An 8.62-μW 75-dB DR_{SoC} Fully Integrated SoC for Spoken Language Understanding

Zhou, Sheng; Li, Zixiao; Cheng, Longbiao; Hadorn, Jerome; Gao, Chang; Chen, Qinyu; Delbruck, Tobi; Kim, Kwantae; Liu, Shih Chii

DOI
10.1109/JSSC.2025.3602936

Publication date
2025

Document Version
Final published version

Published in
IEEE Journal of Solid-State Circuits

An 8.62-μW 75-dB DR_{SoC} Fully Integrated SoC for Spoken Language Understanding

Sheng Zhou, *Graduate Student Member, IEEE*, Zixiao Li, *Graduate Student Member, IEEE*, Longbiao Cheng, *Member, IEEE*, Jérôme Hadorn, Chang Gao, *Member, IEEE*, Qinyu Chen, *Member, IEEE*, Tobi Delbruck, *Fellow, IEEE*, Kwantae Kim, *Senior Member, IEEE*, and Shih-Chii Liu, *Fellow, IEEE*

***Abstract*—We present a sub-10-μW fully integrated SoC for on-device spoken language understanding (SLU). Its analog feature extractor (FEx) applies global and per-channel automatic gain control (AGC) to extend the system's dynamic range (DR)—a critical requirement for real-world scenarios, including far-field operations. The on-chip streaming-mode recurrent neural network (RNN) accelerator exploits temporal sparsity and pooling, reducing its power by 2.3×. By combining hardware-aware training with a behavioral model of the FEx that captures circuit nonidealities, the network is trained to maintain SLU accuracy despite chip-to-chip variation. Fabricated in a 65-nm CMOS process, the SoC occupies 2.23 mm^2 and consumes 8.62 μW for end-to-end SLU. The 16-channel FEx achieves 93-dB DR while dissipating 1.85 μW at 100-Hz feature frame rate. The SoC is evaluated on the 32-class Fluent Speech Commands dataset (FSCD), achieving 92.9% accuracy for 2.8-mV$_{rms}$ inputs while maintaining >85% accuracy over a 75-dB input range.**

***Index Terms*—Automatic gain control (AGC), edge artificial intelligence (AI), feature extractor (FEx), hardware–software co-design, recurrent neural network (RNN), spoken language understanding (SLU), tiny machine learning (TinyML), ultra-low power, voice interface.**

I. Introduction

THE recent trend of edge artificial intelligence (AI) designs sees an increasing number of functionalities being integrated into Internet of Things (IoT) nodes and wearable devices for applications such as always-on health monitoring and smart home automation [1]. Of these functionalities, adding a voice interface to edge devices is useful because it provides a hands-free means of user interaction [2]. However, the limited system power budget of these battery-powered devices poses a challenge for adding new features.

Received 8 May 2025; revised 26 July 2025; accepted 17 August 2025. Date of publication 11 September 2025; date of current version 29 October 2025. This article was approved by Associate Editor Jacques Christophe Rudell. This work was supported in part by the Swiss National Science Foundation Compute-aware Deep Neural Networks on the Edge (CA-DNNEdge) Project under Grant 208227 and in part by the Scientific dynamic vision sensor event camera (SCIDVS) Project under Grant 185069. *(Sheng Zhou and Zixiao Li contributed equally to this work.) (Corresponding author: Shih-Chii Liu.)*

Sheng Zhou, Zixiao Li, Longbiao Cheng, Jérôme Hadorn, 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).

Chang Gao is with the Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands.

Qinyu Chen is with Leiden Institute of Advanced Computer Science (LIACS), Leiden University, 2333 CC Leiden, The Netherlands.

Kwantae Kim is with the Department of Electronics and Nanoengineering, School of Electrical Engineering, Aalto University, 02150 Espoo, Finland.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2025.3602936.

Digital Object Identifier 10.1109/JSSC.2025.3602936

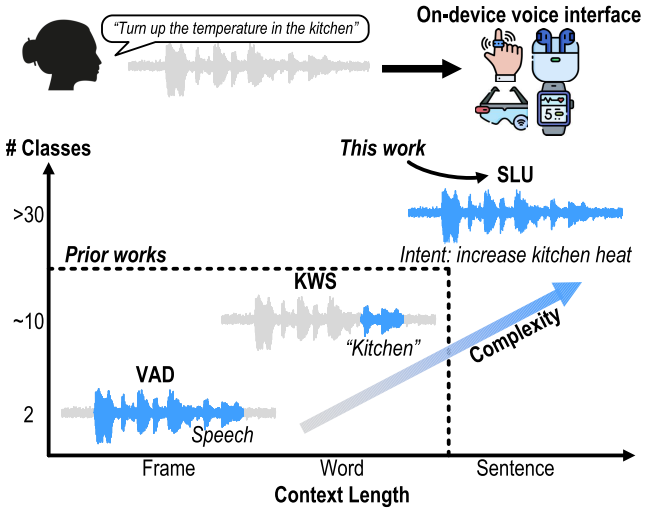


Fig. 1. Audio classification tasks for on-device voice interface.

A typical smart ring, for example, has a battery capacity of 20 mAh and a target battery life of one week, setting a total system power budget of less than 450 μW. The stringent power requirement has led to growing research interest in developing application-specific integrated circuits (ASICs) for on-device voice interface [3], [4], [5]. Fig. 1 shows typical lightweight audio classification tasks for an on-device voice interface. voice activity detection (VAD) designs [6], [7], [8], [9], [10], [11], [12], [13], [14] distinguish speech from non-speech using frame-level context of tens of milliseconds. The more complex keyword spotting (KWS) designs [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34] detect one or a few keywords using word-level context of hundreds of milliseconds. Both VAD and KWS can serve as wake-up stages to activate the more expensive and power-consuming on- or off-chip speech processor. Our recent always-on spoken language understanding (SLU) design [5] aims to infer user intent directly from continuous speech. It utilizes sentence-level context of several seconds to support the detection of more than 30 user intents from an SLU dataset [35]. In addition, end-to-end audio-to-intent classification eliminates the use of more power-consuming general-purpose automatic speech recognition (ASR) processors [36], [37], [38] and natural language understanding (NLU) engines [39], [40] for low-power edge applications.

Fig. 2 depicts the major processing stages of an on-device SLU system. The microphone output is fed to a feature

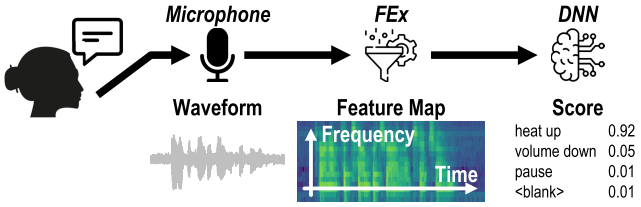


Fig. 2. Processing stages for an on-device SLU system.

extractor (FEx), which produces a feature map representing the time-varying spectral components. A deep neural network (DNN) classifier then processes the feature map to produce the score for each intent. Both the FEx and the DNN classifier pose several design challenges for realizing an ultra-low-power system-on-chip (SoC) for SLU, as explained next.

A conventional digital FEx consists of a cascade of amplifiers, an analog-to-digital converter (ADC), and a digital signal processor (DSP). The digital waveform sampled at Nyquist rate (typically 8–16 kHz) is processed by the DSP using fast Fourier transform (FFT) and frequency-domain filtering to obtain the spectrogram features. However, due to the relatively high ADC sampling rate and large number of DSP operations, the digital FEx accounted for more than 50% of the SoC power in prior designs [21], [27]. For speech understanding, the typical audio feature frame shift is between 10 and 20 ms [21], [22], corresponding to a frame rate of 50–100 Hz, which is much lower than the sampling rate of the raw waveform. Therefore, an analog FEx with direct analog-to-feature conversion can be employed to reduce feature extraction power [6], [7], [8], [9], [15], [16], [17], [18], [19]. However, due to the analog circuit nonidealities such as transistor mismatch, electronic noise, and large-signal distortion, prior designs only demonstrated VAD or KWS on a limited (<15) number of classes and sometimes required a costly off-chip DNN classifier to compensate the FEx nonidealities [18], [19]. In addition, real-world applications require the FEx to have an input dynamic range (DR) of at least 60 dB [41] due to the varying speech volume and speaker distance to the microphone. This large input DR has not been achieved yet by previously reported analog FEx.

SLU requires a DNN capable of processing long, variable-length contexts. In addition, since the start and end of a spoken sentence are unknown a priori, the model should operate in streaming mode to process input continuously. Convolutional neural network (CNN), widely used for KWS [32], [34], have a fixed context length and require extra activation memory for processing long contexts. The Transformer [42] is a powerful architecture for sequence modeling, but its self-attention module is not suitable for streaming operation. In contrast, gated recurrent neural network (RNN) using long short-term memory (LSTM) [43] or gated recurrent unit (GRU) [44] encodes an input temporal sequence into fixed-size hidden states, enabling them to process a theoretically unbounded context window while supporting streaming input. Yet, realizing ultra-low-power RNN is difficult due to the memory-intensive matrix-vector multiplication (MVM) and limited activation sparsity [45]. Therefore, algorithm-hardware co-design is required to reduce energy consumption while retaining SLU accuracy.

To address the aforementioned design challenges, in this work, we describe the first fully integrated SoC for on-device SLU, extending upon our conference paper [5]. The mixed-signal ASIC demonstrates sub-10-μW end-to-end user intent understanding with continuous speech input, achieving 92.9% accuracy for 32-class SLU on the Fluent Speech Commands dataset (FSCD) [35]. It also maintains >85% accuracy over a 75-dB input range. The SoC interfaces directly with an ultra-low-power single-ended (SE) micro-electromechanical systems (MEMS) microphone [46], simplifying system integration. Our design is enabled by the following technical contributions.

1) An analog FEx with integrated global and per-channel automatic gain control (AGC). It achieves the highest reported DR and state-of-the-art Schreier figure of merit (FoM) [17], [19].
2) A Python-based behavioral model of the FEx that incorporates analog circuit nonidealities. By combining the FEx behavioral model with hardware-aware training (HAT), our SLU model is trained without costly chip-in-the-loop training.
3) A temporal-sparsity-aware RNN accelerator operating in streaming mode. By exploiting fine-grained temporal sparsity and temporal pooling, digital power is reduced by 2.3×.

The rest of this article is organized as follows. Section II describes the SoC's architectural design, encompassing the analog FEx and its behavioral model, as well as the algorithmic optimization for the RNN. Section III details the circuit implementation of the SoC building blocks. Section IV presents the measurement results for both the building blocks and the final end-to-end system. Section V concludes this article.

II. SoC Architecture and Algorithmic Optimization

Fig. 3 illustrates the overall architecture of the SLU SoC, consisting of an analog front end (AFE) and a digital backend (DBE). The AFE is driven by an off-chip SE microphone and directly converts this analog input into 16-channel digital features (D_{AFE}) using an analog FEx (see Section II-A). The per-channel features represent the instantaneous log amplitude within a particular frequency band. The feature frame rate is 100 Hz and each D_{AFE} value consists of 8 bits. An FEx behavioral model (see Section II-B) is developed in Python for generating features for network training. It also models the nonidealities of the FEx. Together with HAT, the SLU network is trained to maintain SLU accuracy and to withstand chip-to-chip variations without using costly chip-in-the-loop training. The DBE sequentially processes the incoming stream of feature frames and produces an output stream of per-class probabilities. It employs a temporal-sparsity-aware Δ-GRU network (see Section II-C) and a training pipeline tailored for streaming-mode operation (see Section II-D).

