) [11].

There have been similar approaches to implement the oscillator-based BPFs for audio IoT applications [12], [13]. However, none of them proposed a clear design strategy to implement a time-domain rectifier or demonstrated an audio classification task using the fabricated oscillator-based BPFs.

This article is an extension of a previous work presented in [14]. The integrated chip also includes a switched-capacitor energy harvester circuit, a voltage reference, and a low-dropout regulator. However, in this article, we focus on the new circuit techniques of the KWS core. This article is organized as follows. Section II presents the software modeling of the KWS modules in this work. Section III covers the description of the overall architecture and design details of the implemented circuits. Section IV presents measurement results and a performance summary of the prototype chip. Section V concludes this work.

II. Software Modeling

The architecture of our KWS IC was developed based on prior silicon cochlea and edge audio-inference ICs [7]–[9], [15], [16]. We implemented a Python model of the KWS IC, including the analog FEx, as shown in Fig. 2. Our model implements a bank of BPFs (second-order Butterworth filter) inspired by modeling of biological cochlea [16], an FWR ($|x|$), an averaging block (low-pass filter), a subsampler, and a quantizer. The subsampler was added to realize the relaxed speed requirement of the quantizer, as discussed in Section I. As the GSCD samples have a 16-kHz sampling rate, the number of averaged samples and the rate of subsampling operation were selected to match the target frame shift size (16 ms in our work) of the audio feature vector (FV). In contrast to prior analog FExs [7], [8], we added additional FV processing stages before the FV is fed to the classifier. These stages consist of: 1) a logarithmic compression stage inspired by the adaptive gain compression mechanism of biological cochleas and 2) an input normalization stage that is widely used in DNN models, both of which help to improve the KWS accuracy on GSCD. We chose a gated recurrent unit (GRU)-based RNN classifier for the last stage of our KWS, as it has been frequently used in automatic speech recognition tasks [17].

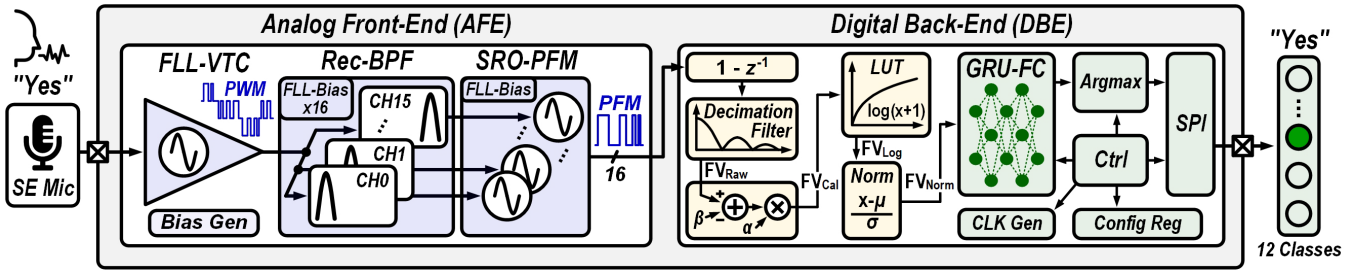


Fig. 3. Overall architecture of the proposed KWS IC.

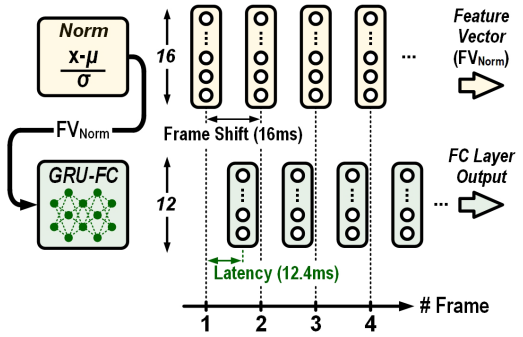


Fig. 4. Timing diagram of the GRU-FC classifier computation according to the input feature vectors.

Fig. 2 shows the accuracy of the software simulation starting from the baseline model, which does not include the compressor and normalizer stages. As seen on the right graph of Fig. 2, the baseline model achieved 77.89%, which increases to 91.35% KWS accuracy on the 12-class GSCD test set with the addition of the two stages. The following design parameters were chosen for our software model. First, we used a 16-channel BPF that was also used in previous works [7]–[9] and a Q-factor of 2 for the BPFs. Second, the center frequencies of the bank of BPFs are distributed according to the Mel scale (from 100 Hz to 8 kHz). The 8-kHz value is also the bandwidth of the analog front end presented in [2]. Note that we oversampled the input speech 2× (from 16- to 32-kHz sampling rate) to avoid the 8-kHz center frequency overlapping with Nyquist frequency (8 kHz with a 16-kHz sampling rate). The third is a 12-bit quantizer (before logarithmic compression in Fig. 2). The fourth is a 16-ms frame shift; the same value was used in [2] and [6]. The fifth is a 10-bit output logarithmic compressor and a 14-bit normalized feature vector (FV_{Norm}) that is fed to a two-layer 48-hidden-unit GRU and an fully connected (FC) layer. Sixth, 14- and 8-bit quantizations were applied to the activations and weights, respectively. The baseline model accuracy shown in Fig. 2 would be higher if using floating-point activations because the 14-bit quantization (6-bit integral part and 8-bit fractional part) cannot cover the dynamic range of the 12-bit unsigned quantizer output.

III. KWS IC With Time-Domain Analog FEx

The overall architecture of our KWS IC is shown in Fig. 3. It is composed of an analog front end and a digital back end. The analog front end is designed to match our software model. The first stage of the analog front end is a frequency-locked loop (FLL)-based voltage-to-time converter (VTC) (see Section III-A). It features a nested analog FLL circuit that linearizes the voltage-to-frequency response of the voltage-controlled oscillator (VCO). A voltage-domain audio input from a single-ended (SE) microphone (Mic) is converted into a time-domain multi-phase PWM output through the VTC. The second stage of the analog front end is a 16-channel rectifying BPF (Rec-BPF). Each channel has a time-domain second-order BPF (see Section III-B) featuring an inherent FWR functionality. The output of each BPF channel is the PWM signals, and they are further converted into PFM signals through a switched-ring oscillator (SRO)-based rate encoder.

In the digital back end, the PFM signals are fed into a digital differentiator ($1 - z^{-1}$). Here, the signal path from the SRO to the digital differentiator builds a first-order $\Delta\Sigma$ time-to-digital converter (TDC) [18], which corresponds to the quantizer in our software model. The output of the digital differentiator is further processed through subsequent stages, including a decimation filter, which performs the averaging and subsampling operation in our software model. It also includes an offset subtractor (β) that removes the free-running frequency component of the SRO, a per-channel gain calibrator (α) that corrects the inter-channel gain mismatch, a logarithmic lookup table (LUT), and an input normalizer. Both μ and σ shown in Fig. 3 are, respectively, the mean and the standard deviation of the output of logarithmic LUT (FV_{Log} in Fig. 3) from our chip with the GSCD training set. With the normalizer, μ is subtracted from the FV_{Log}, and the resulting subtracted output is multiplied by a value $1/\sigma$. The resulting output of the FV_{Norm} is a 16-channel signed 14-bit FV, which is generated every 16 ms of a frame shift, as shown in Fig. 4. For each FV, a two-layer GRU RNN and a one-layer FC digital accelerator output the most probable keyword over 12 classes with a 12.4-ms latency (see Fig. 4).

A. Voltage-to-Time Converter

Previous VAD and KWS ICs have used mainly a differential output microphone interface [2], [8], [9]. However, commercial off-the-shelf differential output micro-electromechanical systems (MEMS) microphones typically consume >100 μW. To realize a system-level low-power audio IoT device, a low-power SE-interface MEMS microphone is preferred because it consumes as little as ∼10 μW (e.g., InvenSense ICS-40310 [19]). However, this approach makes it difficult to obtain good linearity because SE signals do not reject even-order harmonics. In general, a linear FEx is preferred as

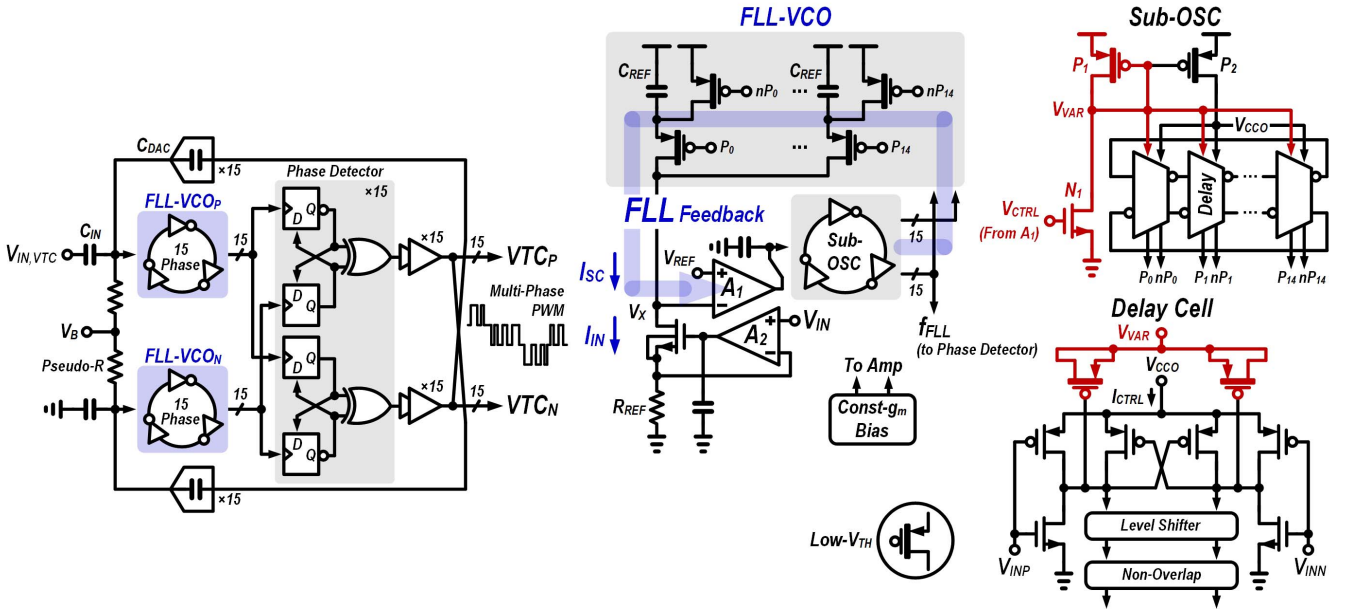


Fig. 5. 0.5-V supply SE input FLL-based VTC.

it makes training the back-end DNN classifier easier and enhances the spectral purity of an audio signal with minimal harmonics and intermodulation distortions. Particularly, the design of an SE-input ring-oscillator-based VTC circuit becomes even more difficult because VCOs exhibit poor linearity compared to a conventional voltage-domain operational transconductance amplifier (OTA).

