13.1 A 0.22mm^2 161nW Noise-Robust Voice-Activity Detection Using Information-Aware Data Compression and Neuromorphic Spatial-Temporal Feature Extraction

Ying Liu*[1], Jie Li*[1], Qining Zhang*[1], Tianhao Zhao[2], Chenhao Shi[1], Ninghui Shang[1], Peiyu Chen[2], Xiaohuan Ge[3], Yufei Ma[1], Linxiao Shen[1], Zhixuan Wang[3], Ru Huang[1], Le Ye[1,2]

[1]Peking University, Beijing, China
[2]Advanced Institute of Information Tecnhology of Peking University, Hangzhou, China
[3]Nano Core Chip Electronic Technology, Hangzhou, China

*Equally Credited Authors (ECAs)

2025 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3315-4101-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISSCC49661.2025.10904569

Nowadays, voice activation detection (VAD), typically consisting of the feature extractor (FE) and the intelligent engine (IE), is crucial for reducing the power consumption of the voice processing system (VPS) (Fig. 13.1.1 top). Normally, the always-on VAD dominates the power consumption while VPS remains inactive [1]. Therefore, the VAD should meet stringent power consumption requirement to extend the battery life of artificial intelligence of things (AIoT) devices. Additionally, achieving excellent inference accuracy and noise robustness is also essential for practical consideration. [1] proposed the analog FE and exploited binary NNs (BNNs), consuming 1μW power but with only 85% accuracy. [2] presented the analog convolutional neural networks (CNNs) to reduce the power of IEs (108nW), but it only achieved about 90% accuracy in high signal-noise ratio (SNR) (>10dB) scenarios. Recently, the bio-inspired spike-based methods [4-5] have shown promising ultra-low-power (ULP) and intelligent potentials across various scenarios. [5] converted Mel-frequency cepstral coefficients (MFCC) features into spikes and trained two fully-connected spiking NNs (SNNs) layers, achieving >90% VAD accuracy across 0~5dB SNR. However, the spikes generation from the static MFCC induces excessive power consumption. [5] demonstrated 90% accuracy for ECG classification with 82nW power consumption. However, the level-crossing coding in [4] will generate more spikes when processing the voice whose frequency is much higher than ECG, resulting in more power consumption. To the best knowledge, a VAD system achieving ultra-low power consumption, excellent accuracy and noise robustness has not yet been presented.

To address this problem, the VAD system in this work (Fig. 13.1.1) introduces an information-aware data compressor (IADC) and a neuromorphic spatial-temporal feature extractor (NSTFE). 1) The IADC compresses the raw voice into spatial-temporal spikes by only filtering out the extreme points (EPs). The generated spikes are coded into 5-bit address-event-representations (AERs) to trigger the NSTFE. As is shown in Fig. 13.1.1(bottom), the straight line between two EPs closely approximates the raw signal at the frequency of the voice. The signal fitted by EPs achieves in average 97% cosine similarity with raw voice across 0~15dB SNR, while compressing the data by 2.6 times, efficiently reducing the power consumption. The implementation of IADC includes the analog extreme point compressor (EPC) and information-aware converter (IAC). The EPC filters out redundant data and only triggers IAC to generate AER where EP appears, thereby significantly reducing the power consumption of the IAC. 2) The NSTFE [6] is employed to further extract and compress spatial-temporal information with trainable weights. Compared with [2], the neurons in NSTFE are integrated with different weights according to the received AERs, replacing multiplication with low-power accumulation while maintaining spatial information. With trainable weights, the NSTFE enables spatial-temporal information extraction from the position and order of EPs. Finally, all EPs are compressed into 32 membranes of integrate neurons. Thanks to the efficient extraction and compression, the required size of ANN is reduced, which saves power and area, simultaneously achieving excellent accuracy.