A. Analog FEx

Speech understanding in natural environments requires a wide DR of more than 60 dB [41], which poses a significant

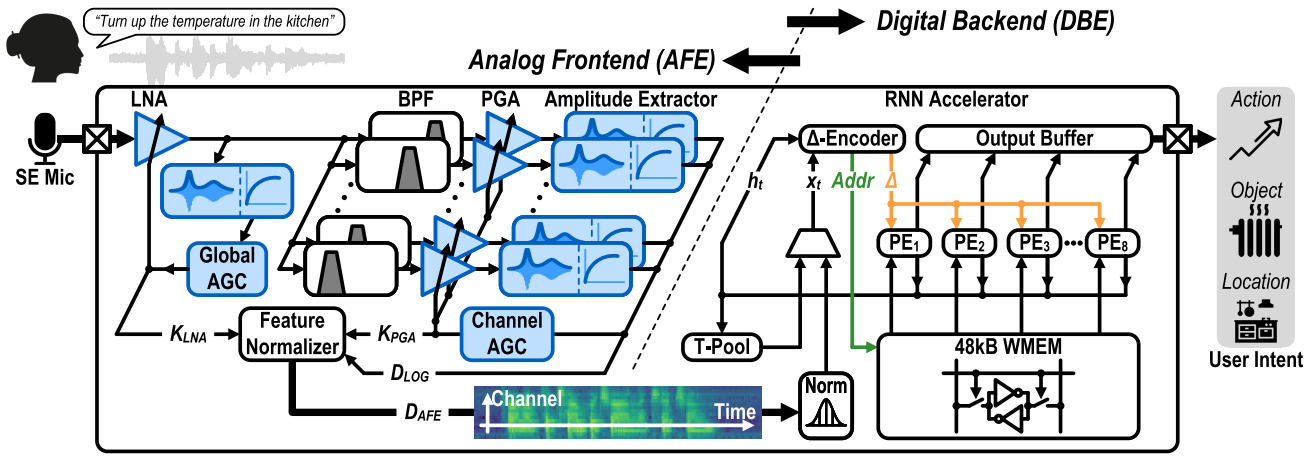


Fig. 3. Architecture of the sub-10-μW mixed-signal SLU SoC, including an AFE and a DBE.

challenge to realize low-power FEx. Lookup table (LUT)-based logarithmic compression has been applied in prior designs [17], [27] to allow the features to encode a wider range using the same number of bits. However, the FEx's DR is still limited by the analog circuits preceding the logarithmic LUT.

The wide DR of over 120 dB in human hearing is enabled through two AGC mechanisms [47]: first, global AGC by the acoustic reflex of the middle ear, and second, local AGC by the outer hair cells of the biological cochlea. Global AGC acts on all spectral components of sound, while local AGC acts on individual frequency bands. Previous silicon cochlea designs [48], [49], [50], [51] have implemented various AGC mechanisms with continuously tunable gains using only analog circuits, and some designs [48], [49], [50], [51] also implemented cross-channel coupling to model the lateral inhibition phenomena in the auditory system. However, these designs [48], [49], [50], [51] required high power (4.5–100 μW) per channel. More recent analog FEx designs [19], [52] adopted capacitively coupled amplifiers with discrete tunable gain levels due to their better power efficiency but relied on an off-chip digital controller.

In this work, we combine logarithmic compression with both global and per-channel digitally assisted AGC to realize a low-power FEx with a wide DR. As shown in Fig. 3, the microphone signal is first amplified by a low-noise amplifier (LNA) and then decomposed into different frequency channels through a bank of 16 bandpass filters (BPFs). The central frequencies of the BPFs are equally spaced on a logarithmic scale from 100 Hz to 8 kHz. Within each channel, a programmable-gain amplifier (PGA) further amplifies the bandpass filtered signal. Thereafter, an amplitude extractor extracts the PGA output amplitude, digitizes it every 10 ms with an ADC, and then log-compresses the ADC output via an LUT. As a result, the amplitude extractor output (D_{LOG}) represents the log amplitude of its channel. The LNA and PGA gains are set by their individual AGC feedback loops. The per-channel AGC loop reuses the amplitude extractor after the PGA, while the global AGC loop requires an additional amplitude extractor to monitor the LNA output. The feature normalizer subtracts the logarithmic gains of the LNA (K_{LNA}) and PGA (K_{PGA}) from D_{LOG}, resulting in the final features, D_{AFE}.

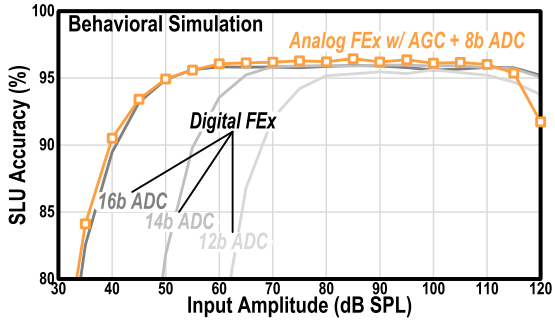


Fig. 4. Simulated SLU accuracy of the analog FEx with AGC versus conventional digital FEx with various ADC resolutions.

To compare the DR of our analog FEx with that of a conventional ADC+DSP digital FEx, we train an RNN model (see Section II-D) using simulated features from each FEx and evaluate SLU accuracy across different amplitudes. The analog FEx features are computed using the behavioral model presented in Section II-B, while the digital FEx features are obtained by quantizing the waveform samples at the selected bit resolution and applying a standard log-Mel filterbank [27]. No AFE circuit nonidealities are included in either case except for quantization. The evaluation results are shown in Fig. 4. Thanks to AGC and analog signal processing, our analog FEx achieves comparable accuracy and input DR to a digital FEx with a 16-kS/s, 16-bit ADC while using only 0.1-kS/s, 8-bit ADCs in the amplitude extractors. In terms of power consumption, a state-of-the-art 5-kHz-bandwidth 16-bit ADC [53] already consumes 4.5 μW without including the power for decimation filter and FFT engine, while our analog FEx with AGC only consumes 1.85 μW (see Section IV-A). Note that the ADC resolution is at least 16 bits [54] for digital FEx used in far-field applications, where the speaker-to-microphone distance is beyond 1–2 m [55].

B. Python-Based Behavioral Model

Analog signal processing using transistors is prone to circuit noise, large-signal distortion, and device-to-device mismatch. These nonidealities may impair the quality of the extracted features, leading to an unacceptable accuracy drop in the downstream algorithm. Reducing these circuit nonidealities

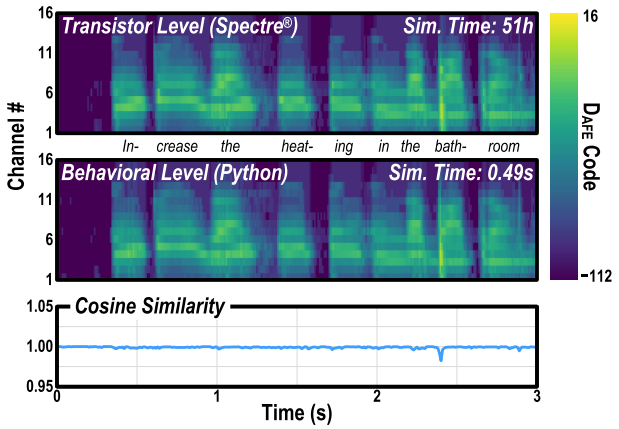


Fig. 5. Simulated features and simulation time for a speech recording using Spectre-based transistor-level model (pre-layout) versus Python-based behavioral model, and the cosine similarity between the simulated features. No circuit nonidealities were included in the shown features.

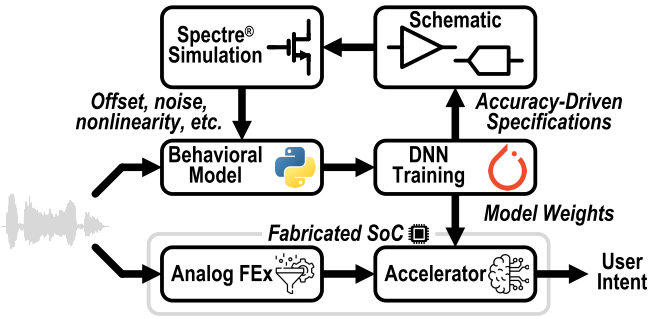


Fig. 6. Workflow for modeling analog circuit nonidealities via Python-based behavioral model and training DNN models resilient to nonidealities by HAT.

typically requires increasing the bias current, supply voltage, and/or circuit area of the amplifiers and filters. To address this power–area–accuracy tradeoff, it is therefore critical to develop a behavioral model for verifying the degree of nonidealities that can be tolerated by the design for a given task. The model can also be used to explore the extensive design space of the analog FEx.

To this end, we have developed a custom Python-based behavioral model of the analog FEx to convert dataset recordings into simulated FEx features. It employs simplified models for the circuit blocks shown in Fig. 3, including amplifiers, filters, amplitude extractors, and gain controllers. Fig. 5 shows the close similarity between simulated features from the behavioral model and those obtained from transistor-level Spectre simulation, even though its simulation time is lower by five orders of magnitude. The cosine similarity between the Spectre- and Python-simulated features, computed using the 16-D feature vectors per frame, is greater than 0.98. The Python model can be easily parallelized, allowing large-scale datasets to be converted within a reasonable time.

To address analog circuit nonidealities and avoid costly chip-in-the-loop training after fabrication, we employ the hardware–software co-design workflow shown in Fig. 6. The dataset recordings are fed to the behavioral model, which incorporates nonidealities in the signal chain. The parameters of the nonideality models [see Fig. 7(b)] are extracted from transistor-level simulation of the schematic designs using Spectre. The simulated features with nonidealities are then used to train the DNN model. Depending on the SLU accuracy of the trained model, we either relax or tighten the specifications of analog circuits (e.g., noise and offset), adjust the schematic design accordingly, and perform the circuit simulations again to extract the nonideality parameters. This iterative process is repeated until all circuit specifications are satisfied by the schematic designs and the SLU accuracy requirements are fulfilled by the DNN trained with nonideal features. After the chips are fabricated, the model weights are directly ported to the accelerator without modification. With HAT, the DNN is able to maintain good classification accuracy across fabricated chips during measurement without requiring expensive chip-in-the-loop training, as reported in Section IV-B.

The feature generation process for HAT is shown in Fig. 7(a). For each clean audio recording of the dataset, we apply data augmentation (see Section II-D) and feed the augmented waveform to the Python-based behavioral model to generate simulated features. The nonideal features are then processed by the DNN (see Section II-C) along with the corresponding groundtruth intent to calculate the loss function and output score (see Section II-D). Prior works [7], [56] also modeled nonidealities in software. However, the nonideal features were used only for evaluation but not during training [7], or the results were limited to simulations [56]. Here, we trained the network with simulated nonideal features and achieved a higher accuracy on measured hardware features.

Fig. 7(b) shows the nonideality models of the analog circuits, covering both deterministic effects such as signal clipping and large-signal nonlinearity, as well as stochastic effects such as noise, offset, gain error, and bias current mismatch. Mismatch-induced effects, including offset, gain error, and BPF center frequency variation, are sampled per recording, while the noise of the amplifiers and comparators is sampled per simulation time step. Regarding large-signal distortion, the output clipping of the amplifiers is modeled using simple thresholding, while the nonlinearity of the BPF (see Section III-C) is modeled using a set of nonlinear ordinary differential equations (ODEs) derived from nodal analysis. The I–V characteristics of the subthreshold transconductors (G_m cells) of the BPF are modeled using the hyperbolic tangent function.

Another important class of analog circuit nonidealities is process, voltage and temperature (PVT) variations. While noise, mismatch, and distortion mainly limit the precision of the circuits, PVT variations could have a more detrimental effect, potentially altering the bias point and affecting the correctness of the circuits. Therefore, instead of using behavioral modeling and HAT to counteract PVT variations, we apply multiple circuit techniques to directly improve the PVT robustness of the circuits: 1) using replicate biasing with local feedback to address process variation; 2) adopting fully differential topology to enhance supply noise rejection; and 3) integrating on-chip proportional to absolute temperature (PTAT) current source to ensure constant transconductance (in weak inversion) across temperature. Further details are provided in Section III.

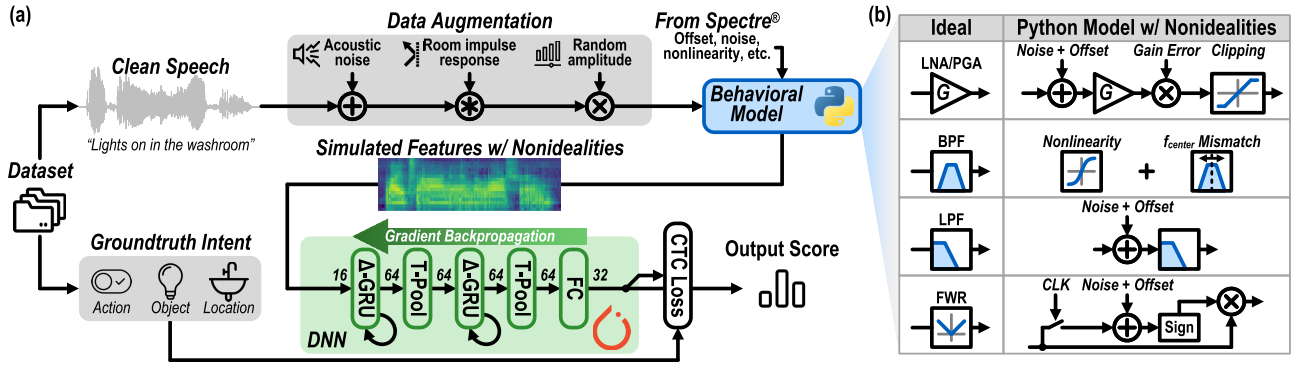


Fig. 7. (a) During HAT, the groundtruth intent directly comes from the dataset, while the input features are generated by applying data augmentation (see Section II-D) followed by the Python-based behavioral model (see Section II-B). Circuit nonidealities are extracted from a transistor-level Spectre simulation. (b) Ideal analog circuits and the corresponding Python model with nonidealities.