In this work, we propose to use a nested analog FLL around the ring-VCO to enhance the linearity of the VTC. Fig. 5 shows the architecture and transistor-level schematic of the FLL-based VTC. The fundamental design principle is adopted from the ring-oscillator-based low-pass filter (LPF) presented in [10]; however, capacitive coupling is used with C_{IN} to isolate the dc bias of the VTC from the microphone. Therefore, its core operation is similar to the capacitively coupled voltage amplifier [20], but the VTC circuit converts the input voltage into the multi-phase PWM output instead of voltage. A pseudo-differential architecture is implemented using a dual-VCO structure along with the phase detector [10]. One input port of the VTC is connected to the SE microphone, and the other input port is tied to the ground. The 15-array phase detector receives a 15-phase frequency-modulated signal out of the VCOs and generates a 15-phase PWM output, which represents the phase difference of the VCOs. Note that exploiting the multi-phase PWM scheme pushes spurious PWM tones to a higher frequency range without necessitating a higher running frequency of the VCO [10]. The outputs of the phase detector are buffered with two inverters and used to close the feedback loop through a 15-array thermometer-coded capacitive digital-to-analog converter (DAC). Since the input node of the VCO acts as a virtual ground, the generated multi-phase PWM signal becomes a time-domain approximated input voltage where the amplitude is encoded into the duty-cycle of PWM. A variation-tolerant pseudo-resistor [21] with a voltage reference V_{B} sets the common-mode dc bias voltage of the VCOs.

The FLL-based VCO includes a single-branch current comparator [22] to operate the analog FLL. As shown in the schematic of the FLL-VCO in Fig. 5, an input current generator drives the input voltage to R_{REF} to generate a low-side current signal $I_{\mathrm{IN}} = V_{\mathrm{IN}}/R_{\mathrm{REF}}$, where A_2 amplifier is designed to have a 34-dB gain. A high-side current $I_{\mathrm{SC}} = 15V_{\mathrm{X}}C_{\mathrm{REF}}f_{\mathrm{FLL}}$ flows through a 15-phase switched-capacitor operation. Here, the multi-phase nature of a ring oscillator is fully utilized to apply the multi-phase interleaving technique at the V_{X} node to minimize the voltage ripple caused by the switched-capacitor operation [22]. The low-V_{TH} devices are used to facilitate 0.5-V low-supply operation for the implementation of the switched-capacitor circuit. The FLL feedback formed through the A_1 amplifier with a 27-dB gain, sub-oscillator (OSC), and switched-capacitor circuit ensures that V_{X} equals the reference voltage V_{REF} while also ensuring that I_{IN} equals I_{SC}. As a result, the output frequency of the FLL-based VCO is set as in (1)

$$f_{\mathrm{FLL}} = \frac{V_{\mathrm{IN}}}{15R_{\mathrm{REF}}C_{\mathrm{REF}}V_{\mathrm{REF}}} \tag{1}$$

$$K_{\mathrm{FLL\text{-}VCO}} = \frac{\partial f_{\mathrm{FLL}}}{\partial V_{\mathrm{IN}}} = \frac{1}{15R_{\mathrm{REF}}C_{\mathrm{REF}}V_{\mathrm{REF}}}. \tag{2}$$

Since f_{FLL} is represented by the input voltage V_{IN} and reference parameters, such as R_{REF}, C_{REF}, and V_{REF}, it leads to an FLL-aided linearization of the VCO, as derived in (2), where $K_{\mathrm{FLL\text{-}VCO}}$ corresponds to the voltage-to-frequency tuning gain. This is because the value of passive elements (R_{REF}, C_{REF}, and V_{REF}) has no dependence on the input signal amplitude (V_{IN}).

The 3-dB bandwidth of the VTC circuit is given in the following, which is similar to the equation of resistive-input and current-feedback ring-oscillator-based filter [10]:

$$f_{3\ \mathrm{dB,VTC}} = \frac{1}{2\pi}K_{\mathrm{FLL\text{-}VCO}}K_{\mathrm{PD}}\beta_{\mathrm{DAC}} \tag{3}$$

$$\beta_{\mathrm{DAC}} = \frac{15C_{\mathrm{DAC}}}{C_{\mathrm{IN}} + 15C_{\mathrm{DAC}}}V_{\mathrm{DD}} \tag{4}$$

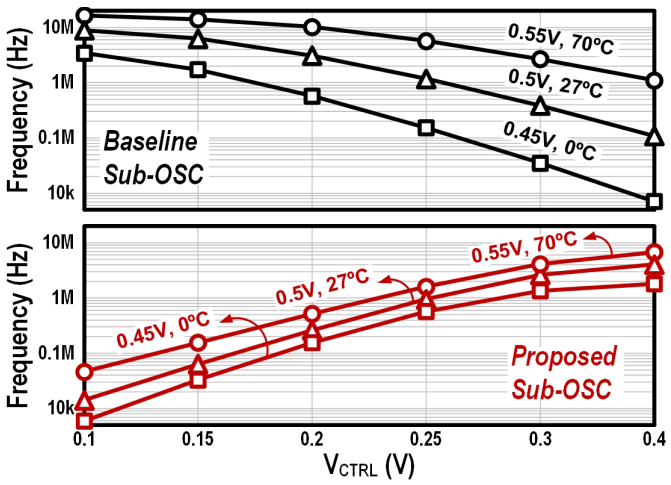


Fig. 6. Simulation result of the supply temperature compensation in sub-OSC.

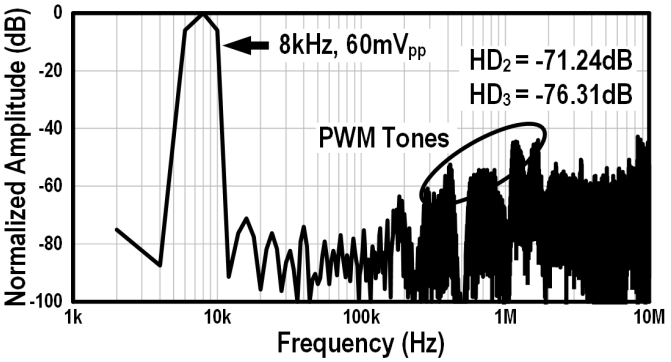


Fig. 7. Simulation result of the FLL-based VTC.

where K_{PD} is the gain of phase detector and β_{DAC} is the feedback factor (time-to-voltage). We designed $f_{3\,dB,VTC}$ to be 17 kHz when the nested FLL feedback has a 158-kHz gain-bandwidth product, which contributes as a nondominant pole to the overall negative feedback loop of the VTC circuit. As shown in Fig. 7, the stability of the VTC is verified with a transient simulation.

To allow 0.5-V low-supply operation for the sub-OSC, a varactor-controlled supply temperature compensator is proposed. This is achieved by sizing the diode-connected transistor P_1 so that V_{VAR} becomes proportional-to-absolute-temperature (PTAT) [23]. Therefore, the capacitance of MOS-varactors in the delay cells [24] adaptively stabilizes the temperature drift of the ring-oscillator frequency. For example, if the temperature increases, then V_{VAR} also increases; therefore, the MOS-varactors are further turned on. This effectively negates the frequency increase in the ring oscillator with a temperature increase. Low-V_{TH} MOS capacitors are used to further enhance the varactor effect. In addition, instead of configuring the sub-OSC as controlled by the gate voltage of P_2 only, the $N_1 - P_1$ path is added to reduce V_{DD} sensitivity based on the fact that $V_{GS,N1}$ is less sensitive to V_{DD} than $V_{SG,P2}$. The simulation results in Fig. 6 show that, with the proposed techniques, the supply temperature variation of the sub-OSC is reduced by 19.98× in the worst case. Note that the baseline sub-OSC refers to the OSC circuit, assuming that