Typically, a low noise amplifier (LNA) is considered to amplify the input signal in the analog front end (AFE) [1-3]. However, the linearity and noise requirements of the LNA induce intensive power consumption. In the IADC, a higher-resolution ADC (8-bit) replaces LNA to directly convert small signals. At 4kHz frequency and 8-bit resolution, the power consumption of ADCs is mainly determined by the digital logic. Thus the 5-8bit resolution increment induces only modest power overhead (4nA in simulation), while providing equivalent gain for small inputs. The 8-bit output is then downscaled to 5-bit AER code, optimizing data size for subsequent processing while retaining information. The 2-bit code is used to determine the scale-down strategy, enabling configurable gain. Moreover, to reduce the power consumption of the 8-bit ADC, the EPC is realized in analog domain to trigger the ADC only when an EP is identified, thereby reducing unnecessary ADC conversions. The circuit implementation of IADC is shown in Fig. 13.1.2. The differential input signals are alternately sampled on two CDAC pairs, with Comparator 1 triggered in each cycle to check the sign of $V[n] – V[n-1]$ ($V[n]$ represents the nth sampled voltage). The sign change of $V[n]-V[n-1]$ indicates an EP at $V[n-1]$, triggering the ADC to convert $V[n-1]$. To minimize the offset of Comparator 1, ensuring accurate detection, an additional input pair is added to calibrate the mismatch. Furthermore, the background calibration is designed to compensate for the offset. An 8-bit SAR ADC is used to convert the detected EPs. With this implementation, if no EP is detected, the scheme requires only one comparison of Comparator 1, compared to eight comparisons of SAR ADC if EPC is realized digitally, making analog domain EPC more power-efficient under sparse EP scenarios. After finishing EP conversion, the IADC generates asynchronous hand-shake signals with NSTFE to send 5-bit AER, which is downscaled from the output of 8-bit ADC.

Figure 13.1.3 (top) illustrates the control flow of the NSTFE, including an IF layer and an integrate layer. Initially, it remains idle until an input AER from IADC appears. To minimize the clock-tree power, the input AER triggers the IF layer using asynchronous valid-ready shaking, which enables the spiking processing logic to integrate the corresponding weights into the membranes. The neurons whose membranes exceed the threshold will fire spikes, then encoded as 5-bit AERs to trigger the integrate layer. The dichotomy is used to optimize the latency of AER generation. The middle part of Fig. 13.1.3 shows an example. 8 neurons are divided into two groups which are checked sequentially by “or” logic. The group is skipped directly if the result of “or” is 0. Otherwise, the group is further divided into two subgroups, and spike detection is repeated. The division and detection continue until each group contains only two neurons, at which point a unique AER is determined. The latency of AER generation is efficiently reduced to 1 cycle. The architecture of NSTFE and ANN engine is shown in the bottom of Fig. 13.1.3. To minimize area cost, the NSTFE and ANN reuse the memory and logic resource. The 2KB global SRAM stores the trained 4-bit weights of both NSTFE and ANN. The 96B buffers are reused for activations in the ANNs or membranes in the NSTFE. The 32 processing units (PU) work in parallel to perform MAC or integration operation. The multipliers in PUs are gated to save power during NSTFE execution. The non-linear function includes firing logic, ReLU logic and decision logic. The decision logic generates the final inference result.

Figure 13.1.4 (top) illustrates the measured output codes of the IADC in both normal ADC mode (w/o EPC) and compression mode (w/ EPC). In compression mode, the generated extreme points are well fitted with the raw voice signal (10dB SNR), while effectively reducing unnecessary data. The IADC spectrum is measured by using sine signal and shown in the middle and bottom of Fig. 13.1.4, demonstrating that the different scale-down strategies can maximize the SNDR of IADC for different input levels, equivalent to the configurable amplification gain of input. 22.6nW total power consumption is measured at 0.8V voltage and 4kHz frequency (Fig. 13.1.4 middle right). The power consumption across different SNR voice inputs is also provided (Fig. 13.1.4 bottom right).