C. Temporal-Sparsity-Aware RNN

Biological neurons operate in an event-driven manner, remaining largely inactive until their membrane potential exceeds a certain threshold, at which point they emit a spike. This threshold-triggered firing results in sparse and efficient communication in the brain. Inspired by this event-driven behavior of biological neurons, spiking neural networks (SNNs) encode information in binary spikes, allowing neurons to fire only when necessary. While promising for ultra-low-power computation, SNNs are typically limited by their binary activations and non-differentiable nature, making them less compatible with gradient-based deep learning workflows. To combine the biological efficiency of sparse updates with the practicality of deep learning, temporal-sparsity-aware RNNs introduce a similar thresholding mechanism into conventional RNN architectures. Unlike SNNs, they retain multi-bit activations and remain fully differentiable, making them more practical for existing deep learning frameworks.

A representative model within this paradigm is the Δ-GRU [45], [57], [58], [59], which builds on standard GRU by incorporating a threshold (Δ_{th}) on the temporal changes, reducing computational complexity while preserving inference accuracy once trained properly. Standard gated recurrent unit (GRU) cells, characterized by update and reset gates, continuously recompute their hidden activations irrespective of the magnitude of changes between time steps. In contrast, Δ-GRU selectively updates its states only when changes surpass Δ_{th}, thus exploiting the temporal sparsity inherent in many sequential inputs.

The mathematical formulation begins by defining the thresholded vectors for T-step-D_{in}-dim input activation sequence $x_{i,t} \in \mathbf{X} = \{\mathbf{x}_t \in \mathbb{R}^{D_{\text{in}}} \mid t = 1, \ldots, T\}$ and T-step-D_{hid}-dim hidden activation sequence $h_{j,t} \in \mathbf{H} = \{\mathbf{h}_t \in \mathbb{R}^{D_{\text{hid}}} \mid t = 1, \ldots, T\}$ as follows:

$$\hat{x}_{i,t} = \begin{cases} x_{i,t}, & |x_{i,t} - \hat{x}_{i,t-1}| \geq \Delta_{\text{th}} \\ \hat{x}_{i,t-1}, & |x_{i,t} - \hat{x}_{i,t-1}| < \Delta_{\text{th}} \end{cases}$$
$$\hat{h}_{j,t} = \begin{cases} h_{j,t}, & |h_{j,t} - \hat{h}_{j,t-1}| \geq \Delta_{\text{th}} \\ \hat{h}_{j,t-1}, & |h_{j,t} - \hat{h}_{j,t-1}| < \Delta_{\text{th}} \end{cases} \quad (1)$$

where i and j represent the input and hidden neuron indices, respectively. The Δ-activations are then computed from the thresholded vectors $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{h}}_t$

$$\Delta\mathbf{x}_t = \hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1}, \quad \Delta\mathbf{h}_t = \hat{\mathbf{h}}_t - \hat{\mathbf{h}}_{t-1}. \quad (2)$$

Δ-GRU updates its internal memory states ($\mathbf{M}_{r,t}$, $\mathbf{M}_{u,t}$, and so on) based on these Δ-activations, cumulatively preserving previous computational efforts and updating only when necessary

$$\begin{aligned} \mathbf{M}_{r,t} &= \mathbf{W}_{xr}\Delta\mathbf{x}_t + \mathbf{W}_{hr}\Delta\mathbf{h}_{t-1} + \mathbf{M}_{r,t-1} \\ \mathbf{M}_{u,t} &= \mathbf{W}_{xu}\Delta\mathbf{x}_t + \mathbf{W}_{hu}\Delta\mathbf{h}_{t-1} + \mathbf{M}_{u,t-1} \\ \mathbf{M}_{xc,t} &= \mathbf{W}_{xc}\Delta\mathbf{x}_t + \mathbf{M}_{xc,t-1} \\ \mathbf{M}_{hc,t} &= \mathbf{W}_{hc}\Delta\mathbf{h}_{t-1} + \mathbf{M}_{hc,t-1}. \end{aligned} \quad (3)$$

Subsequently, the gates and hidden activations are updated through sigmoid (σ) and hyperbolic tangent (tanh) activations

$$\begin{aligned} \mathbf{r}_t &= \sigma\left(\mathbf{M}_{r,t}\right), \quad \mathbf{u}_t = \sigma\left(\mathbf{M}_{u,t}\right) \\ \mathbf{c}_t &= \tanh\left(\mathbf{M}_{xc,t} + \mathbf{r}_t \odot \mathbf{M}_{hc,t}\right) \\ \mathbf{h}_t &= (1 - \mathbf{u}_t) \odot \mathbf{c}_t + \mathbf{u}_t \odot \mathbf{h}_{t-1}. \end{aligned} \quad (4)$$

The theoretical speedup of Δ-GRU is quantified by the proportion of computations skipped via Δ-activation sparsity, which is demonstrated in Fig. 8. The top panel shows the GRU first-layer activations for a spoken sentence, while the bottom panel illustrates the significantly sparser updates triggered by a threshold of $\Delta_{\text{th}} = 0.125$. By incorporating the thresholding mechanism during training, Δ-GRU achieves a significant reduction in computational cost with negligible accuracy loss (see Section IV-B).

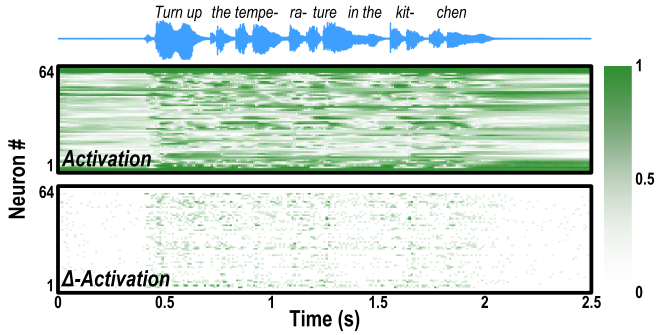


Fig. 8. First-layer GRU activations and the corresponding sparse Δ-activations (with $\Delta_{\text{th}} = 0.125$) in response to a spoken sentence. Darker color indicates larger absolute value.

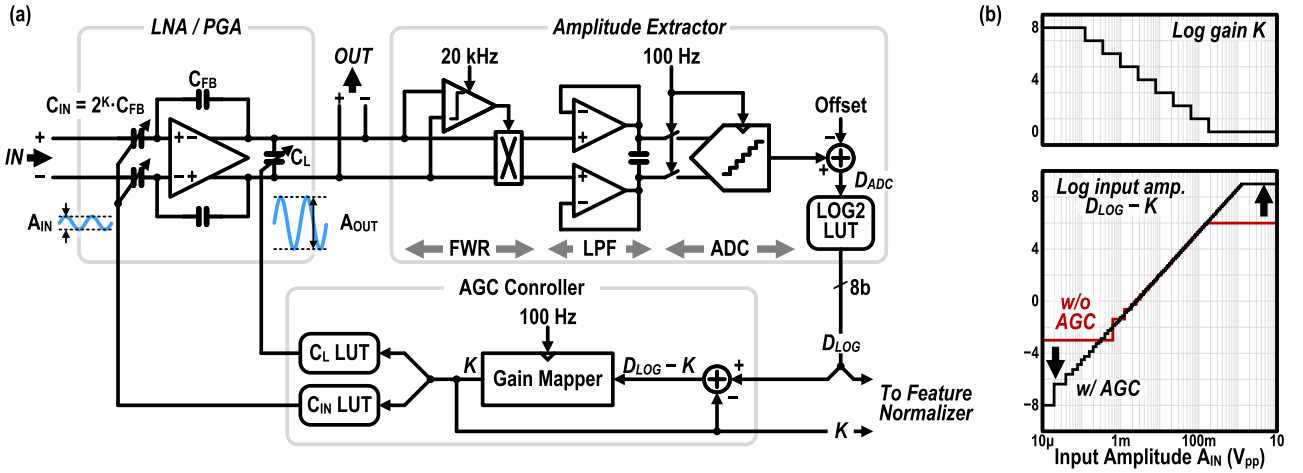


Fig. 9. (a) Global and per-channel AGC feedback loops. (b) Logarithmic gain K and digitized logarithmic input amplitude $D_{\mathrm{LOG}} - K$ versus analog input amplitude A_{IN}. Results from global AGC behavioral model.

In this work, a custom Δ-GRU accelerator is implemented and tightly integrated with the FEx in the SoC. As shown in Fig. 3, it first normalizes the 16-D D_{AFE} every 10 ms, then performs the SLU classification task through a two-layer Δ-GRU network with 64 neurons per layer [see Fig. 7(a)]. A fully connected (FC) layer then outputs the probabilities of up to 32 classes. All inputs, intermediate activations, and weights are quantized to 8 bits via quantization-aware training (QAT), enabling storage of the entire model's weights in 48 kB of on-chip weight memory (WMEM). The implementation details of the accelerator are provided in Section III-D.

D. Streaming-Mode SLU

Our SLU model is trained and evaluated using the publicly available FSCD [35], which consists of 19 h of 16-kHz, single-channel recordings from 97 speakers. Each recording contains a spoken English sentence for controlling a smart home device. Every sentence corresponds to one of the 31 possible user intents, each comprising an action, an object, and a location. As an example, for the sentence "Increase the temperature in the kitchen," the corresponding action, object, and location are "increase," "heat," and "kitchen," respectively. Each intent can be expressed with various sentences. For example, both "Lights on in the washroom" and "Bathroom lights on" express the same intent "activate"+"lights"+"washroom/bathroom." There are 248 different sentences in total.

We apply the data augmentation pipeline shown in Fig. 7(a). First, the clean speech recordings from FSCD are mixed with noise recordings using signal-to-noise ratio (SNR) values between 5 and 30 dB. Then, the signal is convolved with a room impulse response (RIR) to model the reverberation. The noise recordings and RIRs are randomly selected from the OpenRIR dataset [60]. Finally, the amplitude of the noisy-and-reverberant speech is randomly chosen from 8.9 μV_{rms} to 0.28 V_{rms}, corresponding to 30–120-dB sound pressure level (SPL) assuming a microphone sensitivity of −37 dBV (ICS-40310 [46]).

To enable streaming intent classification, we train the Δ-GRU model (see Section II-C) using the connectionist temporal classification (CTC) loss [61] for 500 epochs. It allows the model to handle variable-length output sequences and automatically align them with label sequences of different lengths. CTC loss requires an extra output label <blank> that indicates no intent has been detected. Together with the 31 possible intents, these form the 32 output classes of the Δ-GRU model.

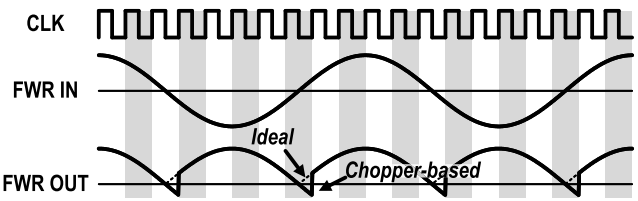


Fig. 10. Behavioral simulation of chopper-based FWR versus ideal FWR.

We also employ an auxiliary cross-entropy (CE) loss during the first 100 training epochs to aid model convergence. During this stage, the model learns to classify the sentence by mapping the final output to the intent label using a softmax classifier. The auxiliary CE loss enables the model to better align the output and label sequences during subsequent CTC training, thereby improving the model's ability to predict the real-time intent in a streaming setting. After training, our SLU model outputs <blank> for both silence/noise and unfinished sentences, and it only outputs the predicted user intent after the sentence is complete, without any assumptions about sentence length.

