An Acoustic Signal Processing Chip With 142-nW Voice Activity Detection Using Mixer-Based Sequential Frequency Scanning and Neural Network Classification

Sechang Oh, *Member, IEEE*, Minchang Cho, *Student Member, IEEE*, Zhan Shi, Jongyup Lim, *Student Member, IEEE*, Yejoong Kim, Seokhyeon Jeong, *Student Member, IEEE*, Yu Chen, Rohit Rothe, *Student Member, IEEE*, David Blaauw, *Fellow, IEEE*, Hun-Seok Kim, *Member, IEEE*, and Dennis Sylvester, *Fellow, IEEE*

***Abstract*— This article presents a voice and acoustic activity detector that uses a mixer-based architecture and ultra-low-power neural network (NN)-based classifier. By sequentially scanning 4 kHz of frequency bands and down-converting to below 500 Hz, feature extraction power consumption is reduced by 4×. The NN processor employs computational sprinting, enabling 12× power reduction. The system also features inaudible acoustic signature detection for intentional remote silent wakeup of the system while re-using a subset of the same system components. The measurement results achieve 91.5%/90% speech/non-speech hit rates at 10-dB SNR with babble noise and 142-nW power consumption. Acoustic signature detection consumes 66 nW, successfully detecting a signature 10 dB below the noise level.**

***Index Terms*— Acoustic event detection, acoustic wakeup detection, audio signal processing, deep neural network, feature extraction, Internet of Things, machine learning, ultra-low power (ULP), voice activity detection.**

I. Introduction

VOICE user interfaces are widely adopted in various devices as the human voice is one of the most natural and information-rich interfaces between humans and machines. Minimizing the power consumption of voice processing is particularly crucial to meet power budgets when the system becomes smaller as the battery size imposes severe power and energy constraints on system design [1]. In many practical applications, acoustic events-of-interest occur infrequently. Constant listening and detection of keywords are very power-hungry. Instead, the use of an always-on voice activity detector (VAD) as a system wakeup mechanism is a popular alternative [2]–[7], and subsequent power-hungry processing is enabled by the VAD to save overall system power, as shown in Fig. 1. The acoustic wakeup detector consumes much less power than constant listening for keywords since it only detects whether an incoming signal contains a human voice. However, since the events occur infrequently, the always-on acoustic wakeup detector typically dominates the system power consumption, and therefore, minimizing the VAD power consumption itself is a critical design challenge.

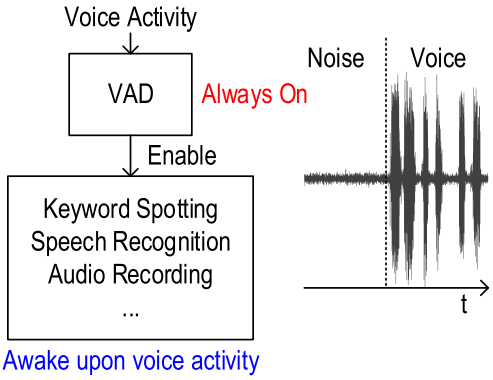


Fig. 1. Always-on voice activity detection as a wakeup mechanism. Advanced processing is enabled upon voice activity detection to save the overall power.

A previous acoustic wakeup detector [8] consumes just 12 nW but it is specifically designed to detect "stationary" events whose signal features are invariant over a relatively long time (a few seconds) and very narrow in frequency (2-Hz bandwidth or 0.5-s extraction time for each feature). The approach in [8] is not applicable to a non-stationary target, such as voice activity containing time-varying features that need to be extracted with a short (tens of ms) interval. Prior VAD chips [2], [3] demonstrated reliable performance but consumed significant power (>20 μW) and lacked an analog front end (AFE), which would further increase the power. More recent analog-domain feature extraction-based VAD chips [4], [5] also reported μW-level power consumption, and their simple decision tree [4] or fixed neural network (NN)-based approach [5] limited broader use for various

Manuscript received May 3, 2019; revised July 22, 2019; accepted August 8, 2019. Date of publication September 12, 2019; date of current version October 23, 2019. This article was approved by Associate Editor Sriram Vangal. This work was supported by the Defense Advanced Research Projects Agency (DARPA) N-Zero Program. *(Sechang Oh and Minchang Cho are co-first authors.)*

S. Oh, M. Cho, Z. Shi, J. Lim, Y. Chen, R. Rothe, D. Blaauw, H.-S. Kim, and D. Sylvester are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: chaseoh@umich.edu; mincho@umich.edu).

Y. Kim and S. Jeong are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48109 USA, and also with Cube Works Inc., Ann Arbor, MI 48109 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2019.2936756

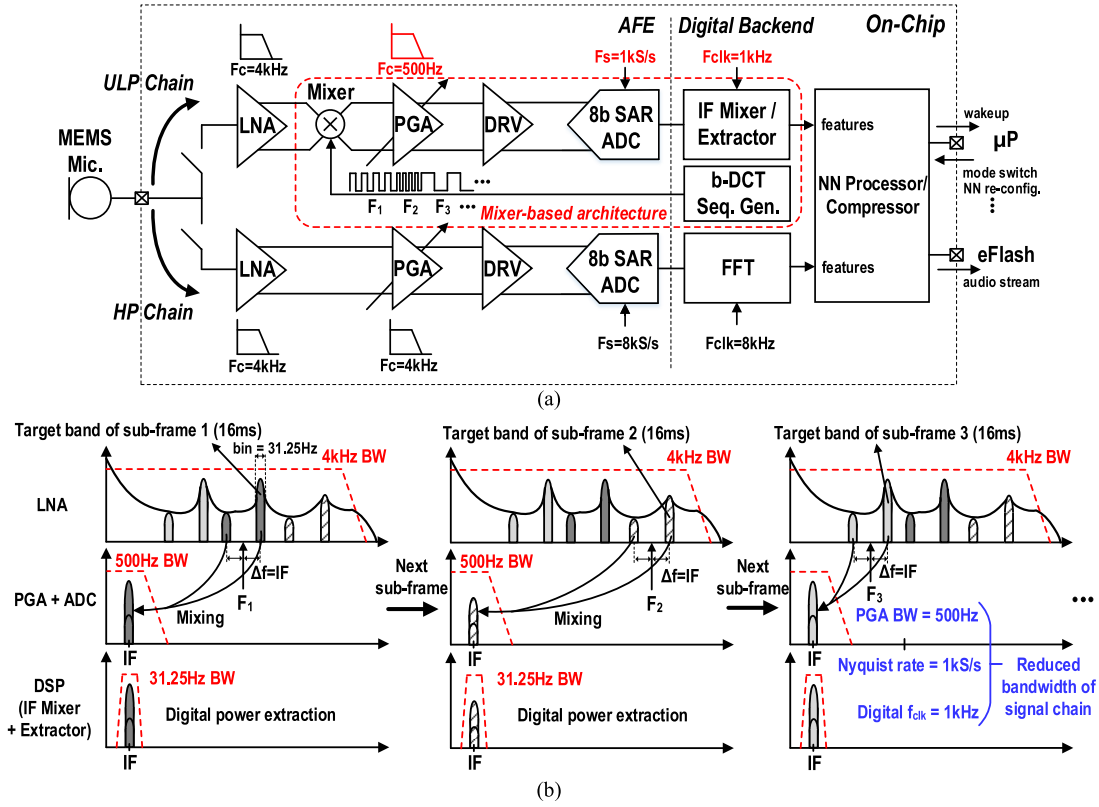


Fig. 2. (a) VAD system architecture. (b) Operating principle of mixer-based sequential frequency scanning.

acoustic event targets. Moreover, the VAD chips [2]–[5] were tested using only electric analog audio signals, rather than actual audio signals. Therefore, additional components and their power consumption overhead need to be added for real audio to electric analog signal conversion.

Typical VADs consist of two parts [9]–[11]. A feature extractor which converts the incoming signal into low-dimensional but dense acoustic features, and a classifier that takes a feature set input and produces a binary decision: Speech or non-speech. Both design of feature extractor and classifier significantly affect overall system power, accuracy, latency, and scalability.