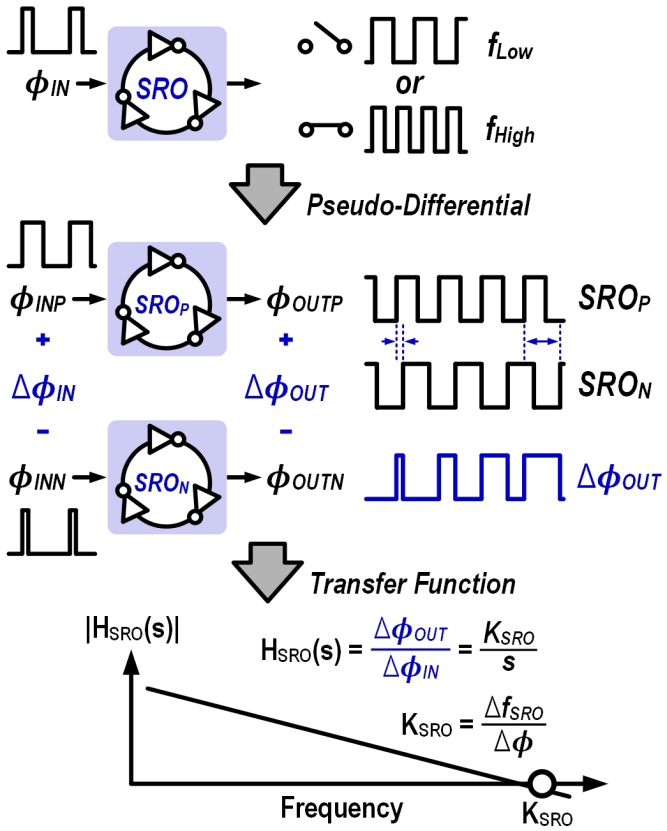


Fig. 8. SRO as an ideal ϕ-to-ϕ integrator.

the added compensation circuits (marked as red in Fig. 5) are removed. In this case, V_{CTRL} is connected to the gate of the P_2 transistor, and therefore, the frequency tuning curve of sub-OSC becomes decreasing function as V_{CTRL} increases. Fig. 7 shows the post-layout simulation result of the VTC. The plotted graph represents the multi-phase PWM signal of VTC output ($VTC_P - VTC_N$). The designed VTC converts a voltage-domain input into a time-domain PWM output while ensuring <-70-dB distortion for dominant harmonics (second and third) even with an SE input. The PWM tones at higher frequencies are filtered out at the following BPF stage.

B. Time-Domain Bandpass Filter

Fig. 8 shows a conceptual diagram of using the SRO [18] as an ideal ϕ-to-ϕ integrator. The SRO switches its running frequency between f_{Low} and f_{High} according to the incoming input PWM signal. The averaged value of SRO frequency is proportional to the duty-cycle of the input PWM signal. If the input PWM signal is configured as a multi-phase format, the possible number of running frequencies also increases. When the dual-SRO is implemented in a pseudo-differential manner, the output phase difference $\Delta\phi_{OUT}$ becomes an accumulated (or integrated) input phase difference $\Delta\phi_{IN}$ over time. Specifically, this integral procedure flows as follows: Input Phase → SRO Frequency → SRO Phase (+Integral) → Output Phase. The phase mathematically represents an integral amount of the frequency within an oscillator. This time-domain accumulation process allows an integral of the signal without boundary as long as the SRO oscillates, unlike voltage-domain designs that

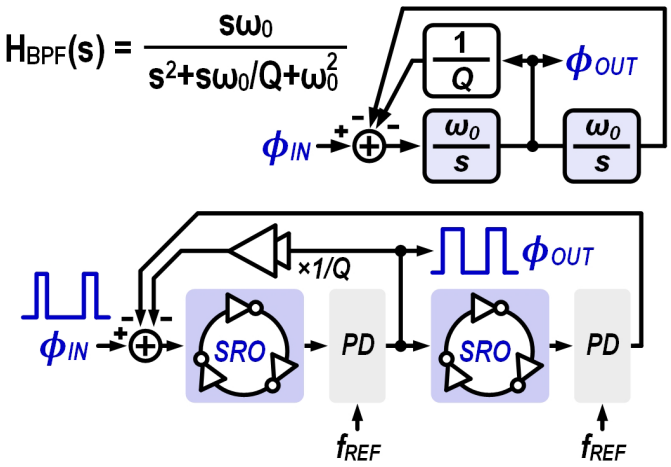


Fig. 9. Block diagram of a time-domain second-order BPF using SRO as a core building block with a half-circuit representation.

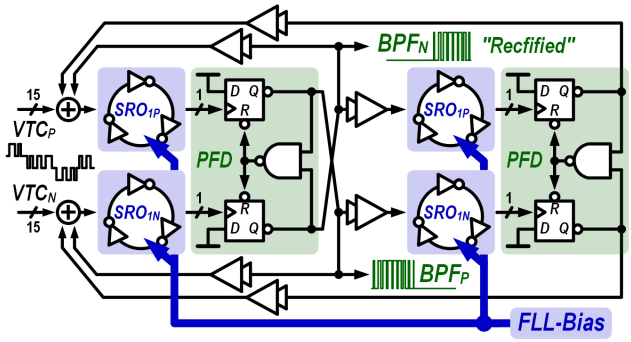


Fig. 10. 0.5-V supply time-domain rectifying BPF.

saturate due to headroom. In other words, it shows an infinite dc gain and acts as a true lossless integrator regardless of the intrinsic gain of transistor or supply voltage level [10]. As shown in the lowermost description of Fig. 8, the ϕ-to-ϕ transfer function is described by K_{SRO}/s, where K_{SRO} is the switching gain of an SRO.

Fig. 9 shows a conceptual diagram for the implementation of a time-domain ϕ-to-ϕ BPF. It adopts the two-integrator-loop Tow–Thomas biquad topology [25], [26] using SRO as a lossless integrator (ω_0/s). The phase detector (PD) extracts the phase difference between two input signals and outputs the phase difference in a PWM signal. The output of PD is used to close the feedback loops where the inner feedback loop ensures the desired Q factor, and the outer feedback loop generates the high-pass shape of the BPF. Overall, the time-domain BPF receives the PWM input and generates the PWM output. Note that an external clock f_{REF} is fed to the PD in Fig. 9 since it is represented as a simplified half-circuit diagram. If the two BPFs are placed in parallel to work as a pseudo-differential configuration, as shown in Fig. 8, the external clock f_{REF} is no longer needed, and the BPF operates in a fully asynchronous way. The same consideration for a pseudo-differential topology also applies to the VTC design, as described in Section III-A.

Fig. 10 shows the block diagram of the proposed time-domain BPF. It receives the multi-phase PWM output of the VTC as an input signal and does not require an external clock. It incorporates four SROs and two phase frequency detector (PFD). The outputs of the BPF are two single-phase PWM signals. The two PFDs implemented in the BPF offer an inherent rectification function in the time domain, which will be discussed in Section III-C. A local FLL-based bias generator provides the required bias voltage, which is shared over the four SROs. This bias voltage is different over 16-channel BPF bank to set different center frequencies. Note that the outputs of first PFD are crossed and connected to the SROs with opposite polarities to realize a subtraction function.

Fig. 11 shows a schematic of the FLL-based bias generator and SROs used in our BPF design. The SRO receives time-domain PWM signals as input, such as VTC and PFD. All the PWM signals are summed at the internal node of the SRO, and these signals drive the buffers that act as an array of current-mode DACs. Therefore, the output frequency of SRO is proportional to the sum of incoming PWM signals. To realize different switching gains of PWM inputs, switching transistors are differently sized. The unit current for current-DAC operation is provided by a local FLL circuit. The FLL acts as a bias generator for the realization of per-channel center frequency designs in BPFs, using a replica biasing scheme. As shown in Fig. 11, the bias voltage V_{VAR} is generated from a diode-connected pFET in the FLL-bias circuit. Therefore, the current-DAC in $\mathrm{SRO}_{1\text{-}2}$ operates as a current mirror when V_{VAR} is shared over the four SROs from the FLL-bias circuit. This means that the switching gain of each PWM input signal is determined by f_{FLL} and the sizing ratio of the current-mode DAC. For example, the switching gain of VTC-port is $K_{\mathrm{IN}} f_{\mathrm{FLL}}$, and the switching gain of PFD_2-port fed into the SRO_2 is $K_2 f_{\mathrm{FLL}}$. Note that we adopt the same circuit structure from the sub-OSC circuit in Section III-A, which allows the BPF to work at 0.5-V low-supply voltage. As discussed in (1), the locking frequency of the FLL circuit is proportional to $1/C_{\mathrm{REF}}$. To cover the target range of BPF center frequencies ranging from 100 Hz to 8 kHz in our design, a coarse-fine approach is used. The output of SRO is divided coarsely by N times using D flip-flop (D-FF), and the C_{REF} of FLL circuit is fine controlled through proper sizing. The complete transfer function $H_{\mathrm{BPF}}(s)$ of the proposed time-domain BPF is given in (5). Its center frequency ω_0 and Q-factor are given in (6), where the Q-factor is designed as 2 for each BPF channel by proper sizing of the switching transistors in the SROs. The stability of the proposed second-order time-domain BPF is verified with a transient simulation

$$H_{\mathrm{BPF}}(s) = \frac{\frac{sK_{\mathrm{IN}} f_{\mathrm{FLL}} K_{\mathrm{PFD}}}{N}}{s^2 + s\frac{K_1 f_{\mathrm{FLL}} K_{\mathrm{PFD}}}{N} + \frac{K_1 K_1 f_{\mathrm{FLL}}^2 K_{\mathrm{PFD}}^2}{N^2}} \quad (5)$$

$$\omega_0 = f_{\mathrm{FLL}} K_{\mathrm{PFD}} \sqrt{\frac{K_1 K_2}{N}} \qquad Q = \sqrt{\frac{K_2}{K_1}}. \quad (6)$$

C. Time-Domain Rectifier

Fig. 12 shows the schematic of FWR. The proposed time-domain FWR is based on a simple PFD circuit consisting of only two D-FFs and one NAND gate. Compared to the

