17.2 A 142nW Voice and Acoustic Activity Detection Chip for mm-Scale Sensor Nodes Using Time-Interleaved Mixer-Based Frequency Scanning

Minchang Cho*, Sechang Oh*, Zhan Shi, Jongyup Lim, Yejoong Kim, Seokhyeon Jeong, Yu Chen, David Blaauw, Hun-Seok Kim, Dennis Sylvester

University of Michigan, Ann Arbor, MI
*Equally-Credited Authors (ECAs)

Acoustic sensing is one of the most widely used sensing modalities to intelligently assess the environment. In particular, ultra-low power (ULP) always-on voice activity detection (VAD) is gaining attention as an enabling technology for IoT platforms. In many practical applications, acoustic events-of-interest occur infrequently. Therefore, the system power consumption is typically dominated by the always-on acoustic wakeup detector, while the remainder of the system is power-gated the vast majority of the time. A previous acoustic wakeup detector [1] consumed just 12nW but could not process voice signals (up to 4kHz bandwidth) or handle non-stationary events, which are essential qualities for a VAD. Prior VAD ICs [2,3] demonstrated reliable performance but consumed significant power (>20μW) and lacked an analog frontend (AFE), which further increases power. Recent analog-domain feature extraction-based VADs [4,5] also reported μW-level power consumption, and their simple decision tree [4] or fixed neural network-based approach [5] limited broader use for various acoustic event targets. In summary, no sub-μW VAD has been reported to date, preventing the use of VADs in unobtrusive mm-scale sensor nodes.

This work presents a 142nW programmable, neural network-based acoustic sensing system for both VAD and non-voice event detection. We use a *time-interleaved mixer-based architecture* that sequentially scans and down-converts the 4kHz bandwidth signal to a ≤500 Hz passband, reducing amplifier, ADC, and DSP power by 4×. The neural network (NN) processor employs computational sprinting, which minimizes static energy dominance in low frequency/voltage regime, providing 12× power reduction in the digital domain. The architecture (Fig. 17.2.1, top) has two signal chains: an ULP chain with 142nW consumption that is always on and a 18μW high performance (HP) chain that wakes upon event detection by the ULP chain. Unlike the ULP chain, the HP chain has a full 4kHz bandwidth AFE while sharing the same digital backend with the ULP chain. In addition to VAD, the system features an inaudible acoustic signature detection mode to enable remote silent system wakeup. With always-on VAD, the system has a 4.5-year lifetime with a 5mm mini coin-cell battery (2mAh) and achieves 91.5% voice detection accuracy.

Figure 17.2.1 shows the *time-interleaved mixer-based architecture* that reduces power consumption of AFE and DSP by lowering their bandwidth and sampling rate to 500Hz and 1kHz, respectively. The incoming signal from the microphone is amplified by an LNA with the full 4kHz bandwidth. At this point the mixer, switched by a binary discrete cosine transform (DCT) sequence, immediately down-converts the frequency of a desired feature to a programmable intermediate frequency (IF) of <500Hz. The digital binary sequence generator supports an arbitrary DCT frequency for the mixer switch control; for example, the 4kHz band can be divided into 31.25Hz frequency bins using a 128-pt DCT, and the energy content of 32 bands is sequentially extracted by sweeping DCT frequencies (F_1, … 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 500Hz bandwidth (via a PGA) and digitized at 1kS/s. Finally, the digital IF mixer down-converts the signal to DC, and the feature power is measured. With a DCT length of 16ms per feature (128-pt DCT with 8kHz binary mixing), 32-feature extraction requires a 512ms frame. The mixer-based structure reduces bandwidth, sampling rate, and clock frequency of AFE and DSP after the mixer; thus, the feature extraction power consumption is decreased from 225nW (simulation) to 60nW. IF is set to ~250Hz to avoid PGA 1/f noise, while the image aliasing issue of non-quadrature mixing is mitigated by a NN trained with image-aliased signals.