The main challenge in reducing the overall VAD power is to reduce the power required for feature extraction since it is typically computation-intensive and operates continuously without duty cycling. Conventional approaches [2], [3], [11] used digital fast Fourier transform (FFT)-based feature extraction, yet FFT itself consumes >2 μW even with extensively relaxed throughput/latency constraints [2]. To reduce power consumption, [4], [5] exploited analog-domain feature extraction techniques. However, the parallel filter bank at the voice-band is still the most power-hungry block, preventing sub-μW operation. Instead of using parallel feature extraction, such as an analog filter bank or digital FFT, a serialized discrete Fourier transform (DFT) on tones-of-interest approach was introduced in our previous work [8] for low-frequency ($<$500 Hz) signal targets. However, applying the same technique to the voice-band (up to 4 kHz) frequency significantly increases the power consumption of both the AFE and the digital feature extractor proportionally with signal bandwidth, limiting the usefulness of this technique.

To improve the accuracy and scalability of the VAD system, the NN-based classifiers have been recently proposed [12]–[16]. Compared to other machine learning classifiers, such as decision tree [4] or support vector machine (SVM) [8], NN-based classifier have shown the improved performance [17], immunity to difficult noise scenarios [18], and strong scalability to multiple acoustic targets [19] and large-scale corpora [20], becoming a strong candidate for real-world applications.

This article presents a programmable acoustic signal processing system for both VAD and non-voice acoustic event detection based on NN classifier. We use a mixer-based architecture that sequentially scans and down-converts the 4-kHz bandwidth signal to a $\leq$500-Hz passband, reducing amplifier, analog-to-digital converter (ADC), and digital signal processor (DSP) power by 4$\times$. The NN processor employs computational sprinting, which minimizes static energy dominance in low frequency/voltage regimes, providing 12$\times$ power reduction in the digital domain. In addition to a VAD, the system features an inaudible acoustic signature detection mode to enable remote silent system wakeup. The proposed always-on VAD consumes 142 nW, which is 8$\times$ lower than that reported in the literature for state-of-the-art works. In this article, Section II describes the overview of the VAD system. Sections III and IV show its circuit implementation, and Section V discusses the measurement results. Finally, Section VI concludes this article.

II. VAD System Overview

Fig. 2(a) shows the overall system architecture with two signal chains: an always-on ultra-low power (ULP) chain and a high performance (HP) chain that wakes upon event detection by the ULP chain. The system has two modes based on the two chains: a 142-nW ULP mode and an 18-μW HP mode. The HP chain is power-gated in the ULP mode, while the ULP chain is always on. When a target event is detected in ULP mode, an off-chip microprocessor (μP) activates the HP mode, which enables more powerful feature extraction and classification to complete additional complex tasks at the cost of power consumption. The HP mode also supports real-time audio compressing and streaming to off-chip eFlash for general purpose post-processing [1]. The HP chain consists of 4-kHz bandwidth and 8-kS/s sampling rate with a conventional AFE architecture consisting of a low-noise amplifier (LNA), programmable amplifier (PGA), ADC driver (DRV), and ADC. In contrast, the ULP chain employs a digitally controlled mixer between the LNA and PGA to shift the desired signal frequency down to 500-Hz bandwidth to lower the Nyquist rate to 1 kS/s after the PGA. Both the ULP and HP chains share the same NN processor, but it operates with a different power scale and network model for each mode. An external clock from a 32-kHz crystal oscillator is divided into 8 and 1 kHz for the HP and ULP chain, respectively.

Fig. 2(b) shows the mixer-based sequential frequency scanning operation that reduces AFE and DSP power consumption in the ULP mode by lowering their bandwidth and sampling rate to 500 Hz and 1 kHz, respectively. The incoming signal from the microphone is amplified by an LNA with the full 4-kHz bandwidth. At this point, the mixer, switched by a binary discrete cosine transform (DCT) sequence, immediately down-converts the frequency of the desired feature to a programmable intermediate frequency (IF) of <500 Hz. The digital binary sequence generator supports an arbitrary DCT frequency for the mixer switch control; for example, the 4-kHz band can be divided into 31.25-Hz frequency bins using a 128-pt DCT, and the energy content of 32 bands out of 128 is sequentially extracted by sweeping the DCT frequencies ($F_1, F_2, \ldots F_{32}$). The 32 bands are chosen during NN training for each target event. The IF down-converted signal is further amplified and low-pass filtered with 500-Hz bandwidth (via a PGA) and digitized at 1 kS/s. Finally, the digital IF quadrature mixer down-converts the signal to dc, and feature power is measured. With a DCT length of 16 ms per feature (128-pt DCT with 8 kHz binary mixing), 32-feature extraction requires a 512-ms frame. The mixer-based structure reduces the bandwidth, sampling rate, and clock frequency of the AFE and DSP after the mixer; thus, the feature extraction power consumption is decreased from 225 nW (simulation; based on LNA and PGA at 4 kHz of bandwidth, DRV and ADC at 8 kS/s of sampling rate, and digital FFT at 8 kHz of clock) to 60 nW (measured; including LNA, PGA, DRV, ADC, IF mixer/extractor, and binary-DCT sequence generator). The programmable IF is set to ~250 Hz to reduce the PGA 1/f noise effect while the image aliasing issue of non-quadrature mixing and imperfection of first-order filtering is mitigated (without noticeable event detection accuracy degradation) by an NN trained with the image-aliased and attenuated signals.

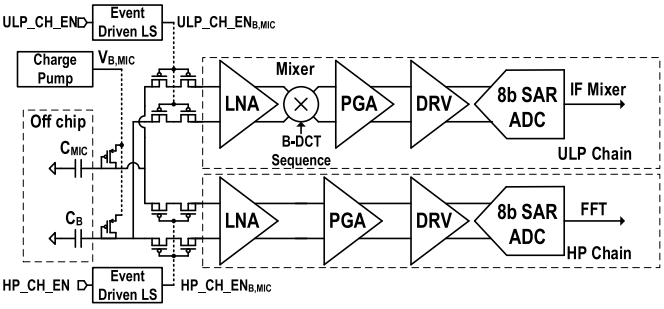


Fig. 3. AFE block diagram with ULP and HP chains.

III. Analog Front-End Implementation

A. Overall Architecture

Fig. 3 shows the AFE circuit diagram with ULP and HP chains. Both chains share a single MEMS capacitive microphone and a charge pump. Depending on the operation mode, the chain selection switches select one chain. The HP chain consists of a 31.3-dB gain LNA, 4.6–31.3-dB gain PGA, 8-bit ADC, and an ADC DRV. The ULP chain also includes all the blocks of the HP chain but operates with lower power consumption as it targets relaxed noise performance and ULP operation. Moreover, the ULP chain has a mixer between the LNA and PGA. The mixer is a passive mixer similar to a typical chopper and is controlled by the binary DCT sequence generator.

B. Charge Pump and 10-V Level Shifter

Microphone sensitivity is proportional to the microphone bias voltage, and therefore, we use a three-stage Dickson charge pump to generate 10-V bias [1]. Because the MEMS microphone is capacitive, the charge pump only needs to drive negligible loads. The charge pump uses the 8-kHz clock to minimize possible clock signal coupling to the signal chain (4-kHz BW) and consumes only 13 nW (measured). The diode-connected PMOS sets the corner frequency of the voltage bias to be well below the microphone response range (<75 Hz) to avoid altering the acoustic response in the system. C_B is an external capacitor to match the input impedance.

To switch the modes between ULP and HP, level shifters shift the control signal voltage level from nominal VDD to 10 V, since the LNA inputs see signals in the 10-V domain. Fig. 4 shows the proposed level shifter. Because 10 V is much higher than the transistor oxide breakdown, coupling capacitors implemented with a metal–oxide–metal (MOM) structure are used to bridge to the high voltage in the level shifter. However, the coupling capacitors may suffer from transistor leakage due to infrequent mode switches. To avoid leakage, the capacitors are periodically refreshed with the clock. It is complementarily switched for continuous operation.