III. Circuit Implementation

A. Global and Per-Channel AGC

Fig. 9(a) shows the design of the global and per-channel AGC feedback loops. A fully differential, capacitively coupled amplifier provides a programmable gain of $C_{\mathrm{IN}}/C_{\mathrm{FB}} = 2^K$ by reconfiguring the input capacitance C_{IN}. The logarithmic gain K has a step size of 1, so the amplifier gain can be set in 6-dB steps. Different amplifier designs are employed in the global and per-channel AGC. The LNA for global AGC interfaces directly with the microphone, so it is optimized for low noise and has a nominal tuning range of 0–48 dB. The PGAs for per-channel AGC contribute negligible amount of noise due to the LNA amplification, so they have less stringent noise

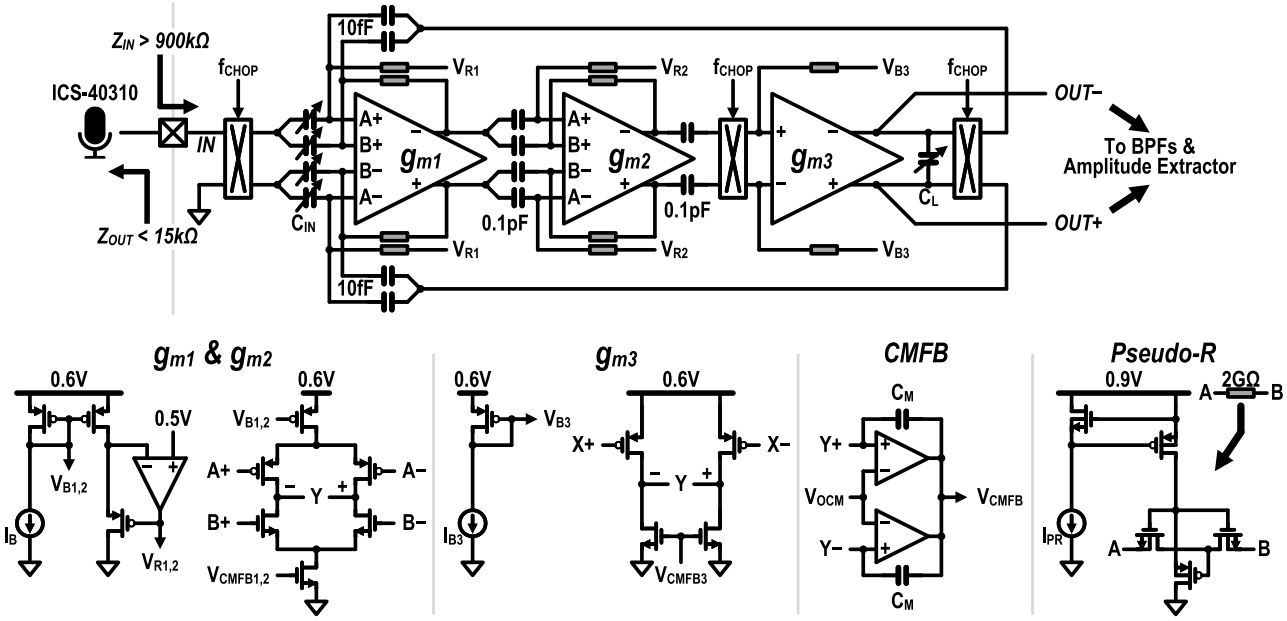


Fig. 11. Three-stage capacitively coupled chopper LNA.

requirements and a smaller nominal tuning range of 0–36 dB. The load capacitance C_L is adjusted along with C_{IN} to set the closed-loop bandwidth and improve the phase margin.

The amplitude extractor converts the analog output of the amplifier into digital values D_{LOG} that represent its instantaneous log amplitude. It consists of a cascade of a full-wave rectifier (FWR), a first-order low-pass filter (LPF), a 10-bit successive approximation register (SAR) ADC, and an LUT. The FWR and LPF down-convert the amplitude of the original high-bandwidth (up to 8 kHz) signal into a low-bandwidth signal (up to 50 Hz), allowing it to be digitized by the ADC at only 100 Hz. The ADC output code is log-compressed by the LUT after offset removal. A unit-length capacitor array [62] is employed for the SAR ADC to minimize its area.

The FWR plays a central role in extracting the instantaneous signal amplitude. The current-domain FWR in [7] employed an open-loop operational transconductance amplifier (OTA) for linear voltage-to-current conversion. Achieving sufficient DR with this approach requires a large OTA bias current, with the FWR contributing over 50% of the AFE's power. In [15], FWR power was reduced by using the nonlinear voltage-to-current conversion of a single subthreshold NMOS, though at the expense of sensitivity to PVT variation. Instead, we employ the low-power chopper-based FWR proposed in [63], which uses a chopper controlled by a dynamic comparator. This approach requires no static bias current and is robust to PVT variation. Fig. 10 shows that the output of the chopper-based FWR deviates from the ideal due to the discrete-time operation of the comparator. This effect is included in the FEx behavioral model [see Fig. 7(b)], allowing the DNN to compensate for the nonideal rectification. The comparator is clocked at 20 kHz, which provides a favorable power–accuracy tradeoff as shown in [63]. The measured per-channel FWR power is 10× lower than [7].

The AGC controller subtracts the logarithmic gain K from the log output amplitude D_{LOG} to compute the log input amplitude $D_{\text{LOG}} - K$. Since all arithmetic is performed in the log domain, no division or multiplication is required. The gain mapper updates K after each ADC conversion using the staircase function in Fig. 9(b), decreasing the amplifier gain for large-amplitude inputs. C_{IN} and C_L are configured via two LUTs according to K. Finally, K and D_{LOG} are sent to the feature normalizer to calculate D_{AFE}. Compared to fixed-gain amplification, AGC enlarges the represented input range by applying the proper gain to prevent ADC overflow or underflow. This enables low-power wide-DR feature extraction by relaxing the noise requirements of all analog circuits following the amplifiers, including the BPFs, PGAs, and amplitude extractors, thereby reducing their static bias currents. Furthermore, clipping and distortion are also avoided at large input amplitudes by reducing the gain on-demand. Therefore, a high power supply is not required to accommodate the otherwise large amplifier output swing.

B. Low-Noise Amplifier

Fig. 11 shows the design of the LNA. A three-stage design is adopted for the core amplifier ($g_{m1,m2,m3}$) to achieve sufficient (>70 dB) open-loop gain. The dominant pole is placed at the last stage since it must drive a large capacitive load due to the BPFs. The first two stages g_{m1} and g_{m2} employ current-reuse inverter-based amplifiers to reduce the bias currents required to achieve the target noise level, while the last stage g_{m3} adopts a simple common-source topology for larger output swing.

A 0.6-V supply is chosen to reduce the LNA power. To ensure robust operation under process variation at this low supply voltage, the PMOS and NMOS input pairs of g_{m1} and g_{m2} are biased independently using PVT-robust pseudo-resistors [64] with an equivalent resistance of 2 GΩ. The PMOS bias voltages V_{R1} and V_{R2} are generated by local feedback loops to maintain a 0.1-V headroom for the tail current source, while the NMOS bias is set by the OTA output via dc feedback to prevent the OTA offset from saturating its

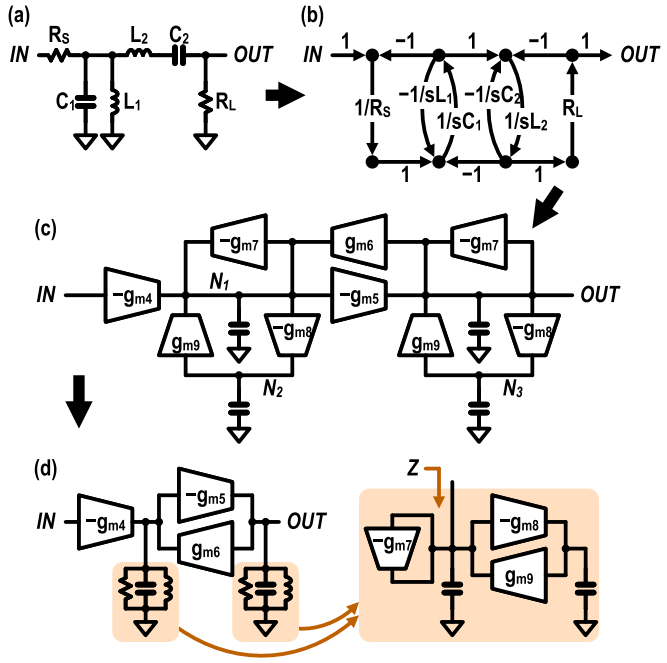


Fig. 12. (a) Doubly terminated LC ladder filter. (b) Corresponding SFG. (c) G_m-C BPF implementing the SFG. (d) Simplified BPF schematic.

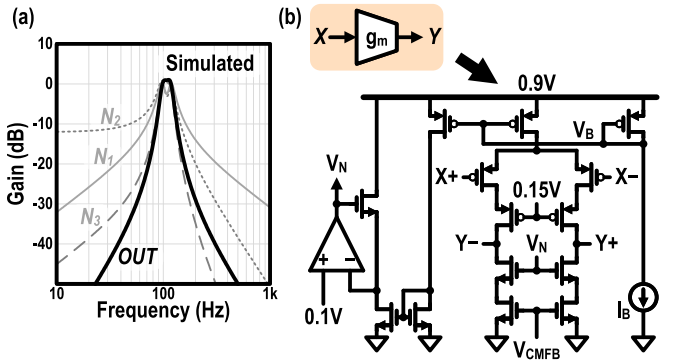


Fig. 13. (a) Simulated gain of the first BPF channel (with 100-Hz central frequency) at its output and internal nodes N_1-N_3 [see Fig. 12(c)]. (b) Design of the G_m cell.

output [65]. To achieve optimal input and output biasing at each stage, AC coupling is employed between stages, and each OTA has an individual common-mode feedback (CMFB) loop. The CMFB amplifier adopts Miller compensation with C_M.

Chopping at f_{CHOP} = 100 kHz is employed to suppress flicker noise, and the chopping ripple is rejected by the AC coupling between g_{m2} and g_{m3} [66]. To minimize microphone output attenuation, the LNA should have a relatively large input impedance $Z_{IN} = 1/(2C_{IN}f_{CHOP})$ compared to the microphone output impedance Z_{OUT}, which is typically less than 15 kΩ [46]. Using two minimum-sized 10-fF metal-insulator-metal (MIM) capacitors for C_{FB}, C_{IN} < 5.2 pF and Z_{IN} > 900 kΩ are achieved, and the microphone attenuation is less than 0.15 dB.

C. Bandpass Filter

The frequency selectivity of the BPF is critical for extracting distinctive spectral features and improving the accuracy of downstream tasks. Behavioral simulation shows that a fourth-order Butterworth BPF with Q = 4 provides good SLU classification accuracy. Comparing with second-order BPFs implemented in prior analog FEx designs [7], [15], [16], [17], [18], [19], the steeper roll-off of the fourth-order BPF improves SLU accuracy by 0.93% from 83.92% to 84.85% on the FSCD validation set.

Fig. 12 illustrates the half-circuit of the fully differential G_m-C BPF, consisting of nine G_m cells, four capacitors, three internal nodes, and an output node. Part of the BPF's output load capacitance is shared with the PGA input capacitance to account for the PGA input impedance and reduce total capacitor area. The topology of the G_m-C filter is derived from a doubly terminated LC ladder filter due to its low sensitivity to component mismatch [67]. Specifically, given the LC ladder filter in Fig. 12(a), we perform nodal analysis and represent the equations using the signal flow graph (SFG) in Fig. 12(b). Then, the SFG is implemented using the G_m-C topology in Fig. 12(c), where the integrators in the SFG are realized by G_m-C integrators. The topology can be further simplified as shown in Fig. 12(d) containing two coupled resonators.

Source-follower-based BPFs have been widely used in prior analog FEx designs [7], [15], [18], [68] due to their simplicity and power efficiency. However, to realize a fourth-order BPF with Q = 4, two biquads with Q > 5.5 are required. The super source follower (SSF)-based BPF [7] has a passband gain > Q at the internal node, while the flipped voltage follower (FVF)-based BPF [15] has a passband gain of Q^2 at the output node. Such high passband gains lead to excessive output distortion. In contrast, as shown in Fig. 13(a), our G_m-C topology allows independent scaling of the G_m cells to ensure near-0-dB gain at the output and all internal nodes, enhancing BPF linearity.

The design of the G_m cell is shown in Fig. 13(b). It achieves >70-dB dc gain using the telescopic cascode topology. The input and output common modes are set to 0.3 V to allow DC coupling with the LNA. Each internal node, as well as the output node, contains an individual CMFB loop similar to that of the LNA. The NMOS cascode bias V_N is generated by local feedback to maintain a 0.1 V headroom for the CMFB NMOS, while the PMOS cascode bias is fixed at 0.15 V to ensure >0.1 V headroom for the input PMOS. The OTA bias currents for CMFB and V_N generation are proportional to the central frequency of the BPF channel, thereby reducing power.