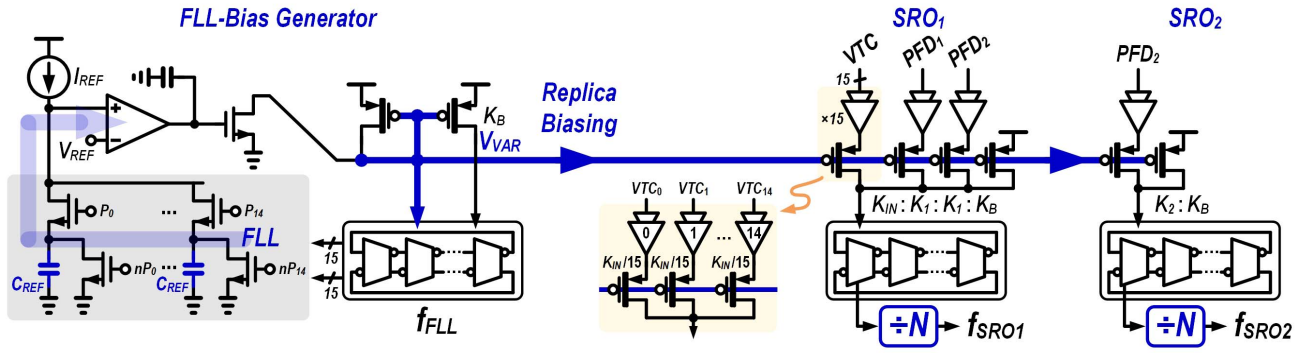


Fig. 11. Schematic of FLL-based bias generator and SRO.

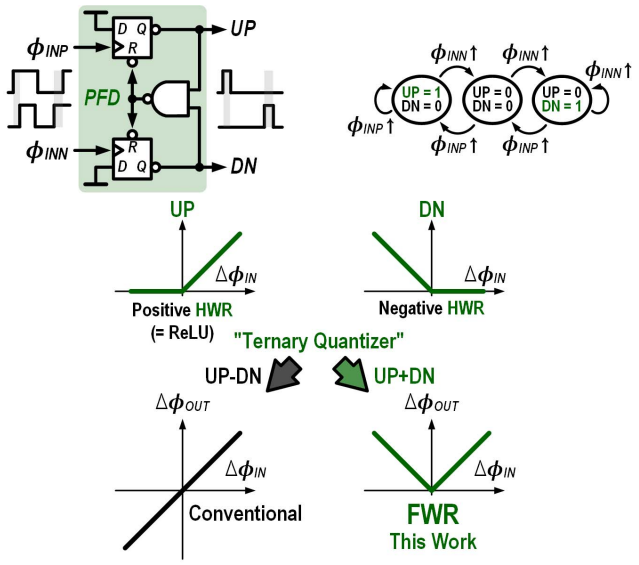


Fig. 12. Proposed time-domain FWR with operational principle of the PFD.

prior voltage-domain design [8] that required several scaling-unfriendly OTAs, references, and passive elements, this work offers an alternative solution that is fully compatible with standard logic gates. Fig. 12 shows a state diagram and an input–output characteristic of the PFD. The PFD circuit extracts the input phase difference $\Delta\phi_{\mathrm{IN}}$ but, at the same time, asynchronously quantizes the phase difference using a ternary code with UP and down (DN) signals. As shown in the state diagram, there are three states that are activated by the rising edges of incoming PWM signals (ϕ_{INP} and ϕ_{INN}). When both UP and DN signals are high, the NAND gate resets two D-FFs immediately, thereby making itself a ternary quantizer. Since the state of PFD stays the same unless a new rising edge arrives, the UP and DN signals represent a positively and negatively half-wave rectified (HWR) phase difference ϕ_{IN}, respectively. Interestingly, the UP signal has the same form as the rectified linear unit (ReLU) activation function widely used in modern DNNs. In conventional usage of such signals like in phase-locked loop (PLL) designs, they are subtracted to derive a linearized phase difference extractor. However, if we add them, a time-domain FWR can be implemented. The PFD-based FWR benefits from its fully time-domain nature, that is, it does not exhibit a headroom-related saturation, assuming that the input signal swing ($\Delta\phi_{\mathrm{IN}}$) is within $\pm 2\pi$ range. As shown in Fig. 10, the proposed time-domain FWR is seamlessly integrated within the time-domain BPF circuit, and thus, the BPF provides an inherent rectification function. The rectified PWM signals ($\mathrm{BPF_{P/N}}$) are summed at the subsequent PFM stage, as described in Fig. 13.

D. Pulse-Frequency Encoder and Time-to-Digital Converter

The analog FEx designs presented in [8] and [9] used an integrate-and-fire (IAF) circuit, which was originally proposed in [27]. The circuit converts an input current into a rate-encoded spiking PFM signal, and its spiking frequency is proportional to the input current magnitude. The IAF circuit can be interpreted as a current-controlled oscillator (CCO) where the core oscillator topology is equivalent to a relaxation oscillator [28]. However, the IAF circuit is scaling-unfriendly because of its voltage-domain integral operation and voltage-domain static amplifier, as discussed in Section I. In this work, we propose to use the SRO as a PFM encoder instead of the IAF circuit. As shown in Fig. 14, the SRO exploits an inherent phase-domain 2π threshold, and its integral operation occurs in the phase domain, which is free from headroom issues. The proposed SRO-based design offers a scaling-friendly implementation using only logic gates and a bias generator without a static amplifier or passive components.

In previous designs [8], [9], an asynchronous ripple-carry counter associated with a multi-bit register was used to quantize and sample the input PFM signal. Interestingly, given the aforementioned interpretation of the IAF circuit as an oscillator, the signal flow from an IAF to a counter builds an VCO/CCO-based $\Delta\Sigma$ modulator [18], [29]. However, the design approaches used in [8] and [9] had two major problems. First, the ripple-carry counter exhibits metastability-induced data corruption when the sampling occurs at the instant of the multibit transition of binary codes. Second, the output digital data from the asynchronous counter, which is $\Delta\Sigma$ modulated, were directly fed to the DNN classifier without filtering of high-pass-shaped quantization noise. Our approach uses arrayed 1-bit XOR differentiators [30] to solve the metastability problem and an oversampling associated with a decimation filter to filter out quantization noise.

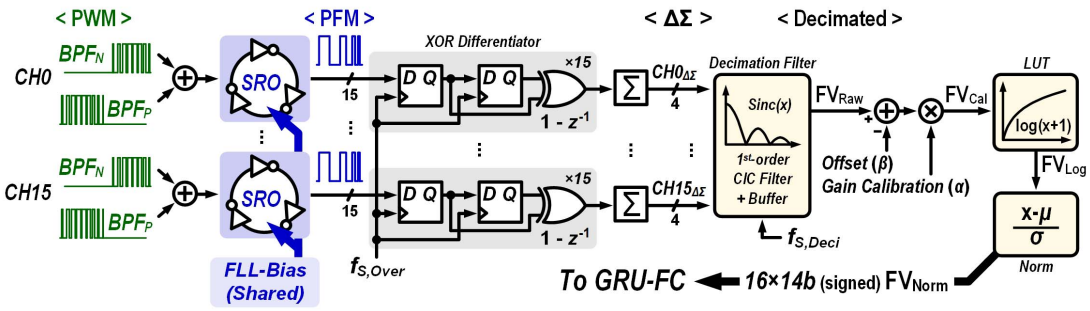


Fig. 13. SRO-based PFM encoder, XOR differentiator, and subsequent post-processing blocks.

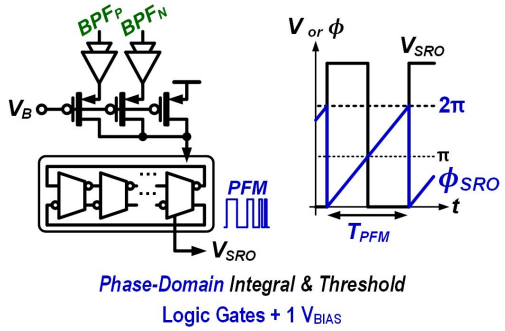


Fig. 14. Proposed SRO-based PFM encoder.

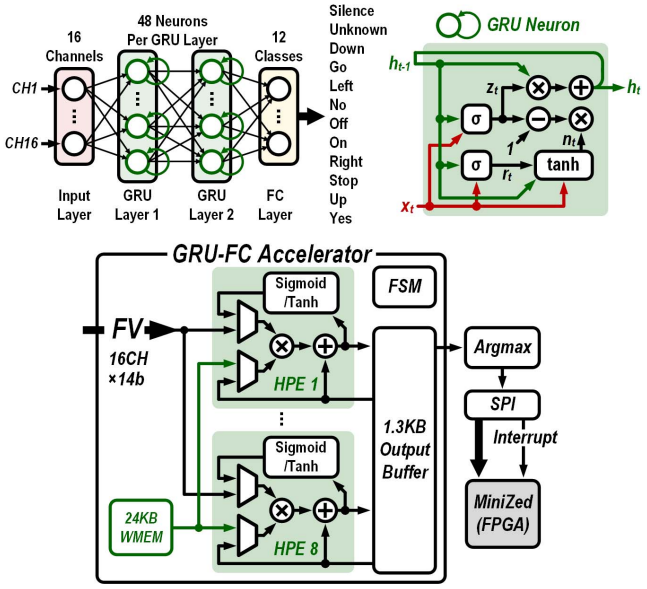


Fig. 15. Architecture of the GRU-FC classification network (upper) and accelerator (lower).