C. Low-Noise Amplifier and Programmable-Gain Amplifier

The first stage amplifier determines the overall noise performance of the analog signal chain as it is the most

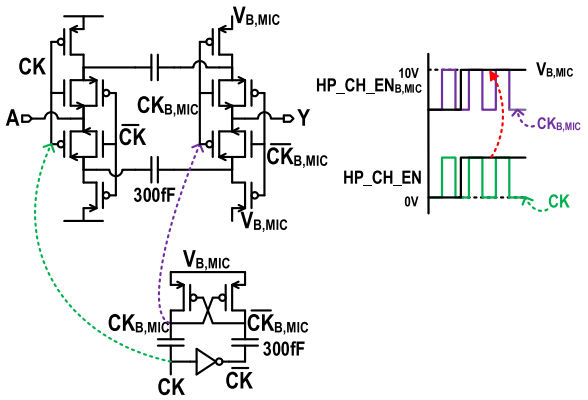


Fig. 4. 10-V level shifter shifts up nominal VDD level to 10 V with periodic refresh. Its waveforms are shown at right.

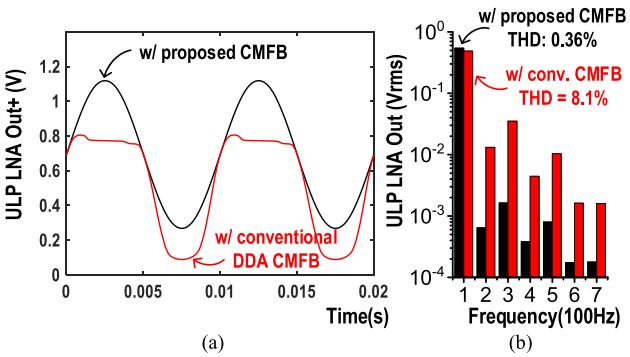


Fig. 6. (a) ULP LNA output waveform with conventional DDA CMFB (red) versus with proposed CMFB consisting of coupling capacitors and DDA (black). (b) Its spectrum (simulated).

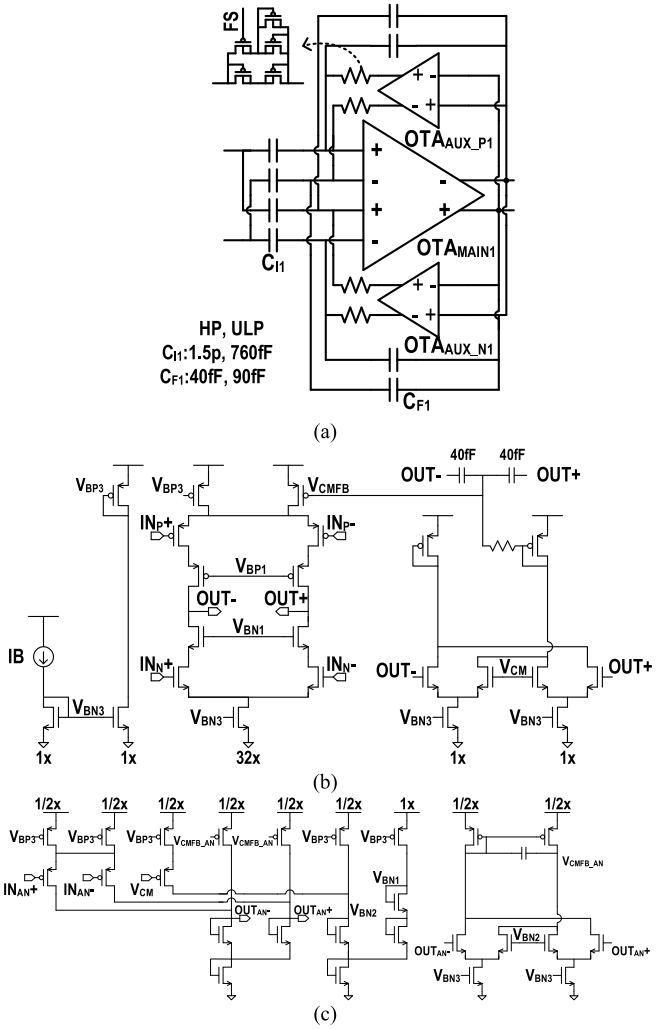


Fig. 5. (a) LNA circuit diagram. (b) $\text{OTA}_{\text{MAIN1}}$. (c) $\text{OTA}_{\text{AUX\_N1}}$ and their bias implementation. $\text{OTA}_{\text{AUX\_P1}}$ are implemented similarly with the opposite type of transistors.

noise-sensitive block in the system. Fig. 5(a) shows the block diagram of the proposed LNA. It uses capacitive feedback and pseudo-resistor dc-servo loops for low-power and small area implementation, respectively. LNA gain is set by the ratios of C_{I1} to C_{F1}. The ULP LNA gain is 18 dB, while the HP LNA gain is set to 31.3 dB to detect smaller acoustic signals. Fig. 5(b) shows the main operational transconductance amplifier (OTA) with common-mode feedback (CMFB). A conventional differential difference amplifier (DDA)-based common-mode feedback shows poor linearity when the signal is large, as shown in Fig. 6 (red line) [21], [22]. To enhance the output range and linearity, we use two different loops for the CMFB. One employs coupling capacitors for high bandwidth and good linearity across the signal amplitude. The other loop uses a DDA with a pseudo-resistor and is only responsible for setting the dc level.

The main OTA adopts an inverter-based cascode amplifier for better noise efficiency [8], [21]. PMOS and NMOS input transistor pairs are separately biased, and hence they have two pairs of C_{I1} and C_{F1} and also have two dc-servo loops. The sizes of the input transistor pairs are determined for balanced 1/f noise and thermal noise. The auxiliary amplifiers (aux-amp) in the dc-servo loops shift the output common-mode voltage of the main OTA to an optimal bias point for each PMOS/NMOS input pair to maximize the LNA output range. The implementation of the aux-amp is shown in Fig. 5(c). Very high resistance ($>$T Ω) can be readily achieved with a pseudo resistor in a small area, but its resistance varies substantially and is nonlinear when the voltage difference between the two terminals is large. In particular, mismatch among parasitic diodes and intrinsic gate diodes causes amplitude-dependent drift that may cause amplifier saturation. The aux-amps attenuate the maximum amplitude seen by the pseudo-resistors and hence, improve the operation range and linearity.

Fig. 7 shows the PGA implementation. Since PGA is less sensitive to noise than LNA, the PGAs main OTA ($\text{OTA}_{\text{MAIN2}}$) uses only a PMOS input pair for the maximum output range. The gain is adjustable between 4.6 and 31.3 dB by changing C_{I2} for both the ULP and HP chains. C_{L2} sets 500-Hz BW and 4-kHz BW for the ULP and HP chains, respectively.

Typical audio systems activated by a VAD could experience front end clipping (FEC), which may result in losing the first portion of each audio segment in passing from noise to voice activity due to the transition time between modes [23]. This effect is exacerbated especially in low power systems with

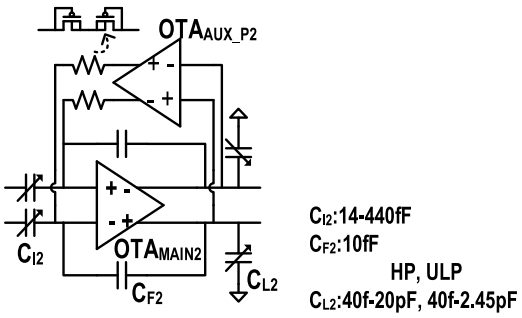


Fig. 7. PGA circuit diagram.

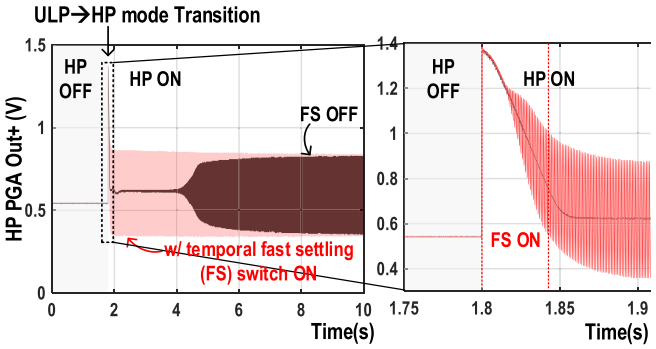


Fig. 8. Measured HP PGA output showing ULP-HP mode transition time. By turning on fast settling switches for 30 ms, the settling time reduces from 6 s (black) to 100 ms (red).