D. RNN Accelerator

Fig. 14 shows the computation steps of the two techniques implemented in our Δ-GRU accelerator (the DBE block in Fig. 3), which collectively reduce the compute workload and therefore improve the energy efficiency. The first technique is to leverage the Δ-activation sparsity (see Section II-C) by using the Δ-Encoder in Fig. 15. The Δ-Encoder calculates the temporal changes of its input and hidden neuron activations ($\mathbf{x}_t$ and $\mathbf{h}_t$) and zeros out the neurons with a temporal change below a preset threshold Δ_{th}. As shown in Fig. 14, these inactive neurons, depicted as white elements in the Δ-activation vectors, have no downstream effects; and their associated multiplications and WMEM accesses are skipped.

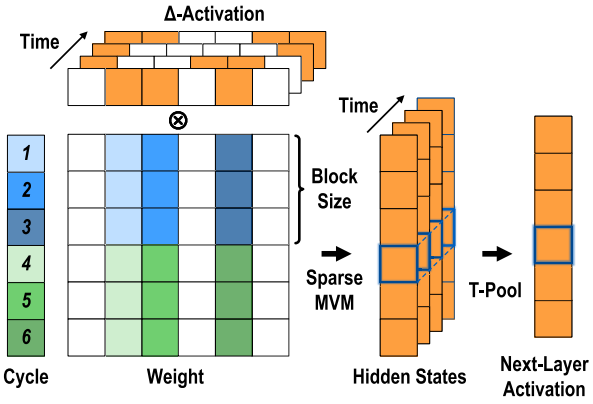


Fig. 14. Computation steps for a Δ-GRU layer. Recurrent feedback and gating are omitted for clarity. The block size, which is equal to the number of PEs and parameters per WMEM address, is 8 in the actual design.

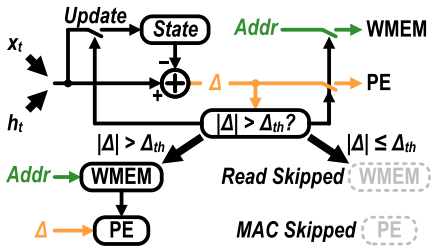


Fig. 15. Design of the Δ-Encoder.

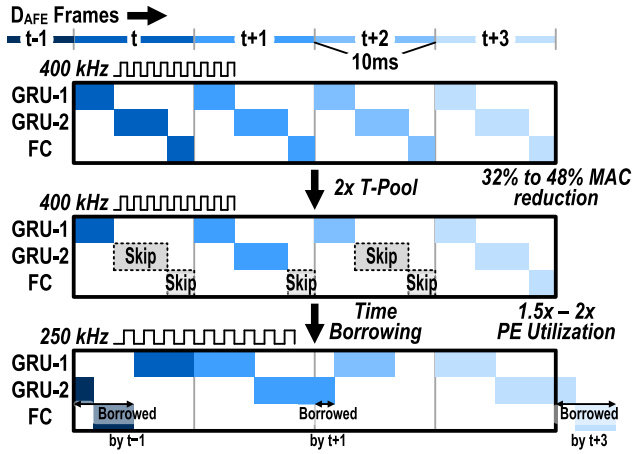


Fig. 16. Time schedule of temporal pooling and time borrowing. As an illustrative example, the pooling window size is $T_p = 2$ in this figure.

Compared to the existing Δ-GRU accelerators [45], [57], [58], [59], we further employ the second technique: adding an average temporal pooling (T-Pool) layer to the Δ-GRU outputs. By collapsing activations across time steps into a single vector, T-Pool reduces both arithmetic operations and memory access in all subsequent layers. Fig. 16 illustrates the time schedule of the temporal pooling. The output of each Δ-GRU layer is averaged over T_p successive frames and activates the next layer's computation only for the final frame, where T_p is defined as the pooling window size. By skipping the preceding $T_p - 1$ frames, this approach reduces multiply-accumulate (MAC) operations by 48% ($T_p = 4$) for the SLU task without accuracy drop. Despite the benefits, temporal pooling results in an imbalanced processing workload across different frames and limits the processing element (PE) utilization.

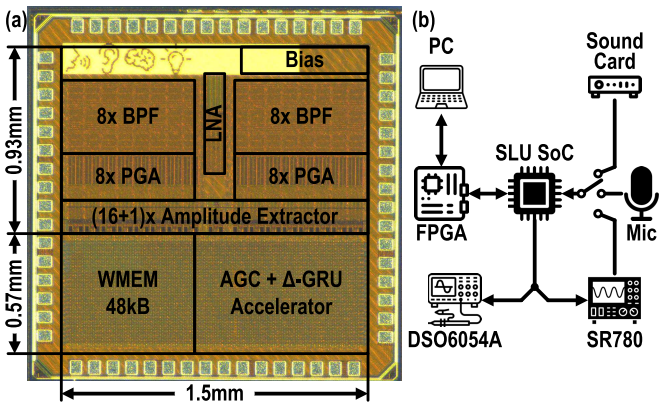


Fig. 17. (a) Chip micrograph. (b) Measurement setup block diagram.

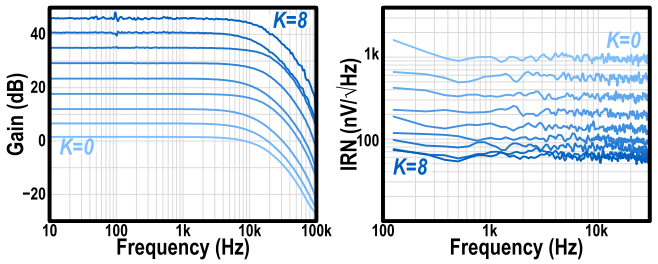


Fig. 18. Measured LNA gain and IRN spectrum.

A time-borrowing scheme alleviates this imbalance by redistributing idle clock cycles from frames processed by only the first Δ-GRU layer to those requiring deeper processing. The optimal scheduling improves the PE utilization by 1.5× to 2× and lowers the clock frequency requirement from 400 to 250 kHz.

With both optimizations in place, eight parallel PEs perform the sparse MVM in a row-level parallel style. Each weight column is partitioned into eight-element blocks, one element for each PE. On each clock cycle, the activation of one active neuron is broadcast to eight PEs and multiplied with a weight block. After processing the first block of this active neuron (cycle 1 in Fig. 14), the engine immediately moves to the first block of the next active neuron (cycle 2). This process continues until all active neurons are serviced. Once the first block of outputs are computed, the engine returns to process the second weight block of the first active neuron. In comparison, prior Δ-GRU designs [45], [57], [58], [59] process an entire weight column before moving to the next neuron and must store column-wide partial sums in a buffer memory. By reducing the partial sum size from the whole column to a single block, the partial sums can be stored using local flip-flops, thereby eliminating the need of a buffer memory and reducing the memory-access power. In addition, each PE includes two LUTs to implement the tanh and sigmoid activation functions used in the GRU model. LUT inputs and outputs are quantized to 8 bits. Synthesis results show that these LUTs occupy a negligible 0.6% of the total DBE area.

IV. Measurement Results

Fig. 17 shows the micrograph of the fully integrated SLU SoC and its measurement setup. The chip is fabricated in

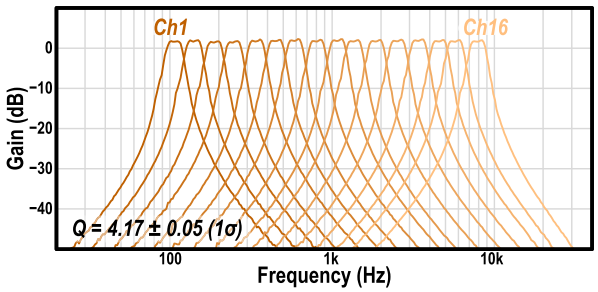


Fig. 19. Measured frequency responses of the BPFs.

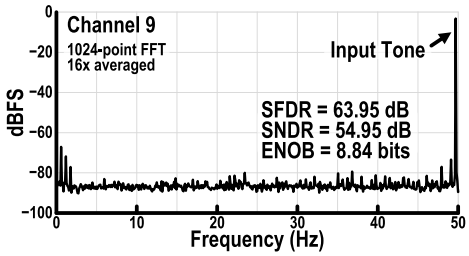


Fig. 20. Measured ADC output spectrum.

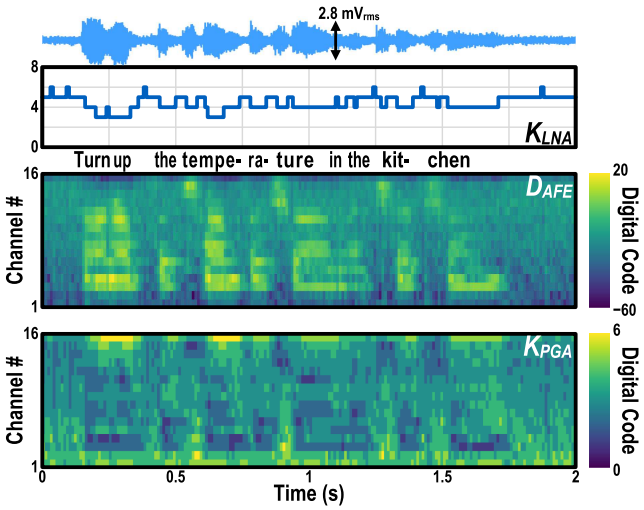


Fig. 21. Measured K_{LNA}, K_{PGA}, and D_{AFE} for an FSCD sample.

TSMC 65-nm CMOS low-power process. Out of the active area of 2.23 mm^2, the AFE and DBE occupy 1.38 mm^2 (62%) and 0.844 mm^2 (38%), respectively. The digital circuits, including the Δ-GRU accelerator and the digital parts of the AGC feedback loops, are implemented with the standard automatic place-and-route flow. An on-chip SPI interface is implemented for configuring the SoC and sending the FEx/DNN outputs off-chip. Source-follower-based test buffers are added between the analog blocks to facilitate circuit characterization and are disabled during end-to-end operation. All bias currents and reference voltages are generated by on-chip circuits.

A. Analog Front End

Fig. 18 shows the measured LNA gain and input-referred noise (IRN) spectrum at different gain settings. The closed-loop gain is tunable as expected and the closed-loop bandwidth is greater than 10 kHz at all gain settings. The IRN density is

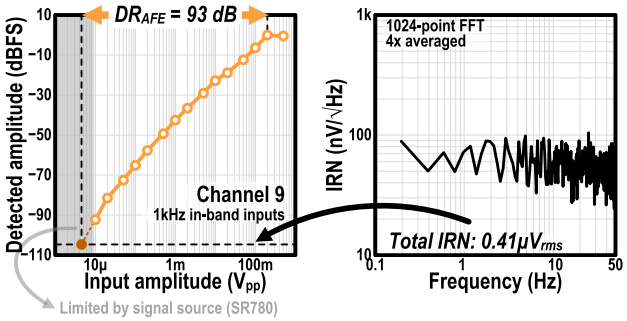


Fig. 22. Measured FEx IRN spectrum and DR.

59.7 nV/$\sqrt{\text{Hz}}$ at the highest gain level. For comparison, the microphone output noise density is >60 nV/$\sqrt{\text{Hz}}$ [46]. The LNA IRN is dominated by thermal noise since flicker noise is up-modulated by chopping. The low IRN of the amplifier allows usable acoustic features to be extracted from soft sound.

Fig. 19 shows the measured transfer functions of the 16 BPFs. The results show that the filters have a flat passband response, in agreement with the transfer function of a fourth-order Butterworth filter. The passbands of adjacent filters do not overlap despite central frequency variation due to transistor mismatch. The measured mean quality factor Q of all channels is 4.17 with a 1σ deviation of 0.05. The frequency responses are well-matched across all channels without per-channel calibration due to the mismatch-insensitive G_m-C topology derived from LC ladder filters.

Fig. 20 shows the measured ADC output spectrum for a near-Nyquist-frequency input tone. The signal-to-noise and distortion ratio (SNDR) is measured to be 54.95 dB, resulting in an effective number of bits (ENOB) of 8.8 bits, which is higher than the 8-bit design target.

Fig. 21 shows the measured features D_{AFE} along with the global (K_{LNA}) and per-channel (K_{PGA}) gain settings for a 2.8-mV_{rms} spoken sentence "Turn up the temperature in the kitchen." The spectral components of the input audio are clearly visible in the D_{AFE} feature map. For example, the four phonemes /k/, /ɪ/, /tʃ/, and /ən/ of the word "kitchen" are discernible at the end of the sentence. K_{LNA} is adjusted according to the instantaneous amplitude of the audio, while K_{PGA} of each channel varies with the signal energy within the corresponding frequency band.