Fig. 13 shows the overview of implemented 16-channel SRO-based PFM encoder, XOR differentiator, and subsequent post-processing stages, also with the signal flow domains at the top. The SRO receives rectified PWM signals from the preceding BPF stage and converts it into 15-phase PFM signals. The same design of the FLL circuit discussed in Section III-B is reused for biasing of the SRO where the generated bias voltage is shared over 16 channels. As the ring-OSC output is represented in the thermometer code, the XOR differentiator ensures the worst case error to be within 1-least significant bit (LSB). In addition, this 1-LSB error is noise-shaped [31], which can be eliminated through oversampling and decimation filtering. The thermometer-coded output data are aggregated to be represented in binary format and then filtered and decimated through a first-order cascaded integrator-comb (CIC) filter. We use 2^{10} decimation size, i.e., $f_{S,\mathrm{Deci}} = f_{S,\mathrm{Over}}/2^{10}$, and $f_{S,\mathrm{Deci}}$ is used in the post-processing blocks, which incorporates a programmable offset (β) subtractor to remove the dc offset due to a free-running component of the SRO-based PFM encoder. A programmable per-channel gain calibrator (α) is used to correct inter-channel gain deviations caused by a mismatch of SROs in the PFM encoder. A logarithmic compression using an LUT and a programmable input normalizer helps to increase the classification accuracy of the following GRU-FC neural network. The post-processing stage is clocked at 61-Hz $f_{S,\mathrm{Deci}}$, and thus, its power dissipation is negligible.

E. Recurrent Neural Network Accelerator

Fig. 15 shows the architectures of the GRU-FC network and accelerator. The network has two GRU layers with 48 units per layer and a final FC layer that generates the confidence scores of the 12 classes. The network model size is entirely buffered within the 24-kB weight memory (WMEM). The accelerator computes the KWS classifier network, and its input comes from the normalizer shown in Fig. 13. The WMEM block is implemented by on-chip static random access memory (SRAM) compiled based on the foundry-provided six-transistor (6T) bit cell. The classifier weights are loaded into WMEM over the SPI interface. The accelerator has eight heterogeneous processing elements (HPEs) controlled by a finite-state machine (FSM). Each HPE has a 14-bit multiplier, a 24-bit accumulator, and an LUT-based sigmoid/tanh unit. Partial sums of the multiply-and-accumulate operations and outputs are stored in a shared 1.3-kB SRAM output buffer. Multiplexers before each multiplier operand select inputs from the normalizer, the sigmoid/tanh unit, the WMEM, and the output buffer to compute element-wise vector multiplication/addition and hyperbolic functions in the GRU RNN. The high V_{TH} device library is used for logic synthesis to reduce leakage current. Output scores of the classifier are fed to the argmax decoder, which outputs the class with the highest score. The classification result is transmitted over the SPI interface with an interrupt flag to the external host, which is a MiniZed board with a Xilinx Zynq-7007S SoC.

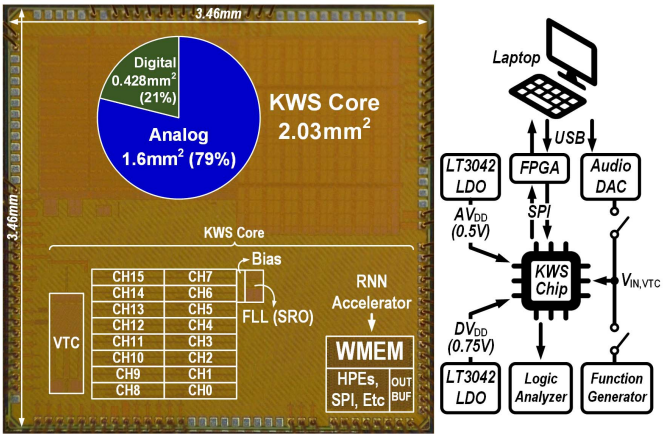


Fig. 16. Chip photograph of the prototyped KWS IC with a block diagram of the measurement setup.

F. Network Training

1) Dataset Preparation: Our GSCD training set is composed of 38 463 samples. The number of samples in the "Silence" class is 4044 that are randomly sampled from the background noise tracks in the dataset. The "Unknown" class also has 4044 samples that are randomly chosen words outside the target 12 classes. As for the test set, we used the standard GSCD test set,[1] which has a roughly equal number of samples (around 400) among the 12 target classes. Thus, the ratio between the training and test sets is around 8:1. As shown in the measurement setup of Fig. 16, the samples from our entire training and test set were played from a laptop to $V_{\text{IN, VTC}}$ through a USB sound card DAC (Sound Blaster E1). We normalized the GSCD samples with the mean and standard deviation of the entire samples such that the amplitude of $V_{\text{IN, VTC}}$ is set to ~250 mV_{PP}. The corresponding FV_{Raw} from all samples was recorded. They were then corrected for the dc offset (β) and the inter-channel gain deviation (α). After applying the logarithmic compression, we then normalize FV_{Raw} with the mean (μ) and standard deviation (σ) of the recorded feature vectors from the entire GSCD training set. This resulting vector called FV_{Norm} (see Fig. 3) is then presented as inputs to the GRU-FC classifier during training. The same μ and σ are applied to FV_{Log} of the test set to generate the corresponding FV_{Norm} for KWS evaluation.

2) Training Schedule: The network is built in the PyTorch 1.8 framework and trained for 200 epochs using the AdamW optimizer [32] with an initial learning rate of $1e-3$ and 0.01 weight decay. The ReduceLROnPlateau learning rate scheduler is used with a decay factor of 0.8 and patience of 3 epochs. The lowest learning rate is $5e-4$. Using quantization-aware training, the activations and weights are quantized to 14 and 8 bit, respectively.

IV. Measurement Results

Fig. 16 shows the KWS IC, which is fabricated in TSMC 65-nm CMOS LP process with an active area of 2.03 mm^2 for the KWS core. The area occupied by the analog and digital circuits is 1.6 mm^2 (79%) and 0.428 mm^2 (21%), respectively, in the KWS core. The GRU-FC neural network accelerator and the associated peripherals in the digital circuits are synthesized from a standard auto place-and-route (P&R) flow.

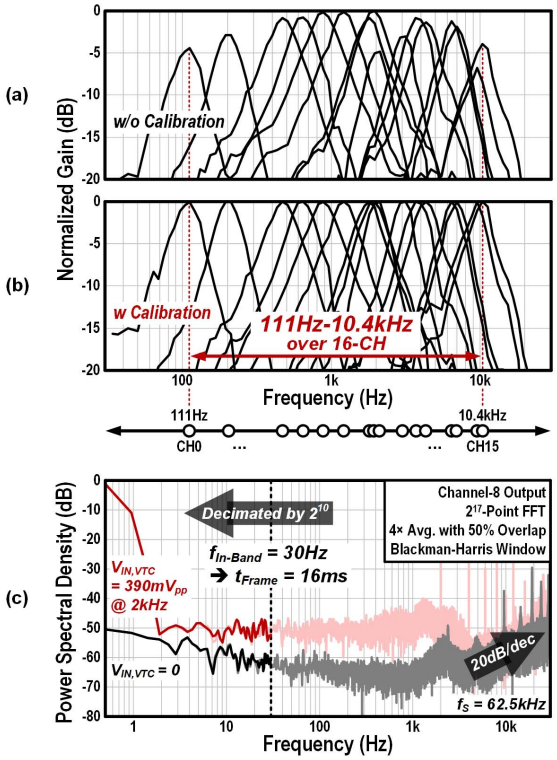


Fig. 17. Measured frequency response of the FEx (a) without per-channel correction, (b) with per-channel gain (α) correction, and (c) output spectrum of FV_{Raw} (after decimation filter) in channel 8, as shown in Fig. 13.

Fig. 17(a) and (b) shows the measured frequency response of the 16-channel FEx with and without per-channel gain calibration. In this case, $V_{\text{IN, VTC}}$ was connected to a function generator. The center frequencies of the 16 BPF channels range from 111 Hz to 10.4 kHz. The center frequencies are distributed according to the Mel scale; therefore, low-frequency (<1 kHz) channels are spaced further apart than high-frequency channels. As shown in Fig. 17(a), the measured gain curve before the calibration shows the inter-channel gain deviations that are caused by systematic mismatches from the SRO-based PFM encoder. The main cause of the gain deviation is the voltage bias (V_{VAR} in Fig. 11), which is generated from a single FLL circuit, and it is shared over the 16-channel SRO, as depicted as "FLL (SRO)" in the chip photograph. We expect that this systematic mismatch due to the distribution of the voltage bias can be improved with a better layout floorplan, for example, with a centralized placement of the bias circuits, while the random mismatch can be improved with larger sizing of the biasing transistors.

Fig. 17(c) shows the measured output spectrum of channel 8 for two different input conditions of the VTC; the black curve is obtained with a zero input condition, while the red curve is obtained with a 2-kHz sinusoidal input of 390 mV_{PP}. Here, the amplitude of the input to the VTC circuit was assumed to be sufficiently large, and our future work will include an additional ultra-low-power pre-amplifier [33] before the VTC

[1] http://download.tensorflow.org/data/speech_commands_test_set_v0.02.tar.gz

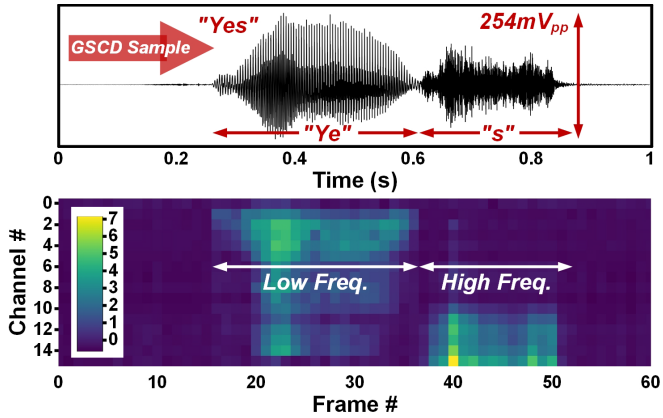


Fig. 18. Measured audio response of the FEx with an applied sample keyword from GSCD.