To show the practical use of the VAD system presented in this paper, this work considers 0~15dB SNR and background noise of the QUT-NOISE-TIMIT dataset from 10 scenarios. Fig. 13.1.5 (top) shows the measurement result of overall hit rate and confusion matrix. The 84% overall hit rate is achieved in the worst case (0dB SNR). 90%, 94% and 98% accuracy are achieved in 5dB, 10dB, 15dB SNR respectively. Compared with state-of-the-arts (SOTAs), this work shows the performance in 0dB SNR and realize higher VAD accuracy in the same SNR level. The measured total 161nW power consumption (Fig. 13.1.5 bottom) shows the ULP characteristic. Furthermore, this work achieves excellent overall performance by considering both accuracy and power consumption. The ability to realize the keywords spotting (KWS) is verified on Hey Snips dataset with 94% accuracy measured.

Figure 13.1.7 shows the micrograph of the chip and summary table. The chip was fabricated in 55nm CMOS technology. The IADC works at 0.8V. The NSTFE and ANNs work at 0.67V. Figure 13.1.6 compares the AFE and chip with SOTAs respectively. For the AFE, this work saves power for efficiently compressing data, showing lower power (23nW) and smaller area (0.05mm^2) metrics among AFEs of SOTAs. For the VAD system, the 0.22mm^2 core area is smaller than other works except for [1]. However, the 55nm chip area of this work is only slightly larger than [1], fabricated in 28nm. Based on practicality and response latency considerations, this work employs a 160ms inference window to avoid the pauses between two adjacent voices being treated as noise. This chip demonstrates the VAD performance across 10 different background noise scenarios. Due to the full utilization of temporal-spatial information, the chip achieves excellent performance under different SNR scenarios. Additionally, our chip shows 84% accuracy at 0dB SNR. The 100kHz VAD system runs at 161nW ultra-low power consumption. In contrast, [2] and [3] also show good power consumption. However, [3] requires a 512ms inference window, which results in more latency. [2] does not include the power overhead required by the more complex clock generation.

Acknowledgement:
This work was supported in part by NSFC Grant 92164301, Grant 62225401; Zhejiang Provincial Key R&D program under Grant 2021C01035; Grant QYJS-2023-2401-B, and Grant QYJS-2023-2402-B. Corresponding authors: Linxiao Shen and Zhixuan Wang (linxiao.shen@pku.edu.cn, zhixuanwang_ee@outlook.com)

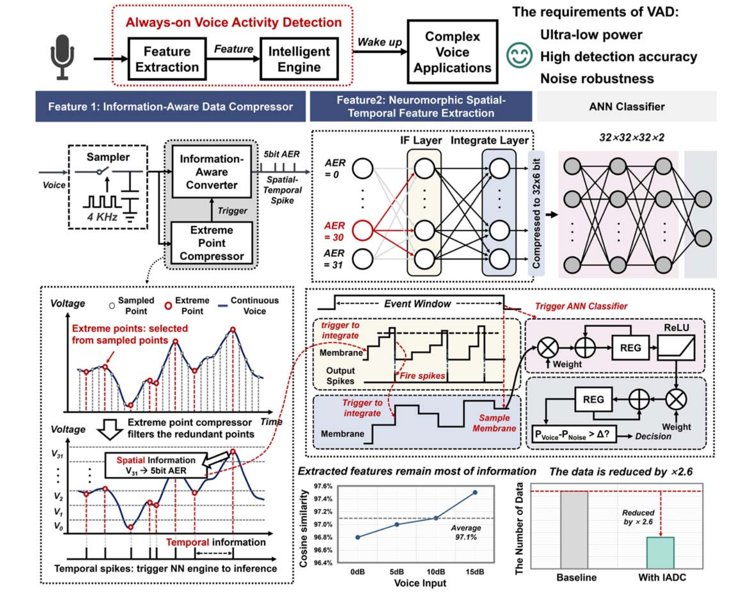


Figure 13.1.1: The architecture of VAD (top). The comparison between extracted feature and raw voice (bottom).