To quantify the DR of the AFE, denoted as DR_{AFE}, we apply sinusoidal test inputs of different amplitudes to the LNA and measure the amplitude detected by the FEx by using K_{LNA}, K_{PGA} and the ADC output codes D_{ADC} of the corresponding channel. Note that D_{ADC} is recorded after ADC offset removal, as shown in Fig. 9. The amplitude detected by the FEx is defined as $D_{ADC}/2^{K_{LNA}+K_{PGA}}$. Fig. 22 shows the detected amplitude versus test input amplitude, indicating that the FEx is able to detect up to 200-mV_{pp} input without saturation, while the low end of the curve is limited by the signal source SR780. The curve is extrapolated toward the IRN floor of the FEx, which is measured to be 0.41 μV_{rms} within the 50-Hz ADC bandwidth, leading to a DR_{AFE} of 93 dB.

Table I compares the state-of-the-art analog FEx designs for voice interface applications. Our analog FEx achieves 38 dB higher DR than the prior state-of-the-art [17], enabled by

TABLE I
COMPARISON OF ANALOG ACOUSTIC FEX

	Symbol	Unit	[7] Yang JSSC'19	[8] Oh JSSC'19	[15] Yang JSSC'21	[17] Kim JSSC'22	[18] Ray JSSC'23	[19] Mostafa ISSCC'24	This work
Process	–	nm	180	180	65	65	65	65	65
Signal Domain	–	–	Voltage	Voltage	Voltage	Time	Time	Time	Voltage
Number of Channels	N	–	16	32	16	16	31	16	16
Supply Voltage	V_{DD}	V	0.6	0.6/1.4	0.6	0.5	0.6	0.4	0.6/0.9
Area/Channel	A_{CH}	mm^2	0.1	0.052	0.056	0.1	0.017	0.0094	0.086
Power	P	µW	0.38	0.06	0.053	9.3	0.08	0.988	1.85
Frame Rate	R	Hz	100	1.95	100	62.5	100	100	100
Frequency Range	f_L – f_H	Hz	100 – 5k	75 – 4k	100 – 5k	111 – 10.4k	100 – 4k	125 – 5k	100 – 8k
Normalized Power	P_{NORM}[1]	µW	0.354	0.037	0.049	4.708	0.047	0.879	1.183
Normalized Energy/Frame	E_{NORM}[1]	nJ	3.54	18.83	0.49	75.32	0.47	8.79	11.83
Dynamic Range	DR_{AFE}	dB	40	47	–	54.89	35	42	**93**
Figures of Merit	FoM_S[2]	dB	121.5	121.2	–	123.1	125.2	119.6	**169.3**
	$FoM_{A,CH}$[3]	dB	135.9	131.2	–	139.1	147.4	147.8	**180.8**
Task	–	–	VAD	VAD	KWS	KWS	KWS	KWS	**SLU**
Number of Classes	–	–	2	2	5[4]	12	11	10	**32**
On-Chip DNN	–	–	✓	✓	✓	✓	✗	✗	✓

[1] From [69]: $P_{NORM} = P \cdot (1-r)/(1-r^N) \cdot 20\,\text{kHz}/f_H$, $r = (f_L/f_H)^{1/(N-1)}$, $E_{NORM} = P_{NORM}/R$.
[2] From [17]: $FoM_S = DR_{AFE} + 10\log_{10}(R/2P_{NORM})$, re-calculated with R in Hz.
[3] From [19]: $FoM_{A,CH} = FoM_S - 20\log_{10}(V_{DD}/1\,\text{V}) - 10\log_{10}(A_{CH}/1\,\text{mm}^2)$, re-calculated with per-channel instead of total area.
[4] Reported in [16], 4 keywords and 1 filler.

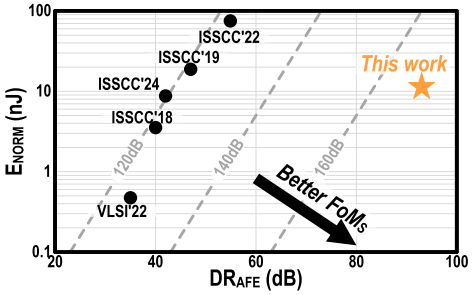


Fig. 23. Comparison of state-of-the-art analog acoustic FEx designs.

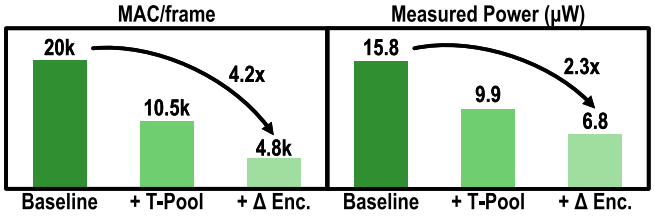


Fig. 24. Measured MAC per frame and DBE power reduction.

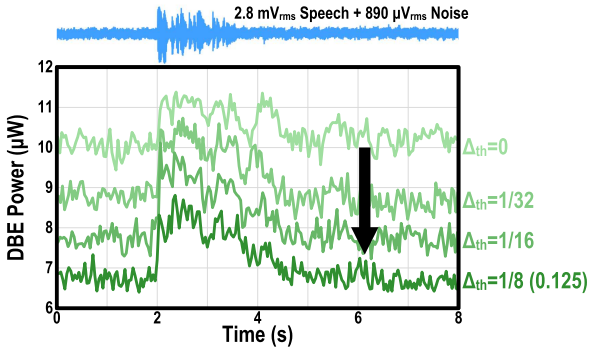


Fig. 25. Measured DBE instantaneous power for different Δ_{th} values for a speech sample in the presence of noise.

the integrated global and per-channel AGC. The normalized energy per feature E_{NORM} [69], which factors in the possibly different frequency range (f_L–f_H), number of channels (N), and frame rate (R) of different FEx designs, is measured to be 11.83 nJ. To fairly compare the energy efficiency of the designs with different DR, we use the Schreier figure of merit FoM_S [17] and its area- and supply-normalized variant $FoM_{A,CH}$ [19]. Our design achieves 169.3-dB FoM_S, 44–50 dB higher than prior designs. Fig. 23 further visualizes different designs in terms of DR_{AFE}, E_{NORM}, and FoM_S. It also achieves the best reported 180.8-dB $FoM_{A,CH}$, 33 dB higher than the prior state of the art [19].

The high energy efficiency of the AFE is achieved by a synergistic combination of multiple techniques. At the *system level*, we use the Python-based behavioral model and HAT (see Section II-B) to define circuit specifications and avoid unnecessary power consumption due to circuit overdesign. At the *architectural level*, we use global and per-channel AGC (see Section II-A) to reduce the bias current and supply voltage of the analog circuits without sacrificing DR. At the *circuit level*, we also employ various optimizations within the analog building blocks to further reduce power (see Section III). In addition, our design is also the first analog FEx that supports audio classification with more than 30 classes.

B. System-on-Chip

Fig. 24 shows the MAC operations and power savings achieved by jointly applying temporal pooling and Δ-encoding. Measured at f_{clk} = 250 kHz, combining pooling window size T_p = 4 with the Δ-threshold Δ_{th} = 0.125 reduces MAC operations by 4.2× and digital power by 2.3× (from 15.8 to 6.8 μW) compared to the baseline (f_{clk} = 400 kHz, T_p = 1, and Δ_{th} = 0). Future work will focus on reducing DBE power consumption. In particular, the use of custom memory cells can help reduce both the DBE power supply and SRAM leakage [22], [27], [33], [59]. The custom SRAM design [59] in the same 65-nm technology reduced read power by

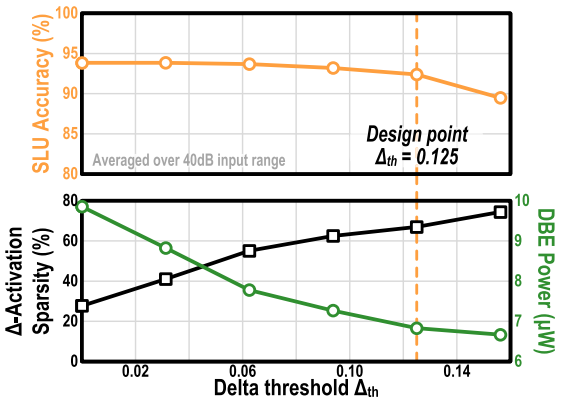


Fig. 26. Measured SLU accuracy, Δ-activation sparsity, and DBE power across different threshold Δ_{th}'s.

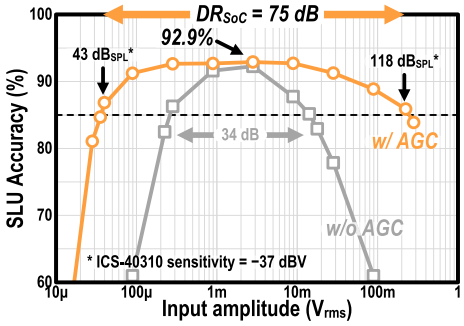


Fig. 27. Measured end-to-end 32-class SLU accuracy on FSCD.

TABLE II
COMPARISON OF SLU ACCURACY WITH AND WITHOUT HAT

Train Features	HAT	Test Set Accuracy (%)[1]		Acc. Drop (%)
		Simulated	Measured	
Simulated	✗	96.0	89.7	6.3
Simulated	✓	93.1	**92.4**	**0.7**

[1] Averaged over 40 dB input range.

6.6× over foundry SRAM and lowered the supply voltage to 0.6 V. Applying this design to our WMEM and using a lower DBE power supply, we expect a reduction of at least half of the current DBE power. Adding an integrated VAD module to selectively activate the accelerator can further reduce the standby power consumption [33].

Fig. 25 shows the measured DBE instantaneous power for a 2.8-mV_{rms} speech sample mixed with white noise at 10-dB SNR. Owing to its temporal-sparsity-aware compute, the DBE power adaptively scales with the input activity, automatically lowering the power consumption for non-speech input by 13.7% compared to speech when Δ_{th} = 0.125.

Fig. 26 shows the measured SLU accuracy, Δ-activation sparsity, and DBE power at different Δ_{th}'s. When Δ_{th} = 0, we observe a non-zero Δ-activation sparsity of 27.7% due to the quantization of the activation values. Using larger Δ_{th} increases Δ-activation sparsity, leading to lower DBE power but also slightly lower SLU accuracy. By sweeping Δ_{th}, we find that Δ_{th} = 0.125 provides a good tradeoff between accuracy and power, achieving a 67% Δ-activation sparsity and 31% power reduction with a minor 1.4% accuracy drop compared to Δ_{th} = 0.

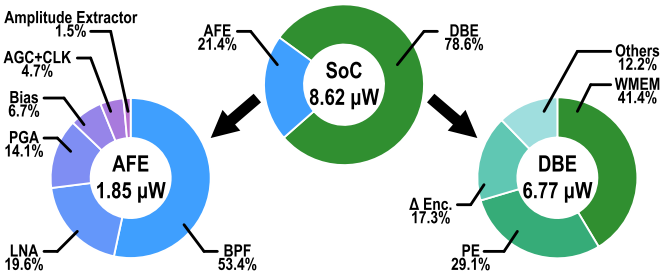


Fig. 28. SoC power breakdown.

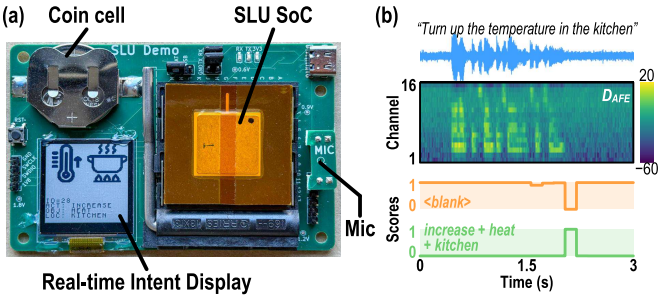


Fig. 29. (a) Coin-cell-powered SLU demo system. (b) FSCD sample, the corresponding measured features (D_{AFE}), and the output scores for <blank> and the correct intent class.

Fig. 27 shows the end-to-end measured 32-class SLU accuracy using the entire FSCD test set. A sound card is used to playback the dataset recordings and the analog signal is fed to the LNA input. The 92.9% accuracy is achieved for 2.8 mV_{rms} input, corresponding to 80-dB SPL with ICS-40310. We define the usable DR of the SoC, denoted as DR_{SoC}, as the input range on which >85% SLU accuracy can be achieved. The measured DR_{SoC} is 75 dB with AGC enabled and 34 dB without. The DR_{SoC} is not limited by the training procedure, which covers the entire measured input range of 90 dB (see Section II-D). AGC boosts DR_{SoC} by 41 dB. It also improves the SLU accuracy by >6% for 280-μV_{rms} inputs, which correspond to normal conversational speech at 60-dB SPL.