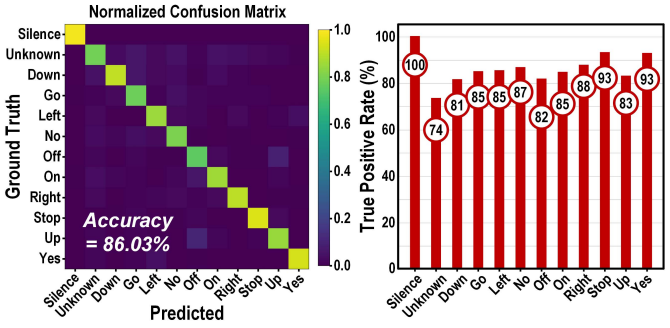


Fig. 19. Measured KWS accuracy on the GSCD test set. A confusion matrix (left), where the magnitudes are normalized between 0 and 1, and a plot of the true positive rates over 12 different classes (right) are shown.

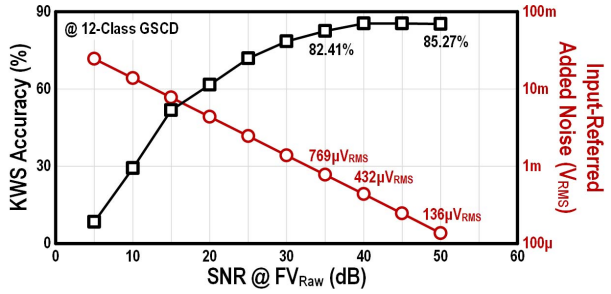


Fig. 20. KWS accuracy obtained over different SNR levels.

circuit. The oversampling clock frequency that is fed into the XOR differentiator is 62.5 kHz. It is clearly seen that the output spectrum has a first-order noise-shaping property with a 20-dB/dec slope for both input conditions. After the feature data are decimated by 2^{10}, the in-band frequency is limited to 30 Hz, which is translated into a 16-ms frame shift or 61-frame/s throughput, and so the 16-channel FV is generated every 16 ms. The integrated in-band noise with zero input is calculated as 248 μV$_{RMS}$, which is dominated by $1/f$ noise. When the input amplitude is increased to 390 mV$_{PP}$, the in-band noise is dominated by thermal noise. We believe that the noise increase is caused by a higher running frequency of the SRO since the phase noise of ring oscillators increases with operating frequency [34].

Fig. 18 shows the measured audio response of the FEx. A 254-mV$_{PP}$ "Yes" keyword sample from GSCD is selected and applied to the VTC while measuring the FEx output.

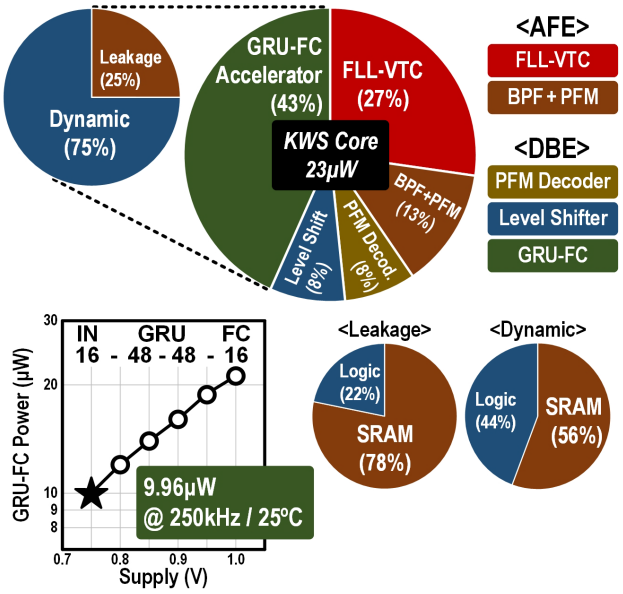


Fig. 21. Power breakdown of the KWS core implemented in the proposed IC (see Fig. 3).

The magnitudes of FV in this figure are normalized by subtracting dc offset and dividing by the standard deviation of the sample clip for better visualization. It is clearly seen that the 16-channel FV has a higher response at low frequencies for the "Ye" sound and at higher frequencies for the "s" sound.

Fig. 19 shows the measured KWS accuracy of the prototype chip obtained using the full 12-class verification flow of the GSCDv2 [35]. The 12 classes include "Silence," "Unknown," and ten target keywords. As shown in the measurement setup in Fig. 16, the generated FVs (FV_{Raw} in Fig. 3) from our time-domain FEx is recorded using the GSCD training set that is fed to the VTC of our chip ($V_{IN,VTC}$ in Fig. 16) to train the classifier network. The 16-ms frame window and the 16-ms frame shift (stride) are used for recording, so there is no overlap between two consecutive frames. The detected class is the most active output at the end of the GSCD sample. The prototype KWS IC achieves an overall 86.03% accuracy with the GSCD test set. The measured true positive rates show that "Silence" is the easiest class with 100% accuracy, and the classifier performed the best on two keywords, "Stop" and "Yes," with 93% accuracy. The most challenging class is the "Unknown" class since it includes 25 non-target keywords, such as "Happy" and "Dog," which requires the classifier to train more parameters with a larger model capacity to improve the accuracy. We expect that the detection accuracy of "Unknown" and, thus, the overall accuracy on this KWS dataset will improve with a larger network model but at the expense of additional power consumption and silicon area. The state-of-the-art accuracy on GSCD using GRU-RNNs is 94.2% [36] with a network, which has 499-kB parameters, and running on a Cortex-M7 microcontroller. This network size would require 21× more on-chip memory, leading to higher power consumption and chip area.

Fig. 20 shows the dependence of the KWS classification accuracy on added noise levels to the recorded feature vector FV_{Raw} (see Fig. 3). We first computed the average

TABLE I
PERFORMANCE COMPARISON TABLE—ANALOG FEX

Analog FEx	M. Yang JSSC 2016 [37]	K. Badami JSSC 2016 [7]	M. Yang JSSC 2019 [8]	S. Oh JSSC 2019 [3]	**This Work**
Process (nm)	180	90	180	180	**65**
Area/Ch. (mm^2)	0.26	0.13	0.1	-	**0.1**
Architecture	g_mC-BPF	g_mC-BPF	g_mC-BPF	Mixer	**OSC-BPF**
Number of Ch.	64×2	16	16	32[A]	**16**
Freq. Range (Hz)	8-20k	75-5k	100-5k	75-4k	**111-10.4k**
Supply (V)	0.5	-	0.6	1.4	**0.5**
Power (μW)	55	6	0.38	0.06	**9.3**
Frame Shift (ms)	-	31.25[B]	10	512[A]	**16**
Dynamic Range (dB)	55	45	40[C]	47	**54.89**
$FoM_{S,DR}$ (dB)	-	82.3	91.5	91.33	**93.11**
Target Task	General Purpose	VAD			**KWS**
Building Blocks	BPF, ADM	LNA, BPF FWR, LPF	LNA, BPF FWR, IAF	LNA, Mixer LPF, DSP	**VTC Rec-BPF PFM**
Support SE Mic	No	✓	No	✓	✓
Parallel FEx	✓	✓	✓	No	✓

[A]With 32-32-16-2 FC neural network [B]f_{LPF} = 16 Hz
[C]Measured by the firing rate range of the IAF only, excluding output noise of the feature vector

$$\mathrm{P_{Norm}} = \frac{P(1-r)}{1-r^n} \cdot \frac{20\mathrm{k}}{f_H} \qquad r = \left(\frac{f_L}{f_H}\right)^{1/(n-1)} \tag{7}$$

$$\mathrm{FoM_{S,DR}} = \mathrm{DR} + 10 \cdot \log_{10}\left(\frac{1}{\mathrm{P_{Norm}} \cdot 2 \cdot \mathrm{FrameShift}}\right) \tag{8}$$

power $P_{\mathrm{Avg,GSCD}}$ using the recorded $\mathrm{FV_{Raw}}$ (see Section III-F). Then, Gaussian noises of different standard deviation values ($\sigma^2 = P_{\mathrm{Avg,Noise}}$) were added to create different SNR values following the equation below:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_{\mathrm{Avg,GSCD}}}{P_{\mathrm{Avg,Noise}}}\right). \tag{9}$$

The noise is randomly generated for each training epoch and test evaluation. For each SNR case, our GRU-FC network is retrained using the noisy training set, and the noisy test set is used to evaluate the classification accuracy. The proposed KWS IC ensures <1% accuracy drop even for noise levels up to 432 $\mu\mathrm{V_{RMS}}$ (input-referred to $V_{\mathrm{IN,VTC}}$) or 40-dB SNR.