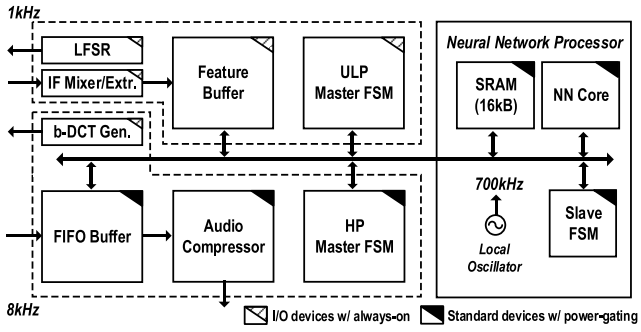


Fig. 9. Digital backend architecture.

pseudo-resistors since their extremely high resistance makes the settling time exceedingly long. In this design, we minimize the ULP-HP transition time by temporarily turning on fast settling switches [see Fig. 5(a)] during the transition. Fig. 8 shows the measured results. The common-mode voltage settling time is reduced from 6 s to 100 ms, proving the effectiveness of this method.

IV. Digital Back-End Implementation

A. Overall Architecture

Fig. 9 shows the digital back-end architecture. In the ULP mode, while the binary DCT mixer sequence generator produces a square wave for the mixer at the AFE, the IF mixer and extractor receive the ADC output to down-convert an IF signal into dc and to extract signal power as features for NN classification. Once a set of features is collected at the feature

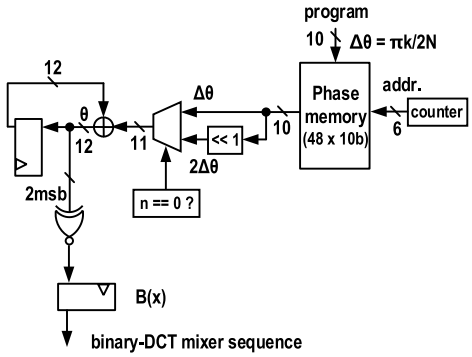


Fig. 10. Binary DCT mixer sequence generator circuits.

buffer during a frame, it is transferred to the NN processor as an input via a bus shared among digital blocks. A linear feedback shift register (LFSR) replaces the binary DCT mixer sequence generator in acoustic signature detection mode, as explained in detail in Section V. In HP mode, the first-in first-out (FIFO) buffer performs the windowing of the ADC samples for both compressions [1] and FFT. The NN processor in HP mode computes the FFT and classification. The always-on ULP modules are implemented with thick oxide I/O devices to suppress leakage, while power-gated HP modules including the NN processor are designed with standard devices. Due to the mixer-based architecture, digital processing after the ADC in ULP mode runs at 1 kHz rather than the 8-kHz Nyquist rate, yielding 41% reduction in digital feature extraction power consumption. While the binary DCT mixer sequence generator runs at 8 kHz, it only consumes 4 nW.

B. Binary DCT Mixer Sequence Generator

In ULP mode, the binary DCT mixer sequence generator shown in Fig. 10 controls the feature frequency band selection by creating a DCT basis waveform to be correlated with the incoming signal. The circuit accumulates a programmable phase, which is expressed as follows:

$$\Delta\theta = \frac{\pi}{2N}k, \quad 0 \le k \le N-1 \tag{1}$$

where N is the DCT size and k is the index of selected scanning frequency bands (i.e., F_k), either as is or doubled by shifting to generate a DCT basis function by the following equation:

$$\cos(\theta) = \cos(\Delta\theta(2n+1)), \quad n = 0, 1, \ldots, N-1 \tag{2}$$

where n is the accumulation step. By simply using 2 MSBs of the accumulated phase instead of the exact cosine calculation, the binarized DCT basis waveform can be obtained by the following equation:

$$B(f(k,n)) = B\left(\cos\left(\frac{\pi}{2N}k(2n+1)\right)\right)$$
$$B(x) = \begin{cases} 1, & x > 0 \\ -1, & x \le 0. \end{cases} \tag{3}$$

The DCT size N determines the resolution of the frequency bins and frame length, and the number of selected feature

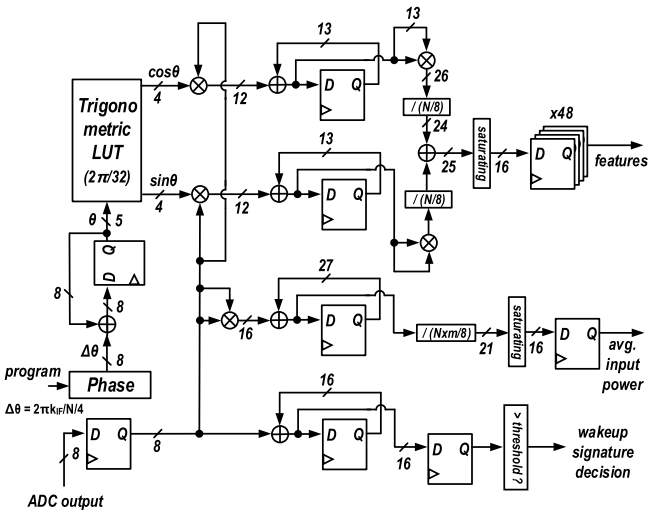


Fig. 11. IF mixer and extractor circuits.

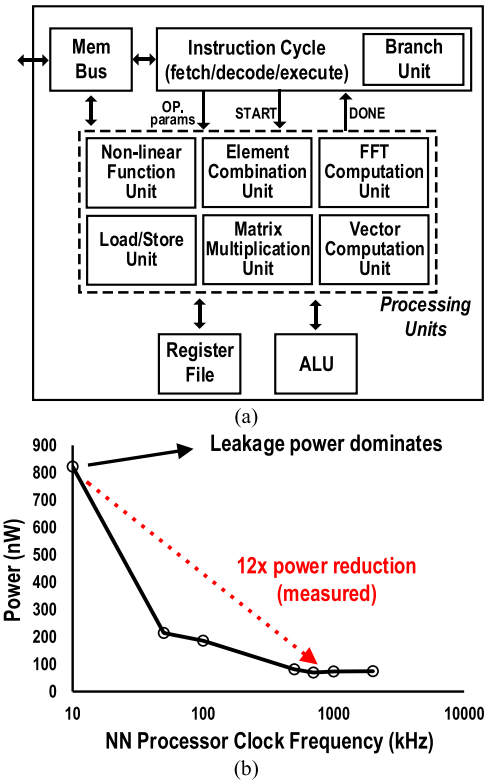


Fig. 12. (a) NN processor core architecture. (b) Measured power reduction from computational sprinting.

frequencies, m, is specified by the number of different accumulation phase values (i.e., the number of different k values). The k values are arbitrarily programmable to set particular scanning frequencies and determined during the NN training process. This design supports $N = 32, 64, \ldots, 1024$, and $m = 16, 20, 32, 48$.

C. IF Mixer and Extractor

Fig. 11 shows the IF mixer and extractor that perform quadrature mixing of the IF signal from the ADC and calculates the power as a scanned frequency feature. The extracted feature can be expressed as follows:

$$\text{feature} = \log((|X[k]|)^2),$$
$$X[k] = \frac{1}{\sqrt{N/8}} \sum_{n=0}^{N/8-1} \left(\frac{1}{8} \sum_{i=8n}^{8n+7} B(f(k,i))x[i] \right) \cdot e^{-j\frac{2\pi}{N/4}k_{\text{IF}}n} \tag{4}$$

where x is the input signal, n is the ADC sample index, and k_{IF} is the index of the IF frequency. Note that the computation inside the parentheses in (4) is done in the AFE by the mixer and the low-pass filter of the PGA. The 4-bit quantized cosine and sine functions are generated by the phase accumulator and lookup table. The phase value can be programmed by the index k_{IF} to set the proper IF frequency, avoiding interference such as 60-Hz noise or other possible ambient acoustic noise. This circuit also computes the average input power per frame to be used for automatic gain control. Last, the ADC output is accumulated for a fixed amount of time with a separate data path that is only turned on during the acoustic signature detection mode, as explained in Section V.