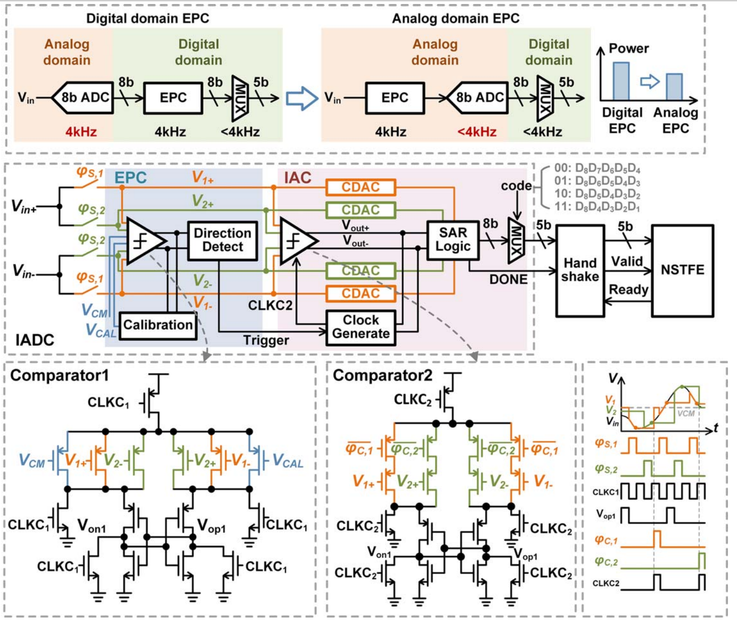
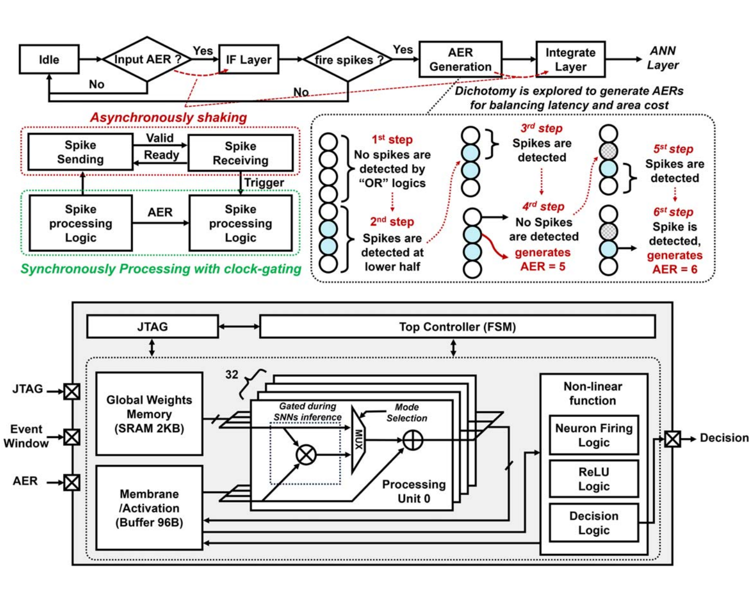


Figure 13.1.2: The architecture (top) and circuit implementation (middle) of IADC.



Idle
Input AER ?
Yes
No
IF Layer
fire spikes ?
AER Generation
Integrate Layer
ANN Layer
Dichotomy is explored to generate AERs for balancing latency and area cost
Asynchronously shaking
Spike Sending
Valid
Ready
Spike Receiving
Trigger
Spike processing Logic
AER
Synchronously Processing with clock-gating
1st step No spikes are detected by "OR" logics
2nd step Spikes are detected at lower half
3rd step Spikes are detected
4rd step No Spikes are detected generates AER = 5
5st step Spikes are detected
6st step Spike is detected, generates AER = 6
JTAG
Top Controller (FSM)
Event Window
Global Weights Memory (SRAM 2KB)
Membrane /Activation (Buffer 96B)
32
Gated during SNNs inference
Mode Selection
MUX
Processing Unit 0
Non-linear function
Neuron Firing Logic
ReLU Logic
Decision Logic
Decision

Figure 13.1.3: The control flow of NSTFE (top). The architecture of IE (bottom).