Fig. 21 shows the power breakdown of the KWS IC. As stated in Section I, this article focuses on the KWS core; only, therefore, the power breakdown in Fig. 21 does not include an energy harvester, low-dropout regulator, and voltage reference circuits. The total power consumption of the KWS core is 23 μW when it is measured at 25 °C room temperature. The GRU-FC neural network accelerator accounts for 43% of the KWS core power. When the 16IN-48H-48H-12C GRU-FC network is updated at 250-kHz clock frequency and 0.75-V supply voltage while performing continual inference on random GSCD samples with 16-ms frame shift, the accelerator consumes 9.96 μW. The accelerator power consumption can be further decomposed into dynamic power (75%) and leakage (or static) power (25%). The leakage power is dominated by the SRAM block (78%), while both logic (44%) and SRAM (56%) contributed rather evenly to the dynamic power. We expect that leakage power can be reduced with custom memory cells [6].

Table I compares the performance of our time-domain FEx with the state-of-the-art voice processing analog FEx [3], [7], [8], [37]. The proposed FEx circuit is the first that demonstrated the ring-oscillator-based BPF topology used for the KWS task. It supports SE microphones, thereby offering a lower system-level power. Unlike sequential FEx, our parallel FEx does not lose frequency-selective information at any time [3]. To allow a fair comparison with previously reported designs with a variety of frame shifts, we derive a Schreier Figure of Merit (FoM) [41] (8), widely used for ADCs. The Schreier FoM considers the tradeoff between DR and bandwidth, also accounting for power consumption. For near dc input ADCs, the bandwidth is replaced with a reciprocal of conversion time [42]. As a bandpass filtered signal is demodulated into baseband (dc) after the rectifier [7], [8] or mixer [3] stage, we consider analog FEx as a dc-input ADC with a pre-processing stage. The FoM equation (8) includes the normalized power consumption P_{Norm} (7) proposed in [37], the DR, and the frame shift. The frame shift is part of the denominator of (8) because the FVs are generated in every frame shift. The amount of integrated in-band noise is reduced with a larger decimation window (i.e., averaged over the longer time interval) in our design and also in [42] where the number of ADC cycles was used for decimation window. The proposed FEx records the best Schreier FoM among the state-of-the-art designs. In addition, the time-domain processing circuits offer better technology scaling and will outperform voltage-domain designs [3], [7], [8], [37] in terms of power and area when implemented in advanced technology nodes.

Table II compares the performance of our KWS IC with other state-of-the-art KWS ICs [2], [6], [9], [38], [39]. This work uses an on-chip analog FEx, while other works needed an off-chip high-resolution (16-bit) ADC [6], [38]. Sometimes, even the digital FEx and ADC were implemented off-chip [39].

TABLE II
PERFORMANCE COMPARISON TABLE—KWS

KWS	S. Zheng TCAS-I 2019 [38]	H. Dbouk JSSC 2021 [39]	W. Shan JSSC 2021 [6]		J. Giraldo VLSI 2019 [2]	D. Wang ISSCC 2021 [9]	**This Work**
-	Off-Chip ADC				On-Chip ADC	On-Chip Analog FEx	
Process (nm)	28	65	28		65	65	**65**
Area (mm^2)	1.29[A]	4.13[A]	0.23[A]		1.52	2.71[B]	**2.03**
SRAM (KB)	52	38	2		32	20	**27**
Clock (Hz)	2.5M	1G	40k		250k	120k	**250k**
FEx	Digital	-	Digital		Digital	Analog Voltage	**Analog Time**
Classifier	CNN	RNN	CNN		RNN	SNN (MLP)	**RNN**
KWS Power (μW)	141[A]	11000[A]	0.51[A]		16.1	0.205-0.570	**23**
Frame Shift (ms)	10	20	16		16	100	**16**
Latency (ms)	10	0.04	64		16	100	**12.4**
Dataset	TIDIGITS	GSCD					
Number of Classes (Keywords)	2	7[C] (6)	2[C] (1)	5[C] (4)	12[D] (10)	5[C] (4)	**12[D] (10)**
Accuracy (%)	96	90.38	97.3[E]	91.7[E]	90.87	90.2	**86.03**
Support SE Mic	Off-Chip ADC				No	No	✓

[A]Excluding off-chip ADC [B]Including SNN chip [40] [C]Excluding "Unknown" word detection as a distinct class
[D]2 non-keywords (Silience/Unknown) + 10 keywords [E]Accuracy is reported from 16-bit GSCD samples; the design excludes the 16-bit ADC

Furthermore, only this work and [2] support the essential "Unknown" class to be detected as a distinct class, which is the most challenging class in the GSCD test set. As such, it implies that concessions would be made in terms of KWS accuracy for [6], [9], and [39], or a larger model size will be required for the classifier to uphold the accuracy, leading to additional power and area costs. In addition, the proposed chip supports SE microphone interface, and the KWS task is verified with an SE input condition. This work shows competitive performance and better system-level power efficiency by using a low-power MEMS SE microphone instead of the differential microphone used in [2]. Last but not least, our prototype chip is the first silicon-verified analog FEx-based voice processing IC that demonstrates 12-class KWS task on GSCD, using an on-chip classifier.

Our belief is that the 5% degradation in the classification accuracy of our KWS IC (86%) compared to the software model accuracy (91%; see Section II) is mainly due to the increased noise floor when the input amplitude is high, as shown in Fig. 17(c). Advanced noise suppression techniques, such as chopper stabilization [43] and dynamic element matching [5] when applied to the front end, will help mitigate the accuracy discrepancy. Our time-domain FEx still needs to address the per-chip gain calibration requirement due to the mismatch of analog circuits, which is not necessary for a fully digital approach [2]. For this, as discussed in the paragraph in Section IV describing Fig. 17(a), improved layout floorplan and larger device sizes accompanied with mismatch-aware DNN training [8] will be another opportunity to remove the calibration requirement.

V. CONCLUSION

We have presented a low-power KWS chip that exploits ring-oscillator-based time-domain processing circuits. Implemented in a 65-nm CMOS process, it consumes 23-μW power dissipation with a power supply of 0.5 V for analog circuits and 0.75 V for digital circuits. The nested analog FLL enhances the linearity of VTC and, thus, facilitates the use of SE microphones, as discussed in Section III-A. The usage of PFD as a time-domain FWR shows significantly reduced implementation cost in comparison with a voltage-domain design. The PFM functionality is realized using an SRO, instead of the conventional IAF circuit to obviate the need for scaling-unfriendly voltage-domain circuits. Table I shows that the proposed time-domain FEx achieves the state-of-the-art DR-based Schreier FoM. The on-chip integrated GRU-FC digital back-end circuit processes incoming audio FVs with a 16-ms frame shift using only $\sim$10-μW power, demonstrating $>$86% classification accuracy with only 12.4-ms latency on the 12-class GSCD KWS task. We expect that the proposed time-domain processing techniques can be further expanded in other domains and, thus, provide various design opportunities for power-efficient circuits, such as the fully time-domain ReLU activation unit shown in Fig. 12 for DNNs. The improvement directions, as discussed in Section IV, along with DR enhancement techniques, such as front-end automatic gain control, will enable the time-domain FEx to be applied to more challenging real-world audio-inference tasks. A 35-class KWS on GSCD [44] or a streaming-mode KWS can be such examples.

ACKNOWLEDGMENT

The authors would like to thank Frank K. Gürkaynak and Beat Muheim from ETH Zürich for their technical support of the digital circuits in this IC technology and Taekwang Jang from ETH Zürich for the valuable discussions on analog frequency-locked loop.

REFERENCES

[1] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[2] J. S. P. Giraldo, S. Lauwereins, K. Badami, H. Van Hamme, and M. Verhelst, "18 μW SoC for near-microphone keyword spotting and speaker verification," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2019, pp. C52–C53.

[3] S. Oh *et al.*, "An acoustic signal processing chip with 142-nW voice activity detection using mixer-based sequential frequency scanning and neural network classification," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3005–3016, Nov. 2019.

[4] K. Kim, J.-H. Kim, S. Gweon, M. Kim, and H.-J. Yoo, "A 0.5-V sub-10-μW 15.28-m$\Omega/\sqrt{\text{Hz}}$ bio-impedance sensor IC with sub-1° phase error," *IEEE J. Solid-State Circuits*, vol. 55, no. 8, pp. 2161–2173, Aug. 2020.

[5] H. Ha *et al.*, "A bio-impedance readout IC with digital-assisted baseline cancellation for two-electrode measurement," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 2969–2979, Nov. 2019.

[6] W. Shan *et al.*, "A 510-nW wake-up keyword-spotting chip using serial-FFT-based MFCC and binarized depthwise separable CNN in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 151–164, Jan. 2021.

[7] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 μW power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan. 2016.

[8] 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 J. Solid-State Circuits*, vol. 54, no. 6, pp. 1764–1777, Jun. 2019.

[9] D. Wang, S. J. Kim, M. Yang, A. A. Lazar, and M. Seok, "A background-noise and process-variation-tolerant 109 nW acoustic feature extractor based on spike-domain divisive-energy normalization for an always-on keyword spotting device," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 160–162.

[10] B. Drost, M. Talegaonkar, and P. K. Hanumolu, "Analog filter design using ring oscillator integrators," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, pp. 3120–3129, Dec. 2012.

[11] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[12] E. Gutierrez, C. Perez, F. Hernandez, and L. Hernandez, "VCO-based feature extraction architecture for low power speech recognition applications," in *Proc. IEEE 62nd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2019, pp. 1175–1178.

[13] N. Goux, J.-B. Casanova, G. Pillonnet, and F. Badets, "A 6-nW 0.0013-mm^2 ILO bandpass filter for time-based feature extraction," *IEEE Solid-State Circuits Lett.*, vol. 3, pp. 306–309, 2020.

[14] K. Kim *et al.*, "A 23 μW solar-powered keyword-spotting ASIC with ring-oscillator-based time-domain feature extraction," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 370–371.