D. Neural Network Processor

The ULP NN processor shown in Fig. 12(a) employs a custom-built instruction set including matrix-vector multiplication, FFT, conditional branch, element-wise vector operation, non-linear activation, and min/max/averaging to support arbitrary network models and various pre/post-processing. The processor has 16 kB of on-chip SRAM storage (see Fig. 9) shared for model parameters (4 bit per weight) and instructions. By leveraging the custom-designed high-V_{th} SRAM cells, the power-gated sleep retention power of the processor is only 440 pW. However, the active state leakage power is >800 nW because the processor core and SRAM peripherals consist of standard-V_{th} devices to meet the performance requirement of the HP mode, and this active leakage power is much higher than the power consumed by ULP feature extraction. Hence, if the processor runs at a slow clock frequency of 1 kHz with the rest of the ULP digital processing modules to minimize dynamic power, then system power consumption would be dominated by NN processor leakage. To suppress this active leakage power, the concept of computational sprinting is adopted, minimizing the active time of the NN processor. Since ULP feature extraction operates sequentially, there is a long interval between classifications of a frame. The NN processor sprints at 700 kHz once the sequential feature extraction is complete and then is power-gated for the remainder of the next feature extraction. When 128-pt DCT, 32-feature, and a 32-32-16-2 NN model configuration are used, a duty cycle of 0.8% (sprint/sleep ratio) is achieved with 512 ms of frame interval, resulting in a 12× power reduction in the NN processor compared with running it at 10 kHz without sprinting, as shown in Fig. 12(b). On the other hand, in HP mode, a 128-80-20-2 NN model configuration is used. The mixer-based sequential frequency scanning feature extraction is replaced by a parallel FFT based approach that

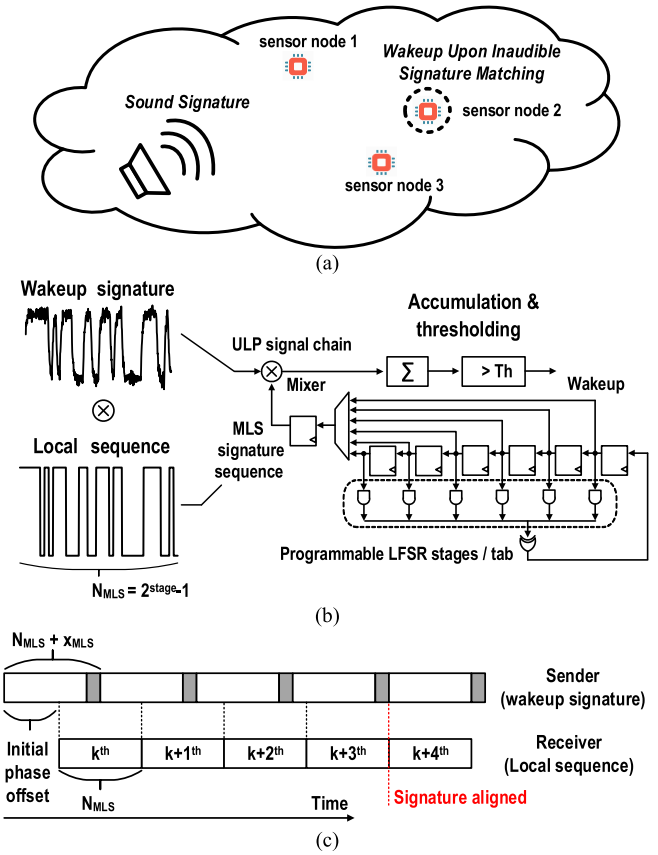


Fig. 13. (a) Inaudible acoustic signature wakeup detection. (b) Local MLS signature generator using programmable LFSR. (c) Time-drift synchronization scheme.

extracts the full 128 features by performing the 256-pt FFT on a 32 ms of the frame. The HP mode operation reduces the latency of feature extraction by a factor of 16× at the cost of 2.47-μW power consumption (measured; for AFE and digital FFT feature extraction) compared to the ULP mixer-based sequential frequency scanning approach. The NN processor stays active running at 700 kHz without duty cycling or clock gating for the HP mode to maintain the 124× increased throughput of 371 kmacs/s, compared with 3 kmacs/s in the ULP mode. Unlike the ULP mode, the active leakage does not dominate the overall power consumption in HP mode.

V. Acoustic Signature Wakeup Detection

The system also features inaudible acoustic signature detection as an alternate wakeup mechanism. This feature enables user-command silent remote wakeup of the sensor node without disturbing other sensors or people around them, as shown in Fig. 13(a). The mixer-based architecture is reused to realize the signature detection, as depicted in Fig. 13(b). An incoming signal is mixed with a local pseudorandom sequence through the mixer in the ULP AFE, and then the (digitally) accumulated value for a full sequence is compared with a threshold to determine the existence of a signature with the circuit shown in Fig. 11, as explained in Section IV-C. In this mode, a programmable LFSR running at 1 kHz replaces the binary DCT mixer sequence generator, producing a maximal length

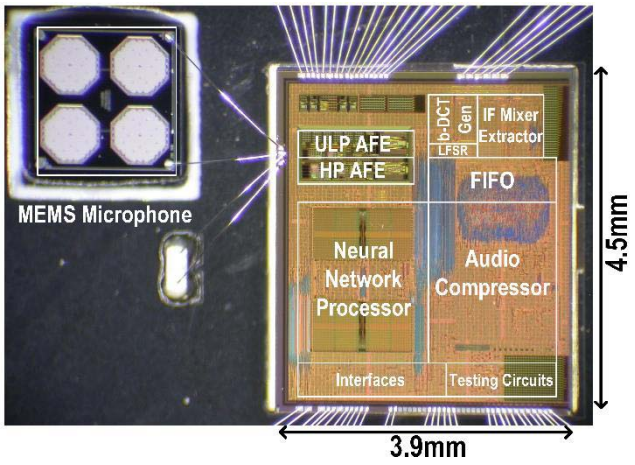


Fig. 14. Die micrograph and system integration with MEMS microphone.

sequence (MLS) to be mixed with the input signal. The length of MLS (N_{MLS}) is determined as $2^{stage} - 1$, where the *stage* is the number of LFSR stages. The LFSR tabs are arbitrarily programmed to allow a dedicated MLS for each sensor node, and to configure bit-stages of LFSR, exploiting the tradeoff between the minimum required SNR and detection latency.

The proposed sequence correlation with simple mixing requires exact phase alignment between the sequence from the transmitter and receiver. However, this phase alignment cannot be guaranteed because each sensor operates on unsynchronized independent clock sources. Running a full correlation at every sample to test all possible phases is computationally expensive. To mitigate this issue, we propose a time-drift synchronization scheme to realize correlation with simple mixing at low power. As shown in Fig. 13(c), the transmitter and receiver use intentionally mismatched sequence lengths of $N_{\mathrm{MLS}} + x_{\mathrm{MLS}}$ and N_{MLS}, respectively. Due to the length mismatch by x_{MLS}, relative phases of two sequences drift over time and periodically align with each other at the beginning of the sequence, and the accumulated mixed-signal produces periodic peaks to trigger wakeup. The period of the peaks, or the worst detection latency, is determined by $N_{\mathrm{MLS}}(N_{\mathrm{MLS}} + x_{\mathrm{MLS}})\ f_{\mathrm{LFSR}}$.