Table II compares the simulated and measured SLU accuracy with and without HAT. Both models are trained with simulated features generated by the Python behavioral model. After training, the models are evaluated using both simulated features as well as measured features from the fabricated FEx. Without HAT, no nonidealities are included in the simulated features, and the model achieves 96.0% accuracy in simulation. However, when tested with measured features, the accuracy is only 89.7%, with a 6.3% drop due to the analog circuit nonidealities. With HAT, the measured accuracy increases to 92.4%, showing the effectiveness of the Python-based FEx model incorporating circuit nonidealities. In addition, our analog FEx design achieves >90% task accuracy using on-chip DNN, while previous designs [18], [19] relied on software DNN. SLU accuracy is further measured on five fabricated chips, all programmed with identical DNN weights obtained through HAT and tested using a 2.8-mV_{rms} input. The average

TABLE III
COMPARISON OF FULLY INTEGRATED AUDIO CLASSIFICATION SOC

	Unit	[16] Wang JSSC'25	[27] Yang JSSC'24	[21] Giraldo JSSC'20	[17] Kim JSSC'22	[33] Tan JSSC'25	[32] Park VLSI'24	This work
Speech Task	–	KWS	KWS	KWS	KWS	KWS	KWS	**SLU**
Context Length	–	Word	Word	Word	Word	Word	Word	**Sentence**
Number of Classes	–	5	7	12	12	12	12	**32**
FEx	–	Analog	Digital	Digital	Analog	CNN	Digital	Analog
Algorithm	–	SNN[1]	Skip GRU	LSTM	GRU	CNN	CNN	Δ-GRU
Process	nm	65	28	65	65	28	65	65
Memory	kB	8.2	18	105	27	16	5	48
Area	mm^2	2.71	0.8	2.56	2.03	0.12	1.32	2.23
Accuracy	%	90.2	92.8	90.9	86.0	91.8	92.7	**92.9**
DR_{SoC}	dB	–	–	–	–	–	–	**75**
FEx Power	µW	0.11	0.77	8.98[2]	13	–	–	1.85
DNN Power	µW	0.46	0.71	3.37[2]	10	–	–	6.77
Total Power	µW	0.57	1.48	16.1	23	1.73	5.6	8.62

[1] Including the SNN chip reported in [70].
[2] Power breakdown reported in [71]. Leakage and clock power excluded.

accuracy is 93.3% with σ = 0.59% despite chip-to-chip variation.

Fig. 28 shows the SoC power breakdown. When running inference on continuous FSCD samples, the SoC consumes 8.62 μW with 1.85 μW from the 0.6-/0.9-V AFE and 6.77 μW from the 0.75-V DBE. The amplifiers and filters together account for 87.1% of the AFE power and AGC only 4.7%. The WMEM is the most power consuming block in the DBE, accounting for 41.4% of its power, followed by the PE array (29.1%) and Δ-Encoder (17.3%).

To demonstrate the real-world applicability of the design, the fabricated SoC is used to build a standalone SLU system powered by a CR2450 coin cell, as shown in Fig. 29(a). The SoC interfaces with an SE analog microphone, extracts the acoustic features, and performs intent classification on-chip. A microcontroller controls the display for output visualization based on the user intent output from the SoC. Fig. 29(b) shows the measured features (D_{AFE}) and the normalized output scores for two of the classes (<blank> and the correct intent "increase"+"heat"+"kitchen") when the SoC processes the corresponding FSCD sentence. For each input frame, the user intent is obtained by taking the class with the highest score.

Our design can also be used in non-speech applications. As an initial study, we train a network on the SPRSound dataset [72] to perform the binary normal-versus-adventitious classification task on respiratory sounds. The network has the same architecture as the SLU network except that the FC layer now has only two output classes. It is again trained using simulated features with HAT while evaluated with measured features. Performance on this task is measured by the harmonic score, which is defined as the harmonic mean of sensitivity and specificity, ranging from 0 to 1. The harmonic score achieved by the SoC is 0.856, which is 0.02 higher than the score achieved in [73] using a larger CNN model with conventional digital FEx.

Table III compares the state-of-the-art end-to-end audio classification SoC, focusing on designs that support at least five output classes. Our design is the first sub-10-μW fully integrated SoC that supports on-device SLU. This task is more challenging than the KWS task supported by prior works in terms of context length and number of classes. In addition, it achieves a competitive accuracy of 92.9% for 32-class classification, exceeding the task accuracy of other designs that support fewer classes. Importantly, input amplitude variation is inevitable in real-world especially for far-field operation and leads to a significant accuracy drop without a wide-DR FEx, as shown in Fig. 27. By incorporating AGC in the FEx, our design maintains >85% SLU accuracy over a wide DR_{SoC} of 75 dB.

V. CONCLUSION

We present a fully integrated SoC for on-device SLU. The mixed-signal ASIC implemented in 65-nm CMOS process occupies 2.23 mm^2 and consumes only 8.62 μW when continuously processing the incoming audio in real-time. By combining global and per-channel AGC with logarithmic compression, the analog FEx achieves the highest reported DR, FoM_S, and $FoM_{A,CH}$. The Python-based behavioral model of the analog FEx enables quick generation of simulated features and therefore design space exploration. When the model is combined with HAT, the RNN model trained with simulated features maintains high SLU accuracy on fabricated chips despite analog circuit nonidealities and chip-to-chip variation. The temporal-sparsity-aware compute exploited by the streaming-mode RNN accelerator reduces its power consumption by 2.3×. The SoC supports end-to-end user intent understanding with up to 32 classes and achieves 92.9% accuracy on FSCD. It also maintains >85% accuracy over a wide DR_{SoC} of 75 dB, enhancing its robustness in real-world applications. The SoC interfaces directly with an SE analog microphone, which significantly reduces the system-level power since off-the-shelf digital or differential analog microphones consume more than 200 μW. To the best of our knowledge, this work is the first demonstration of sub-10-μW fully integrated wide-input-DR on-device SLU. Future design improvements include further optimization of the DBE power as discussed in Section IV-B, as well as using gradient-based methods to jointly optimize the FEx and DNN parameters [56], [74], [75].

ACKNOWLEDGMENT

The authors would like to thank Frank K. Gürkaynak, Beat Muheim, and Zerun Jiang from the Microelectronics Design Center, ETH Zürich, Zürich, Switzerland, for their support on EDA tools and PDK.

REFERENCES

[1] Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki, and A. S. Hafid, "A comprehensive survey on TinyML," *IEEE Access*, vol. 11, pp. 96892–96922, 2023.

[2] P. Spachos, S. Gregori, and M. J. Deen, "Voice activated IoT devices for healthcare: Design challenges and emerging applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 7, pp. 3101–3107, Jul. 2022.

[3] S. Yadav, P. A. D. Legaspi, M. S. O. Alink, A. B. J. Kokkeler, and B. Nauta, "Hardware implementations for voice activity detection: Trends, challenges and outlook," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 3, pp. 1083–1096, Mar. 2023.

[4] Z. Zhu and L. Feng, "A review of sub-μW CMOS analog computing circuits for instant 1-dimensional audio signal processing in always-on edge devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 71, no. 9, pp. 4009–4018, Sep. 2024.

[5] S. Zhou, Z. Li, T. Delbruck, K. Kim, and S.-C. Liu, "An 8.62μW 75dB-DR_{SoC} end-to-end spoken-language-understanding SoC with channel-level AGC and temporal-sparsity-aware streaming-mode RNN," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2025, pp. 238–240.

[6] 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.

[7] 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.

[8] 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.

[9] M. Croce, B. Friend, F. Nesta, L. Crespi, P. Malcovati, and A. Baschirotto, "A 760-nW, 180-nm CMOS fully analog voice activity detection system for domestic environment," *IEEE J. Solid-State Circuits*, vol. 56, no. 3, pp. 778–787, Mar. 2021.

[10] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz, and V. De, "A 2.3 nJ/frame voice activity detector-based audio front-end for context-aware system-on-chip applications in 32-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, Aug. 2013.

[11] M. Price, J. Glass, and A. P. Chandrakasan, "A low-power speech recognizer and voice activity detector using deep neural networks," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 66–75, Jan. 2018.

[12] F. Chen, K.-F. Un, W.-H. Yu, P.-I. Mak, and R. P. Martins, "A 108-nW 0.8-mm^2 analog voice activity detector featuring a time-domain CNN with sparsity-aware computation and sparsified quantization in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3288–3297, Nov. 2022.

[13] J. Lin, K.-F. Un, W.-H. Yu, R. P. Martins, and P.-I. Mak, "A 47-nW voice activity detector (VAD) featuring a short-time CNN feature extractor and an RNN-based classifier with a non-volatile CAP-ROM," *IEEE J. Solid-State Circuits*, vol. 58, no. 11, pp. 3020–3029, Nov. 2023.

[14] Y. Liu et al., "A 0.22 mm^2 161nW noise-robust voice-activity detection using information-aware data compression and neuromorphic spatial–temporal feature extraction," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2025, pp. 1–3.

[15] M. Yang et al., "Nanowatt acoustic inference sensing exploiting nonlinear analog feature extraction," *IEEE J. Solid-State Circuits*, vol. 56, no. 10, pp. 3123–3133, Oct. 2021.

[16] D. Wang, S. J. Kim, M. Yang, A. A. Lazar, and M. Seok, "Background noise and process-variation-tolerant sub-microwatt keyword spotting hardware featuring spike-domain division-based energy normalization," *IEEE J. Solid-State Circuits*, vol. 60, no. 2, pp. 685–694, Feb. 2025.

[17] K. Kim et al., "A 23-μW keyword spotting IC with ring-oscillator-based time-domain feature extraction," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3298–3311, Nov. 2022.

[18] S. Ray and P. R. Kinget, "Ultra-low-power and compact-area analog audio feature extraction based on time-mode analog filterbank interpolation and time-mode analog rectification," *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1025–1036, Apr. 2023.

[19] A. Mostafa, E. Hardy, and F. Badets, "0.4 V 988nW time-domain audio feature extraction for keyword spotting using injection-locked oscillators," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2024, pp. 328–330.

[20] R. Guo et al., "A 5.1pJ/neuron 127.3us/Inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65nm CMOS," in *Proc. Symp. VLSI Circuits*, Jun. 2019, pp. C120–c121.

[21] 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 J. Solid-State Circuits*, vol. 55, no. 4, pp. 868–878, Apr. 2020.

[22] 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.

[23] 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.

[24] Z. Wang et al., "A 148-nW reconfigurable event-driven intelligent wake-up system for AIoT nodes using an asynchronous pulse-based feature extractor and a convolutional neural network," *IEEE J. Solid-State Circuits*, vol. 56, no. 11, pp. 3274–3288, Nov. 2021.

[25] W. Shan, J. Qian, L. Zhu, J. Yang, C. Huang, and H. Cai, "AAD-KWS: A sub-μW keyword spotting chip with an acoustic activity detector embedded in MFCC and a tunable detection window in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 58, no. 3, pp. 867–876, Mar. 2023.

[26] A. Kosuge, R. Sumikawa, Y.-C. Hsu, K. Shiba, M. Hamada, and T. Kuroda, "A 183.4nJ/inference 152.8μW single-chip fully synthesizable wired-logic DNN processor for always-on 35 voice commands recognition application," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2.

[27] H. Yang et al., "A 1.5-μW fully-integrated keyword spotting SoC in 28-nm CMOS with skip-RNN and fast-settling analog frontend for adaptive frame skipping," *IEEE J. Solid-State Circuits*, vol. 59, no. 1, pp. 29–39, Jan. 2024.

[28] F. Tan, W.-H. Yu, K.-F. Un, R. P. Martins, and P.-I. Mak, "A 0.05-mm^2 2.91-nJ/decision keyword-spotting (KWS) chip featuring an always-retention 5T-SRAM in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 59, no. 2, pp. 626–635, Feb. 2024.

[29] C. Li et al., "A 0.61-μW fully integrated keyword-spotting ASIC with real-point serial FFT-based MFCC and temporal depthwise separable CNN," *IEEE J. Solid-State Circuits*, vol. 59, no. 3, pp. 867–877, Mar. 2024.

[30] J. Zhang et al., "ANP-I: A 28-nm 1.5-pJ/SOP asynchronous spiking neural network processor enabling sub-0.1-μJ/sample on-chip learning for edge-AI applications," *IEEE J. Solid-State Circuits*, vol. 59, no. 8, pp. 2717–2729, Aug. 2024.