[15] K. Kim and S.-C. Liu, "Continuous-time analog filters for audio edge intelligence: Review and analysis on design techniques," 2022, *arXiv:2206.02639*.

[16] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-36, no. 7, pp. 1119–1134, Jul. 1988.

[17] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[18] A. Elshazly, S. Rao, B. Young, and P. K. Hanumolu, "A noise-shaping time-to-digital converter using switched-ring oscillators—Analysis, design, and measurement techniques," *IEEE J. Solid-State Circuits*, vol. 49, no. 5, pp. 1184–1197, May 2014.

[19] *Ultra-Low Current, Low-Noise Microphone With Analog Output, ICS-40310 Datasheet*, InvenSense, San Jose, CA, USA, Dec. 2014.

[20] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

[21] D. Djekic, G. Fantner, K. Lips, M. Ortmanns, and J. Anders, "A 0.1% THD, 1-MΩ to 1-GΩ tunable, temperature-compensated transimpedance amplifier using a multi-element pseudo-resistor," *IEEE J. Solid-State Circuits*, vol. 53, no. 7, pp. 1913–1923, Jul. 2018.

[22] T. Jang, S. Jeong, D. Jeon, K. D. Choo, D. Sylvester, and D. Blaauw, "A noise reconfigurable all-digital phase-locked loop using a switched capacitor-based frequency-locked loop and a noise detector," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 50–65, Jan. 2018.

[23] M. Seok, G. Kim, D. Blaauw, and D. Sylvester, "A portable 2-transistor picowatt temperature-compensated voltage reference operating at 0.5 V," *IEEE J. Solid-State Circuits*, vol. 47, no. 10, pp. 2534–2545, Oct. 2012.

[24] W. Zhao *et al.*, "A 0.025-mm^2 0.8-V 78.5-dB SNDR VCO-based sensor readout circuit in a hybrid PLL-$\Delta\Sigma$M structure," *IEEE J. Solid-State Circuits*, vol. 55, no. 3, pp. 666–679, Mar. 2020.

[25] J. Tow, "Active RC filters—A state-space realization," *Proc. IEEE*, vol. 56, no. 6, pp. 1137–1139, Jun. 1968.

[26] L. Thomas, "The biquad: Part I-some practical design considerations," *IEEE Trans. Circuit Theory*, vol. CT-18, no. 3, pp. 350–357, May 1971.

[27] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA, USA: Addison Wesley, 1989.

[28] A. A. Abidi and R. G. Meyer, "Noise in relaxation oscillators," *IEEE J. Solid-State Circuits*, vol. SSC-18, no. 6, pp. 794–802, Dec. 1983.

[29] A. Iwata, N. Sakimura, M. Nagata, and T. Morie, "The architecture of delta sigma analog-to-digital converters using a voltage-controlled oscillator as a multibit quantizer," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 46, no. 7, pp. 941–945, Jul. 1999.

[30] M. Z. Straayer and M. H. Perrott, "A 12-bit, 10-MHz bandwidth, continuous-time $\Sigma\Delta$ ADC with a 5-bit, 950-MS/s VCO-based quantizer," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 805–814, Apr. 2008.

[31] J. Kim, T. K. Jang, Y. G. Yoon, and S. Cho, "Analysis and design of voltage-controlled oscillator based analog-to-digital converter," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 1, pp. 18–30, Jan. 2010.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–19.

[33] D. Han, Y. Zheng, R. Rajkumar, G. S. Dawe, and M. Je, "A 0.45 V 100-channel neural-recording IC with sub-μW/channel consumption in 0.18 μm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 735–746, Dec. 2013.

[34] A. A. Abidi, "Phase noise and jitter in CMOS ring oscillators," *IEEE J. Solid-State Circuits*, vol. 41, no. 8, pp. 1803–1816, Aug. 2006.

[35] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[36] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," 2017, *arXiv:1711.07128*.

[37] M. Yang, C.-H. Chien, T. Delbruck, and S.-C. Liu, "A 0.5 V 55 μW 64×2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2554–2569, Nov. 2016.

[38] S. Zheng *et al.*, "An ultra-low power binarized convolutional neural network-based speech recognition processor with on-chip self-learning," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 12, pp. 4648–4661, Dec. 2019.

[39] H. Dbouk, S. K. Gonugondla, C. Sakr, and N. R. Shanbhag, "A 0.44-μJ/dec, 39.9-μs/dec, recurrent attention in-memory processor for keyword spotting," *IEEE J. Solid-State Circuits*, vol. 56, no. 7, pp. 2234–2244, Jul. 2021.

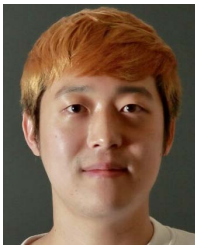
[40] D. Wang *et al.*, "Always-on, sub-300-nW, event-driven spiking neural network based on spike-driven clock-generation and clock-and power-gating for an ultra-low-power intelligent device," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2020, pp. 1–4.

[41] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*, vol. 74. Piscataway, NJ, USA: IEEE Press, 2005.

[42] Y. Chae, K. Souri, and K. A. A. Makinwa, "A 6.3 μW 20 bit incremental zoom-ADC with 6 ppm INL and 1 μV offset," *IEEE J. Solid-State Circuits*, vol. 48, no. 12, pp. 3019–3027, Dec. 2013.

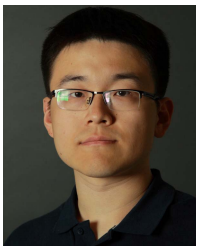
[43] C. C. Enz and G. C. Temes, "Circuit techniques for reducing the effects of op-amp imperfections: Autozeroing, correlated double sampling, and chopper stabilization," *Proc. IEEE*, vol. 84, no. 11, pp. 1584–1614, Nov. 1996.

[44] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming keyword spotting on mobile devices," in *Proc. Interspeech*, Oct. 2020, pp. 1–5.



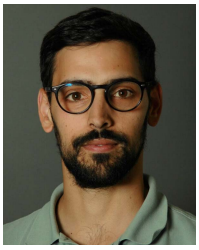
Kwantae Kim (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015, 2017, and 2021, respectively.

From 2015 to 2017, he was with the Healthrian R&D Center, Daejeon, where he designed bio-potential readout IC for mobile healthcare solutions. In 2020, he was a Visiting Student with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, where he has been a Post-Doctoral Researcher since 2021. His research interests include analog/mixed-signal ICs for time-domain processing, in-memory computing, bio-impedance sensor, and neuromorphic audio sensor.



Chang Gao (Member, IEEE) received the B.Eng. degree in electronics from the University of Liverpool, Liverpool, U.K., and Xi'an Jiaotong-Liverpool University, Suzhou, China, in July 2015, the master's degree in analog and digital integrated circuit design from Imperial College London, London, U.K., in November 2016, and the Ph.D. degree from the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, in December 2021.

In August 2022, he joined the Delft University of Technology (TU Delft), Delft, The Netherlands, as an Assistant Professor in digital design. His current research interest includes designing energy-efficient AI hardware for edge computing.



Rui Graça received the B.Sc. and M.Sc. degrees in electrical and computer engineering from the Faculty of Engineering, University of Porto, Porto, Portugal, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland.

From 2016 to 2019, he worked as an Analog Design Engineer at Synopsys, Porto, Portugal, where he worked on the design and verification of mixed-signal circuits for high-speed SerDes. His current research interests include neuromorphic engineering and low-noise design for event-based sensors.

Ilya Kiselev (Member, IEEE) received the Specialist degree in physics from Tambov State University, Tambov, Russia, in 2000, the M.Sc. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Dolgoprudny, Russia, in 2002, and the Ph.D. degree from ETH Zürich, Zürich, Switzerland, in 2021.

He is currently doing his post-doctoral work at the Institute of Neuroinformatics, University of Zürich and ETH Zürich. His research interests include hardware implementations of signal acquisition and processing for traditional and event-based audio processing.

Hoi-Jun Yoo (Fellow, IEEE) graduated from the Department of Electronics, Seoul National University, Seoul, South Korea, in 1983. He received the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1985 and 1988, respectively.

Dr. Yoo has served as a member of the Executive Committee for the International Solid-State Circuits Conference (ISSCC), the Symposium on Very Large-Scale Integration (VLSI), and the Asian Solid-State Circuits Conference (A-SSCC), the TPC Chair of the A-SSCC 2008 and the International Symposium on Wearable Computer (ISWC) 2010, the IEEE Distinguished Lecturer from 2010 to 2011, the Far East Chair of the ISSCC from 2011 to 2012, the Technology Direction Sub-Committee Chair of the ISSCC in 2013, the TPC Vice-Chair of the ISSCC in 2014, and the TPC Chair of the ISSCC in 2015. More details are available at http://ssl.kaist.ac.kr.

Tobi Delbruck (Fellow, IEEE) received the B.A. degree in physics and applied mathematics from the University of California at San Diego, San Diego, CA, USA, in 1983, and the Ph.D. degree in computation and neural systems from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1993.

Since 1998, he has been with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, where he is currently a Professor of physics and electrical engineering. The Sensors Group, which he co-directs with Shih-Chii Liu, currently focuses on neuromorphic sensory processing, control, and efficient hardware artificial intelligence (AI).

Shih-Chii Liu (Fellow, IEEE) received the B.Sc. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1983, and the Ph.D. degree in the computation and neural systems program from the California Institute of Technology, Pasadena, CA, USA, in 1997.

She is currently a Professor with the University of Zürich, Zürich, Switzerland. Her group focuses on audio sensors, in particular, the spiking cochlea and bio-inspired deep neural network algorithms and hardware.