VI. Measurement Results

The chip was fabricated in 180-nm CMOS and integrated with a MEMS microphone, as shown in Fig. 14. The ULP and HP chain amplifiers consumed 31 and 370 nW, respectively. The ULP chain amplifiers have 16 and 62 μVrms measured input-referred noises with the maximum and minimum gain settings, respectively, as shown in Fig. 15(a). The maximum PGA output range that satisfies 8-bit accuracy [<0.4% total harmonic distortion (THD)] is 1.45 Vpp. The HP chain amplifiers have 8.7-μVrms input-referred noise across all PGA gain settings [see Fig. 15(b)]. Fig. 16 shows the measured mixer-based frequency scanning operation and input referred noise spectrum for the 64-pt DCT case. Two different applied tones, 1 and 2 kHz, were mixed down to 250 Hz in the IF, and power was extracted by DSP at two mixing frequencies each: 1) 0.75 and 1.25 kHz for 1-kHz input tone and 2)

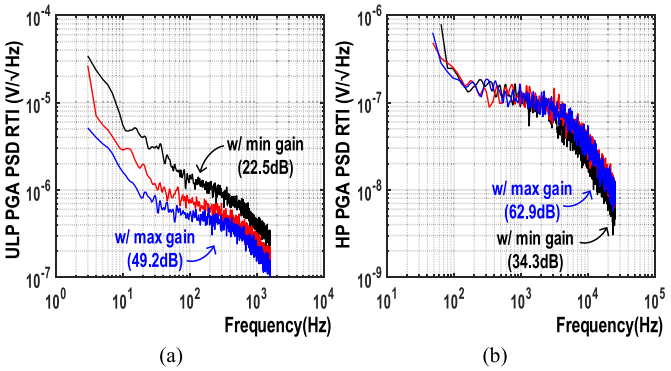


Fig. 15. (a) ULP PGA. (b) HP PGA input referred noise spectrum density with different PGA gain settings (min, mid, and max gain).

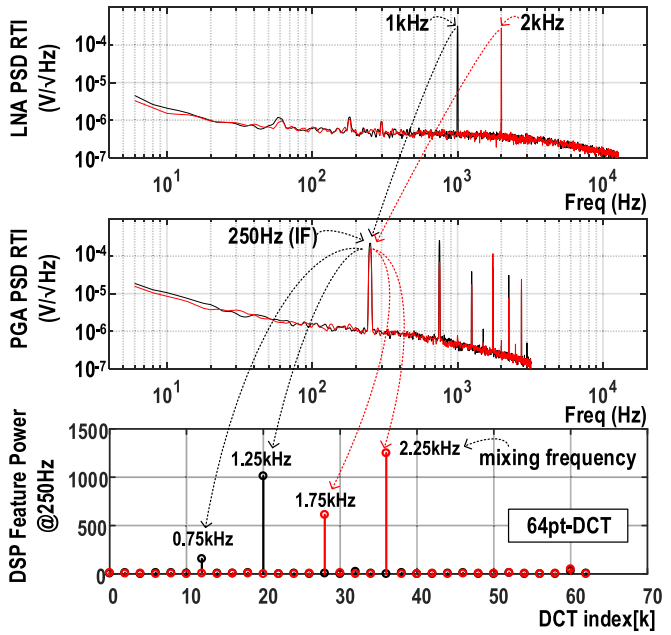


Fig. 16. Power spectral density referred to input (PSD RTI) for LNA, PGA, and DSP. Two different applied tones (1 and 2 kHz) are mixed down to 250 Hz in IF and extracted by DSP at two mixing frequencies each (0.75 and 1.25 kHz for 1 kHz and 1.75 and 2.25 kHz for 2-kHz tone).

1.75 and 2.25 kHz for 2-kHz input tone. Fig. 17 shows the measured ULP and HP mode power breakdown. The total ULP power was measured as 142 nW, and every block power was very balanced, which indicates a well-optimized design. The measured HP power was 18 μW dominated by the digital circuits.

For VAD performance evaluation, 40 min of speech segments were concatenated from the LibriSpeech data set and mixed with babble noise from the NOISEX-92 data set for training. For testing, 10 min of concatenated speech and noise segments were used. Exclusive data sets were used for NN training and evaluation to guarantee no over-fitting occurred.

We first performed electrical testing by inputting signal feeds to the LNA via an electrical connection. Fig. 18(a) shows the measured receiver operating characteristic (ROC) curve with varying SNRs in the ULP mode. The detection

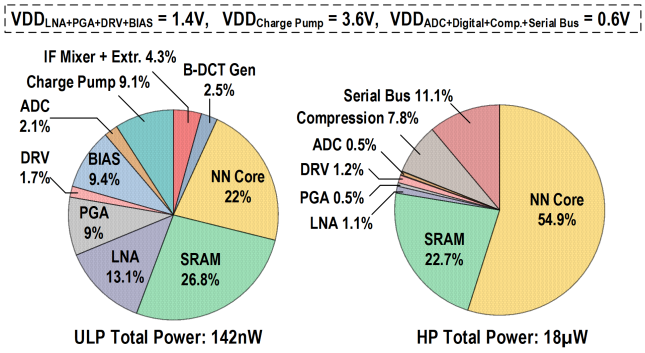


Fig. 17. Measured power distribution of ULP mode (left) and HP mode (right).

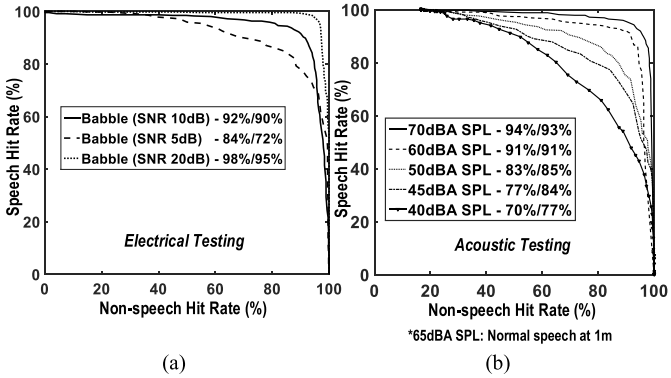


Fig. 18. ROC curves for ULP VAD mode with (a) varying SNRs in the electrical test (electrical connection to LNA) and (b) SPLs in the acoustic test (using speaker/integrated microphone in the sound chamber).

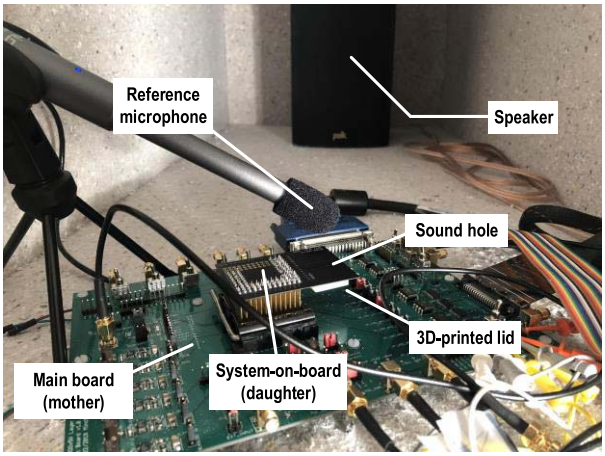


Fig. 19. Acoustic testing setup. Proposed chip was integrated into the system-on-board with a MEMS microphone and 3-D-printed lid and tested in a sound chamber.

threshold is set by the point on the ROC curve that maximizes the rectangular area formed by its coordinates. The system achieves 91.5%/90% speech/non-speech hit rates at 10-dB SNR with babble noise in the ULP mode when programmed with an NN of size 32-32-16-2 neurons with two hidden-layers, exhibiting ~7.5% better hit rate with 7× less power consumption than prior state-of-the-art works.