Figure 17.2.2 shows the circuit diagram of the AFE with ULP and HP chains. Each chain consists of an LNA, PGA, ADC driver, and ADC. Both chains share a single MEMS microphone and charge pump. For high sensitivity, a three-stage Dickson pump is used to bias the microphone at 10V. The microphone switches between the chains by controlling *ULP_CH_EN* and *HP_CH_EN*, which are level-shifted to 10V. As coupling capacitors for the level-shifting may suffer from leakage between infrequent mode switches, we refresh their charge periodically. Capacitive feedback and pseudo-resistor dc-servo loops are used for the ULP LNA (18dB gain). The LNA OTA adopts an inverter-based cascode amplifier for noise efficiency. Its common-mode feedback is composed of two loops. Coupling capacitors provides fast loop response, and the DDA output sets the DC voltage. The auxiliary amplifiers (Aux-amp) in the dc-servo loop shift common-mode voltages for high dynamic range. In addition, Aux-amp attenuates large LNA output due to diode-connected nature, which reduces the maximum amplitude seen by the pseudo-resistors, reducing their amplitude-dependent drift. The mixer is composed of transmission gates switched by the DCT sequence generator. Unlike the LNA OTA, PGA OTA uses only a PMOS input pair for the maximum output range. By tuning cap C_{i2}, the gain is adjustable between 4.5 and 31.2dB, and C_L sets 500Hz BW for ULP mode. The ADC driver is followed by an 8b SAR ADC. The HP blocks are similar to the ULP counterparts except that they are scaled for low noise and full 4kHz bandwidth. We minimize the ULP-HP transition time by temporarily turning on the fast settling switches during the transition. This helps to set the common-mode voltage very quickly (100ms vs 6s, measured).

Figure 17.2.3 (top) depicts the digital backend architecture. Always-on modules are implemented with thick oxide I/O devices to suppress leakage, while power-gated modules (NN processor, FIFO, and audio compressor [6]) are designed with standard devices. The NN processor uses a 16kB custom ultra-low retention leakage SRAM for the 4-bit weights storage, and its ISA includes matrix-vector multiplication, non-linear activation, FFT, conditional branch, element-wise vector operation, and min/max/averaging. In the ULP mode, the processor sprints at a relatively high frequency clock (700kHz, Fig. 17.2.3, bot right) when the sequential feature extraction is complete (every 512ms) and then power-gates to minimize the leakage power, resulting in 12× power reduction (sprint/sleep ratio of 0.008). In HP mode, the processor computes a full (non-sequential) FFT and a larger NN without duty cycling for the improved latency (32ms) or hit rate at the cost of higher power consumption (14μW). The binary mixer sequence generator (Fig. 17.2.3, bot left) is programmable for different DCT sizes, feature frequency resolutions, and number of features. Due to the mixer-based architecture, digital processing runs at 1kHz (vs. 8kHz Nyquist rate), yielding 41% reduction of feature extraction power.

The system also features inaudible acoustic signature detection to enable silent remote system wakeup (Fig. 17.2.4). The binary mixing sequence is replaced with a maximal length sequence (MLS) signature generated by a 1kHz programmable LFSR. Correlation between the incoming wakeup signal and the local sequence is performed through the ULP mixer and PGA. To synchronize the wakeup and local sequence, we employ a time-drift synchronization scheme that uses intentional frequency mismatch between the two so that they naturally time-synchronize periodically. This inaudible (–10dB SNR) signature detection consumes only 66nW with 4s worst case latency.