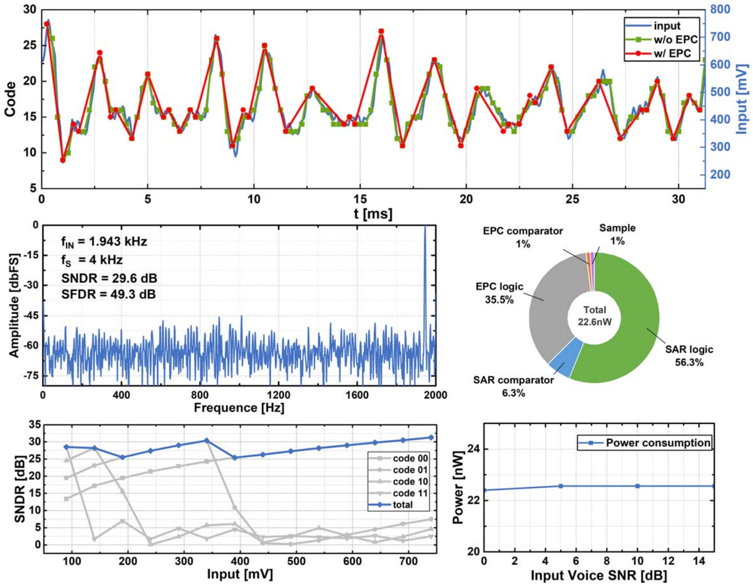


Figure 13.1.4: IADC outputs (EPC on/off), output spectra, SNDR vs input, power breakdown, and power vs SNR.

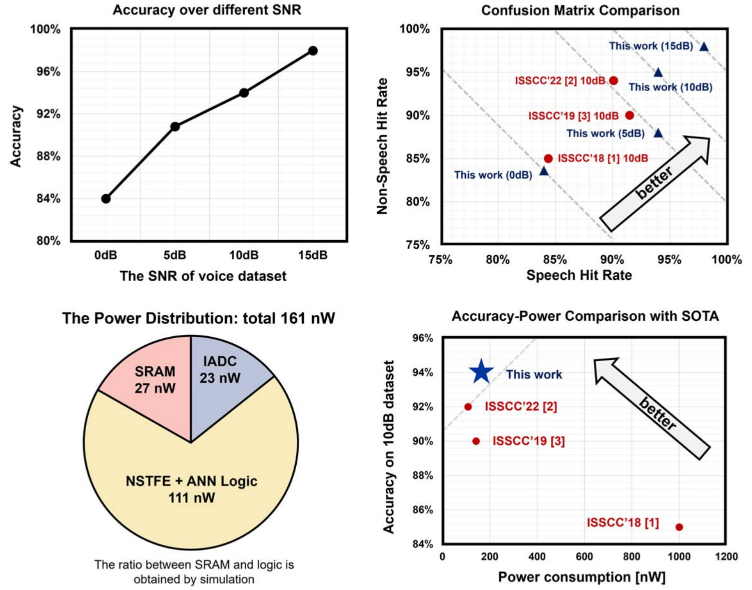


Figure 13.1.5: Accuracy and power breakdown (left). Comparison of confusion matrix and accuracy-power(right).

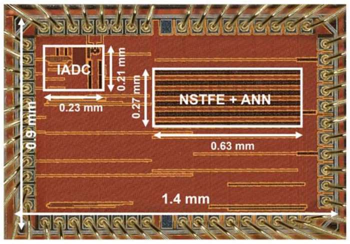
Analog Front End	ISSCC'2022 [2]	ISSCC'19 [3]	ISSCC'18 [1]	This Work
Technology	28 nm	180 nm	180 nm	55 nm
Feature Extractor	Time-Domain CNN	Mixer-based AFE	Analog-to-Event Filter Bank	Information-Aware Data Compressor
Area	0.055 mm^2	0.56 mm^2	1.6 mm^2	0.05 mm^2
Dynamic Range	N. A.	47 dB	40 dB	49 dB
Power Consumption	73 nW	60 nW	380 nW	23 nW
w/ Compression	×	×	×	✓
FoM[a]	1	1	1	2.52