Unlike prior-art, we also performed an acoustic VAD test with the setup shown in Fig. 19. The proposed chip was

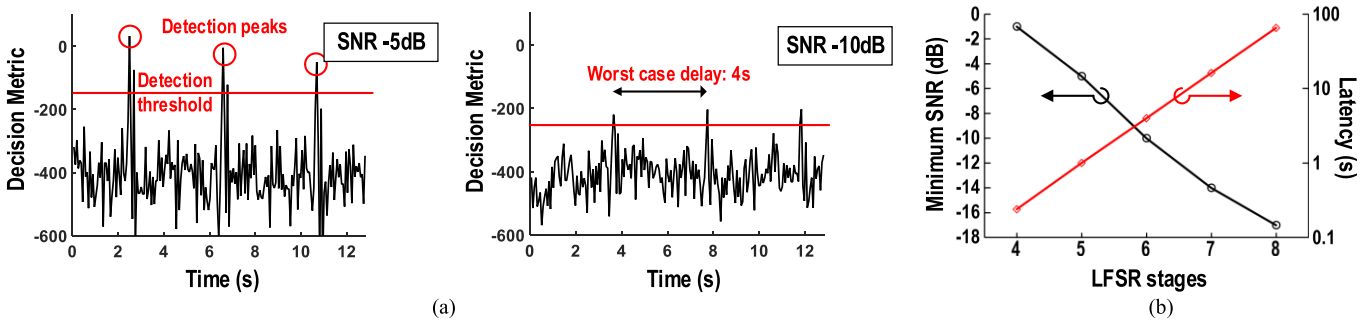


Fig. 20. Measurement results of acoustic signature wakeup detection. (a) With MLS sequence of six stages, $N_{\mathrm{MLS}} = 63$, and $x_{\mathrm{MLS}} = 1$ at various SNRs, showing detection down to −10-dB SNR. (b) With various LFSR stages, showing the tradeoff between the minimum required SNR versus worst case detection latency.

integrated with a MEMS microphone in the daughterboard, which includes a sound hole, and is then covered by a 3-D-printed custom lid to provide an acoustic cavity for the microphone and protect electronics at the same time. Then, the daughterboard was connected to the motherboard and placed within the sound chamber to achieve very low ambient noise, around 35-dBA sound pressure level (SPL). For acoustic testing, we concatenated speech segments without mixing any background contextual noise to measure the effect of circuit noise only. The measurement results show >83%/85% speech/non-speech hit rates with a signal level down to 50-dBA SPL, as shown in Fig. 18(b). The measured AFE equivalent input noise (EIN) is 45- and 44-dB SPL (no weighting) for ULP (500-Hz BW) and HP (4-kHz BW) chains, respectively.

The measurement of acoustic signature wakeup detection was also performed. As shown in Fig. 20(a), the system wakes up under exposure to as little as −10-dB SNR of white-noise-like sound when MLS signature of 6-stages, $N_{\mathrm{MLS}} = 63$, and $x_{\mathrm{MLS}} = 1$ is used, consuming 66 nW. The detection threshold of the decision metric is set to 10 dB to measure minimum SNR. These results prove that the system can be awoken by a signature buried in ambient noise that is inaudible to humans near the receiver. Moreover, Fig. 20(b) shows that the increased stages of LFSR allow more relaxed SNR requirement at the cost of increased detection latency. Note that every added stage achieves around 3 dB of SNR gain, but pays ∼4× increased latency.

Fig. 21 shows measured logic analyzer output of overall system operation. The acoustic system stayed in the ULP mode when there was no voice. The system clock ran at 1 kHz, and NN output data was observed every 512 ms. Once a voice activity was detected, the proposed acoustic chip sent an interrupt request to an external microcontroller via an inter-chip serial interface [24]. Then, the microcontroller sequentially waked up HP AFE chain and HP digital back-end via the serial interface. The 100-ms delay was given for AFE signal settlement before the digital back-end operation. The system clock was switched to 8 kHz, and frame length of HP NN was 16 ms (measured in 128-pt FFT and 64 features case). The acoustic system also compressed audio with a frame length of 24 ms in the HP mode. The HP detection threshold was set to achieve a high non-speech hit rate and accurate false alarm removal (97%/25% non-speech/speech hit rates, measured with a 128-80-20-2 NN model and 256-pt FFT). When there was no voice for long enough time, the acoustic system returned to the ULP mode.

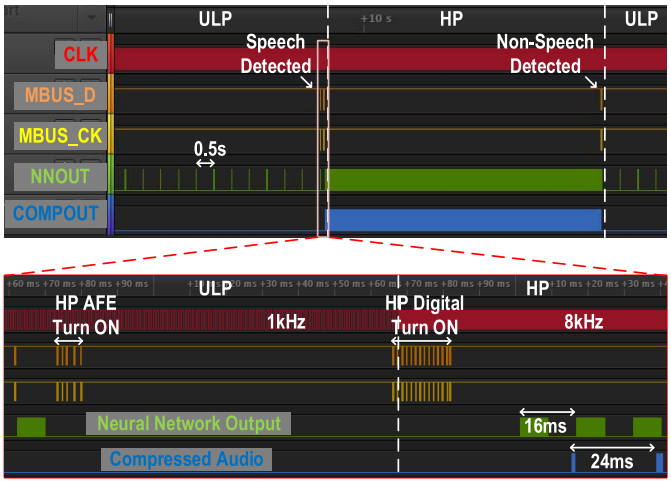


Fig. 21. Measured waveform of the acoustic system that switches between ULP and HP modes.

TABLE I
COMPARISON OF FEATURE EXTRACTOR

Feature Extractor	This Work	[5]	[7]	[4]
Technology (nm)	180	180	180	90
Feature Extraction Type	Mixed-signal	Analog to events	Digital	Analog
Channel Number	16-48	16	4-16	16
Frequency Range (Hz)	75-4k	100-5k	0.2-470	75-5k
Power (nW)	60	380	10	6000
Normalized Power (nW) [a]	5	71	34	1186
Dynamic Range (dB)	47	40	N/A	40
Building Blocks	LNA, Mixer, LPF, DSP	LNA, BPF, FWR, IAF	DSP	LNA, BPF, FWR, LPF

[a] Normalized power is calculated according to the equation in [5], normalized to 4kHz.

TABLE II
COMPARISON OF VOICE ACTIVITY DETECTOR (VAD)

Voice Activity Detector	This Work	[5]	[3]	[4]	[2]
Technology (nm)	180	180	65	90	32
Acoustic Input	Analog/passive mic. w/ gain stage	Analog mic. w/ gain stage	Assumed digitized	Analog mic. w/ gain stage	Assumed digitized
Classifier	Neural network	Neural network	Neural network	Decision tree	Energy-based
Classifier Topology Programmability	Yes	No	Yes	No	No
Dataset [a]	LibriSpeech + NOISEX-92	AURORA4 + DEMAND	AURORA2	NOISEUS	N/A
Latency (ms)	512	10	10	<100	10
Power (μW)	0.142	1	22.3	6	300
Accuracy SP/Non-SP hit rate (electrical test)	91.5%/90% [b] @ babble 10dB SNR	84%/85% @ restaurant 10dB SNR	90%/90% [c] @ unspecified context 7dB SNR	89%/85% @ babble 10dB SNR	97% accuracy @ unspecified context unspecified SNR
Acoustic Testing Performed	Yes	No	No	No	No

[a] All datasets are similar in speech quality.
[b] Measured at ULP mode with 128pt-DCT, 32 feature channels, and 250Hz IF.
[c] Converted from EER in [3]

VII. CONCLUSION

This article demonstrated the design of a sub-μW voice and non-voice acoustic activity detection chip. By using mixer-based sequential frequency scanning operation, the feature extraction power is reduced by 4×. Table I compares the proposed feature extractor with prior works. Although [8] shows the lowest power consumption, the signal bandwidth is limited to under 500 Hz. This article achieves the lowest normalized power consumption, calculated in the same manner as in [5], which reflects the power normalized to the number of channels and signal bandwidth. Moreover, this article achieves the best front end dynamic range, thanks to the proposed amplifier design.

Table II shows the comparison of this article with prior state-of-the-art VAD systems. While this design consumes the lowest power, it is worthwhile to also consider the latency or throughput. For example, [5] has better energy efficiency in terms of classification/W/s than this article. However, it is not always possible to scale the power consumption of feature extraction to a lower power level with relaxed latency since it is an always-on block. In addition, there are still useful applications (e.g., compressed speech recording after VAD) that can tolerate this latency, given the normal speech rate are 120–160 words per minute. Moreover, the digital backend design of this article offers greater flexibility to use various model topologies compared to [5], making this design a better approach for applications that are extremely power-constrained yet require mapping to various target events, such as miniaturized battery-operated IoT sensor nodes.

REFERENCES

[1] M. Cho *et al.*, " A 6×5×4 mm^{3}general purpose audio sensor node with a 4.7 μW audio processing IC," in *Proc. Symp. VLSI Circuits*, 2017, pp. C312–C313.

[2] 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.

[3] 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.

[4] K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "Context-aware hierarchical information-sensing in a 6 μW 90 nm CMOS voice activity detector," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.

