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

Minchong Cho*, Sechang Oh*, Zhan Shi, Jongyup Lim, Yejoong Kim, Seokyoung 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, 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-event detection. We use a time-interleaved mixer-based architecture that sequentially scans and down-converts the 4kHz bandwidth signal to a ≤500Hz passband, reducing amplifier, ADC, and DSP power by 4×. The neural network (NN) processor employs computational pruning, 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 channel 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 audible acoustic signature detection mode to enable remote silent system wake-up. 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 the point of interest, the mixer is 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 segments using a 128-point DCT. The incoming signal from the microphone is sequentially extracted by sweeping DCT frequencies (F<sub>1</sub>, … F<sub>32</sub>). The 32 bands are separately band-pass filtered with 225nW (simulation) to 60nW. Both chains share the same 180nm ULP LNA (180 mAh), which has a single-pole/zero gain adjustability between 4.5 and 31.2dB, and C<sub>0</sub> 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 interface power consumption 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).

The chip is fabricated in 180nm CMOS and integrated with a MEMS microphone (Fig. 17.2.2, top) that reduces the maximum amplitude seen by the pseudo-resistors, reducing their amplitude-dependent drift. The mixer is composed of transmission gates switched by the DC sequence generator. Unlike the LNA OTA, PGA OTA uses only a PMOS input pair for the maximum output range. By tuning cap C<sub>0</sub>, the gain is adjustable between 4.5 and 31.2dB, and C<sub>0</sub> 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 interface power consumption 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).

Data correlation between the incoming signal and the local sequence is performed using a maximal length sequence (MLS) signature generated by a 1kHz programmable LFSR. The binary mixing sequence is replaced with 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 system also features an audible acoustic signature detection to enable silent remote system wake-up (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 signal and the local sequence is performed using a matched filter and ADC. To synchronize the wake-up and local sequence, we employ a time-drift synchronization scheme that uses intentional frequency mismatch 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.

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. 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 shows 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 50dB 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:
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 (bottom 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).

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

---

**Feature Extractor**

- ISCC '19 (B)
- ISCC '18 (B)
- ISCC '19 (B)

**VAD (B)**

- Out of sample
- FF: 2.9MHz
- SNR: 2.4dB

**Performance Summary**

- ULP AFE
- ISCC '19 (B)
- ISCC '18 (B)
- ISCC '19 (B)
- ISCC '19 (B)

**Technology (top)**

- ISCC '18 (B)
- ISCC '19 (B)
- ISCC '19 (B)
- ISCC '19 (B)

**Acoustic Input**

- Analog/Digital
- Analog/Digital
- Analog/Digital
- Analog/Digital

**Classifier Topology**

- SVM
- SVM
- SVM
- SVM

**ROC Curves**

- VAD: 95%
- VAD: 95%
- VAD: 95%
- VAD: 95%

---

*Note: All values are measured under standard conditions.*
Figure 17.2.7: Die micrograph and system integration with MEMS microphone.