Chip	ISSCC'2022 [2]	ISSCC'2022 [4]	ISSCC'19 [3]	ISSCC'18 [1]	This Work
Task	VAD	ECG Classification	VAD	VAD	VAD
Dataset	TIMIT-NOISEX-92	MIT-BIH Arrhythmia	LibriSpeech data + NOISEX-92	AURORA4 + DEMAND	QUT-NOISE-TIMIT
Inference Window	10 ms	1s	512 ms	10ms	160 ms
Core Area	0.16 mm^2	> 3 mm^2	> 12 mm^2	N.A.	0.22 mm^2
Feature Extractor	Time-Domain CNN	Level-Crossing Sampling	Mixer-based AFE	Analog-to-Event Filter Bank	Information-Aware Data Compressor + Neuromorphic Feature Extractor
Inference Engine	BNN	SNN	NN	BNN	NN
Power Consumption	108 nW	<350 nW	142 nW	1 μW	161 nW
The # of Noise Scenarios	N.A.	N.A.	1	2	10
Overall Hit Rate	<86% @ 4dB SNR 92.5% @ 10dB SNR < 93% @ 16dB SNR	90.5%	78% @ 5dB SNR 91% @ 10dB SNR 96.5% @ 20dB SNR	85% @10dB SNR	84% @ 0dB SNR[b] 90% @ 5dB SNR 94% @ 10dB SNR 98% @ 15dB SNR
Speech /Non-Speech Hit Rate	90.1%/94% @ 10dB	N.A.	91.5%/90% @ 10dB	84.4%/85% @ 10dB SNR	94%/95% @ 10dB SNR

[a] FoM = (1 - information loss ratio)×Compression Ratio. The AFEs of [1-3] do not compress any data, thus the information loss equals 0 and compression ratio equals 1.

[b] The VAD accuracy at 0dB SNR is firstly shown

Figure 13.1.6: The comparison table of analog front end (top) and chip (bottom).



Process [nm]	CMOS 55nm
Core Area [mm^2]	0.22
Voltage [V]	Analog @ 0.8
	NSTFE + ANN @ 0.67
Accuracy	98% @ 15dB SNR
	94% @ 10dB SNR
	90% @ 5dB SNR
	84% @ 0dB SNR
Power [nw]	161 nW

Figure 13.1.7: The micrograph (left) and summarization (right) of chip.

References:

[1] M. Yang et al., “A 1μW voice activity detector using analog feature extraction and digital deep neural network,” *ISSCC*, pp. 346-347, Feb. 2018.

[2] F. Chen et al., “A 108nW 0.8mm^2 Analog Voice Activity Detector (VAD) Featuring a Time-Domain CNN as a Programmable Feature Extractor and a Sparsity-Aware Computational Scheme in 28nm CMOS,” *ISSCC*, pp. 1-3, Feb. 2022.

[3] M. Cho et al., “A 142nW Voice and Acoustic Activity Detection Chip for mm-Scale Sensor Nodes Using Time-Interleaved Mixer-Based Frequency Scanning,” ISSCC, pp. 277-278, Feb. 2019.

[4] Y. Liu et al., “An 82nW 0.53pJ/SOP Clock-Free Spiking Neural Network with 40μs Latency for AIoT Wake-Up Functions Using Ultimate-Event-Driven Bionic Architecture and Computing-in-Memory Technique,” *ISSCC*, pp. 372-374, Feb. 2022.

[5] F. Martinelli et al., “Spiking Neural Networks Trained with Backpropagation for Low Power Neuromorphic Implementation of Voice Activity Detection,” *ICASSP*, pp. 8544-8548, 2020.

[6] Y. Liu et al., “Sparsity-Aware In-Memory Neuromorphic Computing Unit with Configurable Topology of Hybrid Spiking and Artificial Neural Network,” *TCAS I*, vol. 71, no. 6, pp. 2660-2673, June 2024.

979-8-3315-4101-9/25/$31.00 ©2025 IEEE