The chip is fabricated in 180nm CMOS and integrated with a MEMS microphone (Fig. 17.2.7). The ULP and HP chain amplifiers consume 31nW and 370nW with 16μVrms and 9.1μVrms input-referred noise, respectively. Figure 17.2.5 (top left) shows the measured mixer-based frequency scanning operation and input referred noise spectrum. Figure 17.2.5 (bot left) shows the measured ULP chain power breakdown. For VAD evaluation, speech from the LibriSpeech dataset is mixed with babble noise from the NOISEX-92 dataset. NN training and evaluation use exclusive datasets. Figure 17.2.6 compares the system with prior work. The system achieves 91.5%/90% speech/non-speech hit rates at 10dB SNR with babble noise (electrical test, Fig. 17.2.5 top right) in ULP mode when programmed with a NN of size 32-32-16-2 neurons, exhibiting ~7.5% better hit rate at 7× less power consumption than prior state-of-the-art. Unlike prior-art, we also report acoustic VAD test results measured in a sound chamber, showing >83%/85% speech/non-speech hit rates with a signal level down to 50dBA SPL (Fig. 17.2.5, bot right).

Acknowledgements:
This work was supported by Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References:
[1] S. Jeong, et al., "A 12nW Always-on Acoustic Sensing and Object Recognition Microsystem Using Frequency-Domain Feature Extraction and SVM Classification," *ISSCC Dig. Tech. Papers*, pp. 362-363, Feb. 2017.
[2] A. Raychowdhury, et al., "A 2.3 nJ/Frame Voice Activity Detector-Based Audio Front-End for Context-Aware System-On-Chip Applications in 32nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963-1969, Aug. 2013.
[3] M. Price, et al., "A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating," *ISSCC Dig. Tech. Papers*, pp. 244-245, Feb. 2017.
[4] K. Badami, et al., "Context-Aware Hierarchical Information-Sensing in a 6μW 90nm CMOS Voice Activity Detector," *ISSCC Dig. Tech. Papers*, pp. 430-431, Feb. 2015.
[5] M. Yang, et al., "A 1μW Voice Activity Detector Using Analog Feature Extraction and Digital Deep Neural Network," *ISSCC Dig. Tech. Papers*, pp. 346-347, Feb. 2018.
[6] M. Cho, et al., "A 6×5×4mm³ General Purpose Audio Sensor Node with a 4.7μW Audio Processing IC," *IEEE Symp. VLSI Circuits*, pp. 312-313, June 2017.

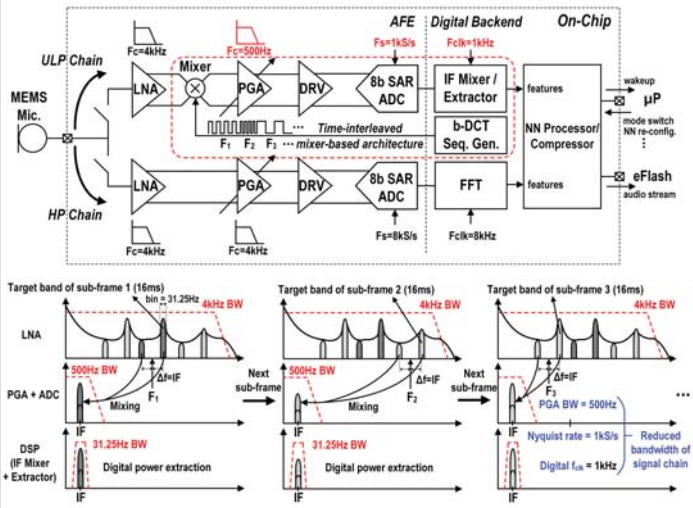

Figure 17.2.1: Acoustic sensing system architecture (top), and operation principle of time-interleaved mixer-based frequency scanning (bottom).

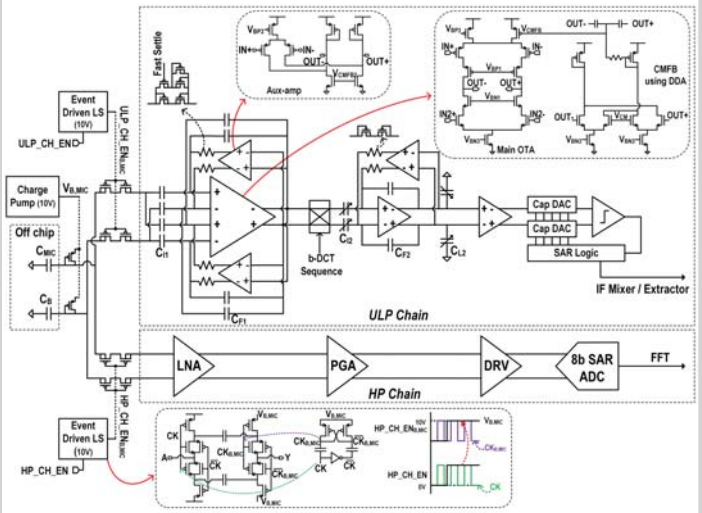

Figure 17.2.2: Circuit diagram of the analog front-end with ULP and HP chains.

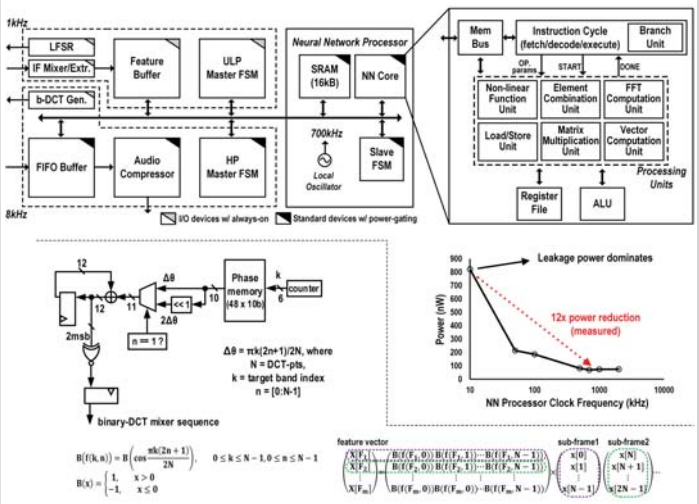

Figure 17.2.3: Digital backend architecture including neural network processor (top), measured power reduction from computational sprinting (bottorm right), and binary DCT mixer sequence generator (bottom left).

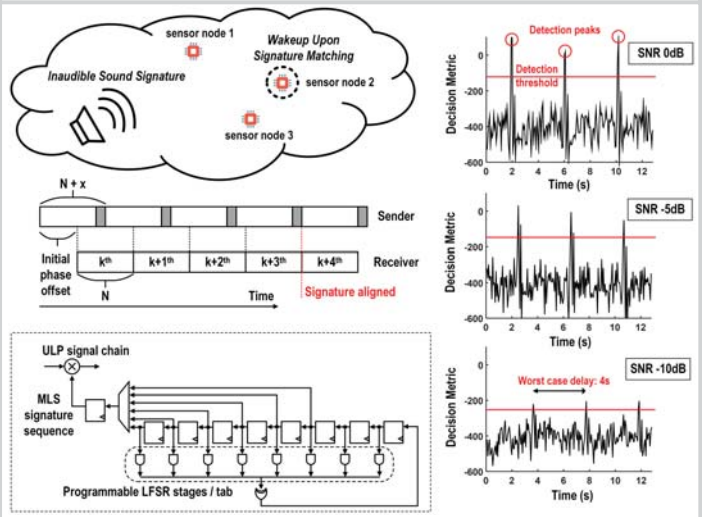

Figure 17.2.4: Acoustic signature wakeup detection (left), and measurement results with 6 stages, 63-length sequence at various SNRs (right), showing detection down to -10dB SNR.

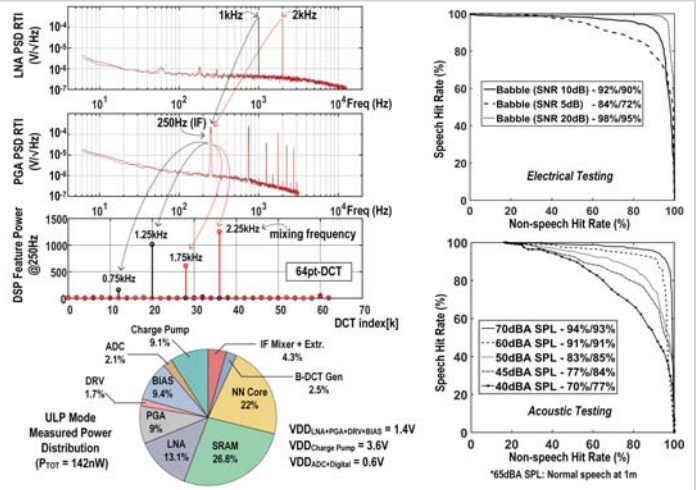

Figure 17.2.5: Chip measurement results. Power spectral density for LNA, PGA, and DSP (top left). Two different applied tones are mixed down to 250Hz in IF and extracted by DSP at two mixing frequencies each. ULP mode power distribution (bot left), and ROC curves for VAD (right).

Feature Extractor	This Work	ISSCC'18 [5]	ISSCC'16 [1]	ISSCC'15 [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)	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

ULP AFE(LNA+PGA) Summary	
Gain(dB)	22.5-49.2
Out Noise (mVrms)	0.83@min gain 4.6@max gain
Max Output (Vpp)	1.45(<0.4%THD)
Dynamic Range(dB)	55.8@min gain 40.9@max gain
Power (nW)	31
Bandwidth (Hz)	500
VDD(V)	1.4

Voice Activity Detector	This Work	ISSCC'18 [5]	ISSCC'17 [3]	ISSCC'15 [4]	JSSC'13 [2]
Technology (nm)	180	180	65	90	32
Acoustic Input	Analog/passive mic. w/ gain stage	Analog mic. w/ gain stage	Assume digitized	Analog mic. w/ gain stage	Assume digitized
Classifier	Neural network	Neural network	Neural network	Decision tree	Energy-based
Classifier Topology Programmability	Yes	No	Yes	No	No
Dataset**	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%*** @ babble 10dB SNR	84%/85% @ restaurant 10dB SNR	90%/90%**** @ unspecified context 7dB SNR	89%/85% @ babble 10dB SNR	97% accuracy @ unspecified context unspecified SNR
Acoustic Testing Performed	Yes	No	No	No	No

*calculated according to the equation in [5], normalized to 4kHz; **all datasets are similar in speech quality;
***measured at ULP mode with 128pt-DCT, 32 feature channels, and 250Hz IF. ****converted from EER in [3].

Figure 17.2.6: Comparison table for feature extractor (top left), VAD (bottom), and performance summary of ULP AFE (top right).

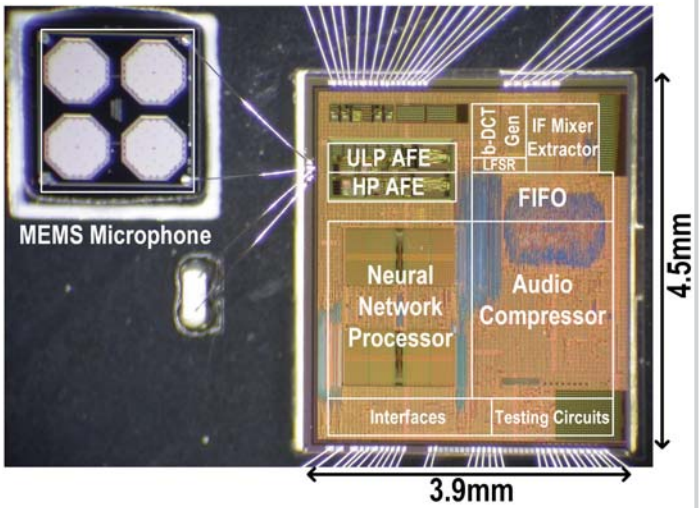

Figure 17.2.7: Die micrograph and system intgration with MEMS microphone.

978-1-5386-8531-0/19/$31.00