[31] S. Mourrane, B. Larras, S. Clerc, A. Cathelin, and A. Frappé, "A sub-400-nW real-time event-driven spectrogram extraction unit in 28-nm FD-SOI CMOS for keyword spotting application," *IEEE J. Solid-State Circuits*, vol. 60, no. 6, pp. 2060–2071, Jun. 2025.

[32] S. Park et al., "A 5.6μW 10-keyword end-to-end keyword spotting system using passive-averaging SAR ADC and sign-exponent-only layer fusion with 92.7% accuracy," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2024, pp. 1–2.

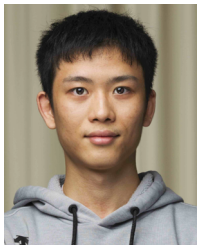
[33] F. Tan, W.-H. Yu, J. Lin, K.-F. Un, R. P. Martins, and P.-I. Mak, "A 1.8% FAR, 2 ms decision latency, 1.73 nJ/decision keywords-spotting (KWS) chip incorporating transfer-computing speaker verification, hybrid-IF-domain computing and scalable 5T-SRAM," *IEEE J. Solid-State Circuits*, vol. 60, no. 3, pp. 1103–1112, Mar. 2025.

[34] H.-J. Lee, K. Pyo, T. Jang, M. Seok, and S. Cho, "A 13.5μW 35-keyword end-to-end keyword spotting system featuring personalized on-chip training in 28nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2025, pp. 620–622.

[35] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. Interspeech*, Sep. 2019, pp. 814–818.

[36] D. Kadetotad, S. Yin, V. Berisha, C. Chakrabarti, and J.-S. Seo, "An 8.93 TOPS/W LSTM recurrent neural network accelerator featuring hierarchical coarse-grain sparsity for on-device speech recognition," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1877–1887, Jul. 2020.

[37] T. Tambe et al., "A 16-nm SoC for noise-robust speech and NLP edge AI inference with Bayesian sound source separation and attention-based DNNs," *IEEE J. Solid-State Circuits*, vol. 58, no. 2, pp. 569–581, Feb. 2023.
[38] Y.-H. Tsai et al., "A 28-nm 1.3-mW speech-to-text accelerator for edge AI devices," *IEEE J. Solid-State Circuits*, vol. 59, no. 11, pp. 3816–3826, Nov. 2024.
[39] T. Tambe et al., "A 12nm 18.1TFLOPs/W sparse transformer processor with entropy-based early exit, mixed-precision predication and fine-grained power management," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2023, pp. 342–344.
[40] B. Keller et al., "A 95.6-TOPS/W deep learning inference accelerator with per-vector scaled 4-bit quantization in 5 nm," *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1129–1141, Apr. 2023.
[41] F.-G. Zeng et al., "Speech dynamic range and its effect on cochlear implant performance," *J. Acoust. Soc. Amer.*, vol. 111, no. 1, pp. 377–386, Jan. 2002.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[44] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, 2014, pp. 1–9.
[45] S.-C. Liu, S. Zhou, Z. Li, C. Gao, K. Kim, and T. Delbruck, "Bringing dynamic sparsity to the forefront for low-power audio edge computing: Brain-inspired approach for sparsifying network updates," *IEEE Solid State Circuits Mag.*, vol. 16, no. 4, pp. 62–69, Apr. 2024.
[46] *Ultra-low Current, Low-Noise Microphone with Analog Output*, InvenSense Inc., San Jose, CA, USA, Dec. 2014.
[47] R. F. Lyon, *Human and Machine Hearing: Extracting Meaning From Sound*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[48] R. Sarpeshkar et al., "An ultra-low-power programmable analog bionic ear processor," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 4, pp. 711–727, Apr. 2005.
[49] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "A biomimetic, 4.5 μW, 120+ dB, log-domain cochlea channel with AGC," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 1006–1022, Mar. 2009.
[50] B. Wen and K. Boahen, "A silicon cochlea with active coupling," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 6, pp. 444–455, Dec. 2009.
[51] G. Yang, R. F. Lyon, and E. M. Drakakis, "A 6-μW per channel analog biomimetic cochlear implant processor filterbank architecture with across channels AGC," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 1, pp. 72–86, Jan. 2015.
[52] I. Kiselev, C. Gao, and S.-C. Liu, "Spiking cochlea with system-level local automatic gain control," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 5, pp. 2156–2166, May 2022.
[53] H. Chandrakumar and D. Marković, "A 15.2-ENOB 5-kHz BW 4.5-μW chopped CT $\Delta\Sigma$-ADC for artifact-tolerant neural recording front ends," *IEEE J. Solid-State Circuits*, vol. 53, no. 12, pp. 3470–3483, Dec. 2018.
[54] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, vol. 46, pp. 605–626, Nov. 2017.
[55] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proc. IEEE*, vol. 109, no. 2, pp. 124–148, Feb. 2021.
[56] J. Hu, Z. Zhang, C. S. Leow, W. L. Goh, and Y. Gao, "LearnAFE: Circuit-algorithm co-design framework for learnable audio analog front-end," *IEEE Trans. Circuits Syst. I, Reg. Papers*, early access, Jun. 19, 2025, doi: 10.1109/TCSI.2025.3578606.
[57] D. Neil, J. H. Lee, T. Delbrück, and S. Liu, "Delta networks for optimized recurrent network computation," in *Proc. 34th Int. Conf. Mach. Learn.*, Aug. 2017, pp. 2584–2593.
[58] C. Gao, A. Rios-Navarro, X. Chen, S.-C. Liu, and T. Delbruck, "EdgeDRNN: Recurrent neural network accelerator for edge inference," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 4, pp. 419–432, Dec. 2020.
[59] Q. Chen et al., "DeltaKWS: A 65nm 36nJ/decision bio-inspired temporal-sparsity-aware digital keyword spotting IC with 0.6 V near-threshold SRAM," *IEEE Trans. Circuits Syst. Artif. Intell.*, vol. 2, no. 1, pp. 79–87, Mar. 2025.
[60] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5220–5224.
[61] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[62] P. Harpe, "A compact 10-b SAR ADC with unit-length capacitors and a passive FIR filter," *IEEE J. Solid-State Circuits*, vol. 54, no. 3, pp. 636–645, Mar. 2019.
[63] S. Zhou, X. Chen, K. Kim, and S.-C. Liu, "High-accuracy and energy-efficient acoustic inference using hardware-aware training and a 0.34nW/Ch full-wave rectifier," in *Proc. IEEE 5th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2023, pp. 1–5.
[64] D. Djekic, G. Fantner, K. Lips, M. Ortmanns, and J. Anders, "A 0.1% THD, 1MΩ-to-1GΩ 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.
[65] L. Shen, N. Lu, and N. Sun, "A 1-V 0.25-μW inverter stacking amplifier with 1.07 noise efficiency factor," *IEEE J. Solid- State Circuits*, vol. 53, no. 3, pp. 896–905, Mar. 2018.
[66] H. Chandrakumar and D. Markovic, "A simple area-efficient ripple-rejection technique for chopped biosignal amplifiers," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 2, pp. 189–193, Feb. 2015.
[67] K. Laker and W. Sansen, *Design of Analog Integrated Circuits and Systems* (Electrical Engineering Series). New York, NY, USA: McGraw-Hill, 1994, pp. 858–862.
[68] K. Kim and S.-C. Liu, "Continuous-time analog filters for audio edge intelligence: Review on circuit designs [feature]," *IEEE Circuits Syst. Mag.*, vol. 23, no. 2, pp. 29–48, 2023.
[69] 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.
[70] 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.
[71] J. S. P. Giraldo, S. Lauwereins, K. Badami, H. Van Hamme, and M. Verhelst, "18μW SoC for near-microphone keyword spotting and speaker verification," in *Proc. Symp. VLSI Circuits*, Jun. 2019, pp. C52–C53.
[72] Q. Zhang et al., "SPRSound: Open-source SJTU paediatric respiratory sound database," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 5, pp. 867–881, Oct. 2022.
[73] N. Babu, J. Kumari, J. Mathew, U. Satija, and A. Mondal, "Multiclass categorisation of respiratory sound signals using neural network," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2022, pp. 228–232.
[74] N. Zeghidour, O. Teboul, F. D. C. Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16.
[75] Q. Fu, Z. Teng, J. White, M. E. Powell, and D. C. Schmidt, "FastAudio: A learnable audio front-end for spoof speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3693–3697.



Sheng Zhou (Graduate Student Member, IEEE) received the bachelor's degree in computer science from The Hong Kong University of Science and Technology, Hong Kong, and the master's degree in data science from ETH Zürich, Zürich, Switzerland. He is currently pursuing the Ph.D. degree with the Sensors Group, Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich.

His research interests include mixed-signal circuit design for ultra-low-power edge applications.

Zixiao Li (Graduate Student Member, IEEE) received the bachelor's degree in information science and engineering from Southeast University, Nanjing, China, and the master's degree in electrical engineering and information technology from ETH Zürich, Zürich, Switzerland. He is currently pursuing the Ph.D. degree with the Sensors Group, Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich.

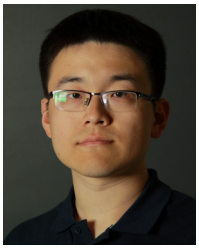
His research interests include neural network acceleration for inference and training, and ASIC design.

Longbiao Cheng (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2022.

He conducted his Ph.D. research at the Institute of Acoustics, Chinese Academy of Sciences, Beijing, where he was also a Post-Doctoral Researcher from July to December 2022. He is currently a Post-Doctoral Researcher with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland. His research interests include compute-efficient neural networks, multichannel audio signal processing, and 3-D audio, with a focus on real-time processing for edge platforms.

Jérôme Hadorn received the bachelor's degree in computer science with a minor in neuroinformatics from the University of Zurich, Zürich, Switzerland. He is currently pursuing the master's degree with the University of Zürich and ETH Zürich, Zürich, in the joint Neural Systems and Computation Program.

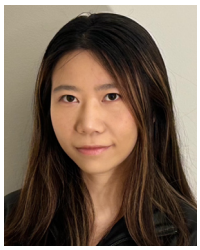
He is currently working on his master's thesis with the Sensors Group, Institute of Neuroinformatics, University of Zürich and ETH Zürich.



Chang Gao (Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong-Liverpool University, Suzhou, China, in 2015, the M.Sc. degree from Imperial College London, London, U.K., in 2016, and the Ph.D. degree (Hons.) in neuroscience from the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, in 2022.

Since August 2022, he has been an Assistant Professor with the Microelectronics Department, Delft University of Technology, Delft, The Netherlands, where he leads the Laboratory of Efficient Machine Intelligence. His research focuses on neuromorphic algorithm-hardware co-design for edge AI, with applications in audio, vision, and wireless communications.

Dr. Gao received the 2022 Misha Mahowald Early Career Award, the Marie Curie Postdoctoral Fellowship, and the 2023 NWO Veni grant. He was named an MIT Technology Review Innovator Under 35 Europe in 2023.



Qinyu Chen (Member, IEEE) received the Ph.D. degree in electronic science and technology from Nanjing University, Nanjing, China, in 2021.

She is an Assistant Professor at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands. She was a Post-Doctoral Researcher at the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, from 2022 to 2024. Her research interests include the seamless brain-inspired AI system at the edge, and its application in healthcare, AR/VR with a focus on event-based processing.

Dr. Chen received the 2022 SNSF and Innosuisse Bridge Fellowship and the 2024 NWO Veni Talent Program Grant.

Tobi Delbruck (Fellow, IEEE) received the B.Sc. degree in physics from the University of California, San Diego, CA, USA, in 1986 and the Ph.D. degree from Caltech, Pasadena, CA, USA, in 1993.

Since 1998, he has been a Professor of physics and electrical engineering with the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland. He co-directs the Sensors Group, Institute of Neuroinformatics. His group's research currently focuses on neuromorphic sensory processing, efficient hardware AI, and neural control.

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

He is an Assistant Professor with the Department of Electronics and Nanoengineering, School of Electrical Engineering, Aalto University, Espoo, Finland. He was a Visiting Student in 2020 and a Post-Doctoral Researcher from 2021 to 2023 at the Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland, and an Established Researcher at the Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, from 2023 to 2024. His research interests include analog/mixed-signal ICs and full-custom memory ICs for neuromorphic signal processing, biomedical sensors, and in-memory computing.

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

She is currently an Adjunct Professor with the Faculty of Science, University of Zürich, Zürich, Switzerland. She co-directs the Sensors Group, Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich. Her group's research focuses on sensor integrated circuit designs, including the spiking silicon cochlea and bio-inspired auditory sensors; and real-time energy-efficient hardware systems that combine both sensor and event-driven low-compute deep neural network algorithms, targeting always-on edge AI, and wearable applications.