[5] M. Yang, C. H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1 μW voice activity detector using analog feature extraction and digital deep neural network," in *IEEE Int. Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 346–348.

[6] M. Cho *et al.*, "A 142 nW voice and acoustic activity detection chip for mm-scale sensor nodes using time-interleaved mixer-based frequency scanning," in *IEEE Int. Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 2019, pp. 278–280.

[7] Y. Chen, M. Cho, S. Jeong, D. Blaauw, D. Sylvester, and H. Kim, "A dual-stage, ultra-low-power acoustic event detection system," in *Proc. IEEE Int. Workshop Signal Process. Syst. (SiPS)*, Oct. 2016, pp. 213–218.

[8] S. Jeong *et al.*, "Always-on 12-nW acoustic sensing and object recognition microsystem for unattended ground sensor nodes," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 261–274, Jan. 2018.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[10] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *Proc. Interspeech*, 2013, pp. 728–731.

[11] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2624–2633, Nov. 2011.

[12] I. Tashev and S. Mirsamadi, "DNN-based causal voice activity detector," in *Proc. Inf. Theory Appl. Workshop*, 2016.

[13] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza, "A deep neural network approach for voice activity detection in multi-room domestic scenarios," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2015, pp. 1–8.

[14] S. M. R. Nahar and A. Kai, "Robust voice activity detector by combining sequentially trained deep neural networks," in *Proc. Int. Conf. Adv. Inform. Concepts, Theory Appl. (ICAICTA)*, 2016, pp. 1–5.

[15] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

[16] Z. Fan, Z. Bai, X. Zhang, S. Rahardja, and J. Chen, "AUC optimization for deep learning based voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6760–6764.

[17] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech Language Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[18] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, 2014, pp. 1534–1538.
[19] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 796–800.
[20] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4500–4504.
[21] S. Oh, T. Jang, K. D. Choo, D. Blaauw, and D. Sylvester, "A 4.7 μW switched-bias MEMS microphone preamplifier for ultra-low-power voice interfaces," in *Proc. Symp. VLSI Circuits*, 2017, pp. C314–C315.
[22] P. Harpe, H. Gao, R. van Dommele, E. Cantatore, and A. H. M. van Roermund, "A 0.20 mm^{2}3 nW signal acquisition IC for miniature sensor nodes in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 240–248, Jan. 2016.
[23] F. Beritelli, S. Casale, and A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 9, pp. 1818–1829, Dec. 1998.
[24] P. Pannuto *et al.*, "MBus: A system integration bus for the modular microscale computing class," *IEEE Micro*, vol. 36, no. 3, pp. 60–70, May 2016.

Sechang Oh (S'12−M'17) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2014 and 2017, respectively.

He is currently a Research Fellow with the University of Michigan. His current research interests include low-power sensor nodes and IoT sensor systems.

Dr. Oh was a recipient of the Jeongsong Fellowship.

Minchang Cho (S'09) received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree with the University of Michigan, Ann Arbor, MI, USA.

From 2010 to 2015, he was with Samsung Electronics Company Ltd., Hwasung, South Korea, where he was involved in the low-power high-speed circuit design for DRAM products fabricated in 17- to 25-nm DRAM processes. His current research interests include low-power digital circuits design, efficient architectures for machine learning and signal processing, and autonomous Internet of Things (IoT) sensor systems.

Zhan Shi received the B.Eng. degree in microelectronics from Sun Yat-sen University, Guangzhou, China, in 2016, and the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2018.

He is currently with Apple Inc., Cupertino, CA, USA, where he is a Pixel IP Design Engineer with the Silicon Engineering Group. His research interests include architecture development and SoC implementation for machine learning and multimedia processing.

Jongyup Lim (S'17) received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2016, and the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2018, where he is currently pursuing the Ph.D. degree.

His current research interests include neural recording systems, energy-efficient deep learning hardware, clock generation, and ultra-low-power sensor node design.

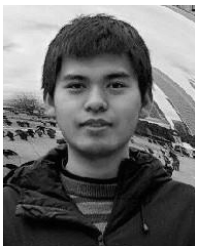
Dr. Lim was a recipient of the Doctoral Fellowship from Kwanjeong Educational Foundation, South Korea.

Yejoong Kim received the bachelor's degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2008, and the master's and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2012 and 2015, respectively.

He is currently a Research Fellow with the University of Michigan and the Vice President of research and development with CubeWorks Inc., Ann Arbor. His research interests include subthreshold circuit designs, ultra-low-power SRAM, and the design of millimeter-scale computing systems and sensor platforms.

Seokhyeon Jeong (S'12) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2011, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2017.

He is currently with the University of Michigan and CubeWorks Inc., Ann Arbor, where he is involved in researching and developing interface circuits for ultra-low-power sensor nodes. His research interests include ultra-low-power temperature sensors, A/D converters, and the design of millimeter-scale computing systems.



Yu Chen received the B.S. degree in electrical engineering from Xian Jiaotong University, Xian, China, in 2014, and the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2016, where he is currently pursuing the Ph.D. degree.

His research interests include deep learning, reinforcement learning, computer vision, and their applications in ultra-low-power systems.

Rohit Rothe (S'19) received the B.Tech. and M.Tech. degrees in electrical engineering from IIT Bombay, Mumbai, India, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Michigan, Ann Arbor, MI, USA.

His current research interest includes ultra-low-power analog VLSI design.

David Blaauw (M'94−SM'07−F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, USA, in 1986, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, Urbana–Champaign, IL, USA, in 1991.

Until August 2001, he was with Motorola Inc., Austin, TX, USA. Since August 2001, he has been on the faculty of the University of Michigan, Ann Arbor, MI, USA, where he is currently a Kensall D. Wise Collegiate Professor of electrical engineering and computer science (EECS). He is the Director of the Michigan Integrated Circuits Laboratory. He has published over 600 articles, and he holds 65 patents. He has extensive research in ultra-low-power computing.

Dr. Blaauw serves on the IEEE International Solid-State Circuits Conference's Technical Program Committee.

Hun-Seok Kim (S'10–M'11) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2010.

From 2010 to 2014, he was a Technical Staff Member with Texas Instruments Inc., Dallas, TX, USA. He is currently an Assistant Professor with the University of Michigan, Ann Arbor, MI, USA. His research focuses on system analysis, novel algorithms, and VLSI architectures for low-power/high-performance wireless communication, signal processing, computer vision, and machine learning systems.

Dr. Kim was a recipient of the 2018 Defense Advanced Research Projects Agency (DARPA) Young Faculty Award. He serves as an Associate Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING and the IEEE SOLID-STATE CIRCUITS LETTERS.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 1999.

He was the Founding Director of the Michigan Integrated Circuits Laboratory (MICL), a group of ten faculty and 70 over graduate students. He was a Research Staff with the Advanced Technology Group, Synopsys, Mountain View, CA, USA, Hewlett-Packard Laboratories, Palo Alto, CA, USA, and a Visiting Professor with the National University of Singapore, Singapore, and Nanyang Technological University, Singapore. He is currently a Professor of electrical engineering and computer science with the University of Michigan, Ann Arbor, MI, USA. He has published over 500 articles along with one book and several book chapters. He holds 48 U.S. patents. He co-founded Ambiq Micro, a fabless semiconductor company developing ultra-low-power mixed-signal solutions for compact wireless devices. His research interests include the design of millimeter-scale computing systems and energy efficient near-threshold computing.

Dr. Sylvester was a recipient of the NSF CAREER Award, the Beatrice Winner Award at the IEEE International Solid-State Circuits Conference (ISSCC), an IBM Faculty Award, the Semiconductor Research Corporation (SRC) Inventor Recognition Award, the University of Michigan Henry Russel Award for distinguished scholarship, and ten best paper awards and nominations. His dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley Electrical Engineering and Computer Science Department (EECS). He was named one of the top contributing authors at ISSCC, most prolific author at the IEEE Symposium on VLSI Circuits. He was the IEEE Solid-State Circuits Society Distinguished Lecturer. He serves on the Technical Program Committee of the IEEE International Solid-State Circuits Conference and the Administrative Committee of the IEEE Solid-State Circuits Society. He serves/has served as an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS, the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN, and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS.