14.4 A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating

Michael Price[1,2], James Glass[1], Anantha P. Chandrakasan[1]

[1]Massachusetts Institute of Technology, Cambridge, MA
[2]Analog Devices, Cambridge, MA

The applications of speech interfaces, commonly used for search and personal assistants, are diversifying to include wearables, appliances, and robots. Hardware-accelerated automatic speech recognition (ASR) is needed for scenarios that are constrained by power, system complexity, or latency. Furthermore, a wakeup mechanism, such as voice activity detection (VAD), is needed to power gate the ASR and downstream system. This paper describes IC designs for ASR and VAD that improve on the accuracy, programmability, and scalability of previous work.

Figure 14.4.1 shows the components of our embedded speech recognizer. To support low-duty-cycle operation, the external memory should be non-volatile; today's consumer devices use MLC flash, with higher energy-per-bit than DRAM (typically 100pJ/b). This memory can easily consume more power than the ASR core itself. As shown by [1], modern ASR modeling and search techniques can be modified to reduce memory bandwidth without compromising accuracy. These techniques include deep neural networks (DNNs), weighted finite-state transducers (WFSTs), and Viterbi search over hidden Markov models (HMMs).

Most recent work on neural network hardware (e.g. [2]) targets convolutional networks (CNNs) for computer vision applications. IC implementations of keyword detection with DNNs have achieved power consumption as low as 3.3 mW [3], but DNNs have not yet been incorporated into a standalone hardware speech recognizer. We tailor DNNs for low-power ASR by using limited network widths and quantized, sparse weight matrices.

Our feed-forward DNN accelerator uses a SIMD architecture shown in Fig. 14.4.2. A compressed model is streamed in, and the decoded parameters are broadcast to 32 parallel execution units (EUs) that each process one feature vector. This reduces memory bandwidth by up to 32× at some expense in latency. Each EU has enough local memory to handle NN layers with up to 1k hidden nodes. As shown in Fig. 14.4.3, the EUs are organized in eight "chunks", which can be reconfigured to handle networks with up to 4k hidden nodes while disabling some of the EUs. Sparse weight matrices are supported by storing a run-length encoding of the nonzero coefficient locations at the beginning of each row; for an acoustic model with 31% nonzero weights, this allows a 54% reduction in memory bandwidth. Quantization tables stored in SRAM allow the weights to be stored with 4-12 bits each. The EU supports sigmoid or rectified linear nonlinearities in each layer. The sigmoid function is approximated by a 5th order Chebyshev polynomial, evaluated using Horner's method. By design, the DNN accelerator's throughput is limited by the 8b memory interface, with a typical cycle overhead of 2% for dense weight matrices and 14% for sparse weight matrices.

Our architecture for Viterbi beam search prioritizes memory locality. The active state list is dynamically resized to allow quick sequential scans when the load factor is small. Hypotheses for the "source" (t) and "destination" ($t+1$) frames are stored together for each active state, providing an extra 13% area savings because most states remain active for several frames at a time. Variable-length WFST states are cached in a circular buffer, using a "least recently missed" eviction strategy to prevent fragmentation and maximize cache utilization. The maximum cacheable state length can be adjusted in 1B increments up to 255B, allowing the cache to be tuned for the task and memory characteristics. An on-chip word lattice captures word arcs generated during search, reducing write bandwidth by at least 8× relative to per-frame state list snapshots.

Previous work such as [4] provided micropower VADs that can be used in quiet environments or in applications that tolerate false alarms. In our application, false alarms will unnecessarily wake up a larger downstream system, increasing time-averaged power consumption and impacting the user experience. Hence, we prioritize VAD accuracy, even if it results in larger area and power for the VAD itself. Our test chip provides three VAD algorithms–energy-based (EB), harmonicity (HM), and modulation frequency (MF)–allowing us to evaluate the interaction of algorithm and circuit performance.

Our VAD architecture is shown in Fig. 14.4.4. The EB algorithm identifies frame-level energy differences and weights them with an estimate of the SNR. The HM algorithm identifies the periodic (voiced) component of the signal, which is a more robust indicator of speech presence than energy alone. The MF algorithm extracts features with a long temporal context and classifies them with a stripped-down version of the DNN evaluator described above; this proved to be the most robust algorithm in challenging noise conditions. The VAD supports downsampling (so ASR and VAD can use different sample rates) with a 65-tap FIR antialiasing filter, and buffers input samples for later retrieval by the ASR module (which is powered down until speech is detected).

Our ASR/VAD test chip is shown in Fig. 14.4.7. This chip performs all stages of ASR transcription from audio samples to text. ASR is functional with logic supply voltages from 0.60V (10.2MHz) to 1.20V (86.8MHz), and VAD is functional from 0.50V (1.68MHz) to 0.90V (47.8MHz). The core is partitioned into five voltage areas; SRAM can be operated at 0.15-0.20V above the logic supply (up to a limit of 1.20V) for best efficiency, or all five voltage areas may be powered from the same supply for simplicity. Latch-based clock gates were inserted explicitly at 76 locations in the design (Fig. 14.4.5), resulting in a 30-40% reduction in core power at full load, and reducing ASR clock tree overhead from 157pJ to 14pJ per cycle. The acoustic model requires 16-56pJ per nonzero neuron weight, depending on supply voltages. Search efficiency varies from 2.5-6.3nJ per hypothesis, compared to 16nJ in [5].

A summary of ASR and VAD performance results is shown in Fig. 14.4.6. A variety of ASR tasks, with vocabularies ranging from 11 words to 145k words, can be run in real-time on this chip. Core power scales by 45× from the easiest to the hardest task, and memory bandwidth scales by 136×. On the WSJ eval92-5k task that was demonstrated by [5], we obtained 4.1× fewer word errors (3.12% vs. 13.0%), 3.3× lower core power (1.78mW vs. 6.0mW), and 12.7× lower memory bandwidth (4.84MB/s vs. 61.6MB/s). Our framework is designed to interoperate with the open-source Kaldi tools [6], allowing software recognizers trained in Kaldi to quickly be ported to the hardware platform. We hope that these contributions will facilitate the deployment of high-quality speech interfaces in low-power devices.

Acknowledgements:
This work was funded by Quanta Computer via the Qmulus Project. The authors would like to thank the TSMC University Shuttle Program for providing chip fabrication.

References:
[1] M. Price, et al., "Memory-Efficient Modeling and Search Techniques for Hardware ASR Decoders," *Interspeech*, pp. 1893–1897, 2016.
[2] B. Moons, et al., "A 0.3-2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets," *IEEE Symp. VLSI Circuits*, 2016.
[3] M. Shah, et al., "A Fixed-Point Neural Network for Keyword Detection on Resource Constrained Hardware," *IEEE Int'l Workshop on Signal Processing Systems*, 2015.
[4] K. Badami, et al., "Context-Aware Hierarchical Information-Sensing in a 6 µW 90nm CMOS Voice Activity Detector," *ISSCC*, pp. 430-431, Feb. 2015.
[5] M. Price, et al., "A 6mW 5K-Word Real-Time Speech Recognizer Using WFST Models", *ISSCC*, pp. 454-455, 2014.
[6] D. Povey, et al., "The Kaldi Speech Recognition Toolkit," *IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

 978-1-5090-3758-2/17/$31.00

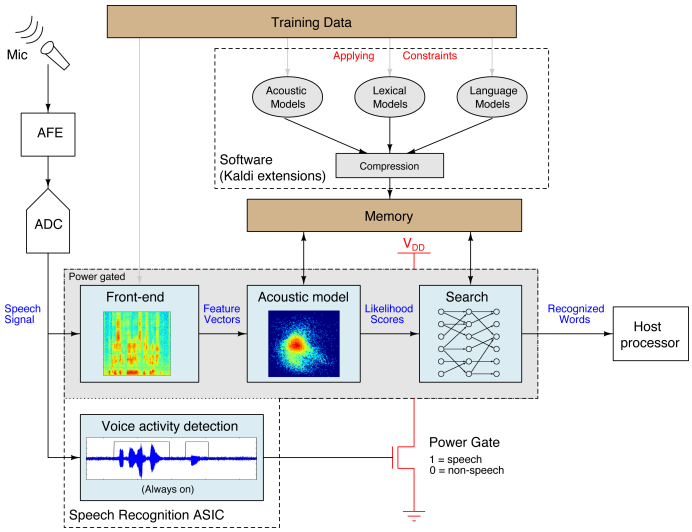


Figure 14.4.1: Power gated speech recognizer concept (gray region is power gated by VAD decision).

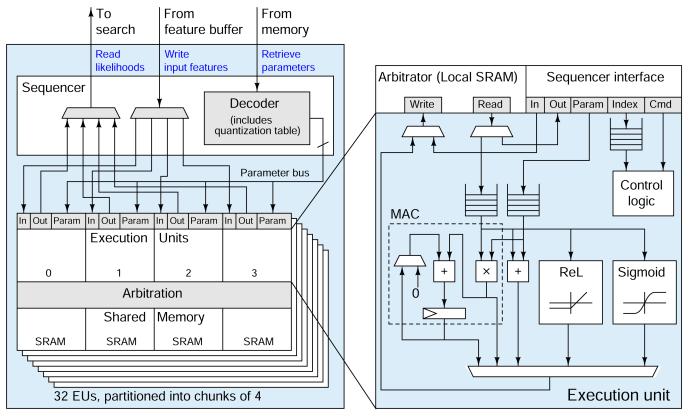


Figure 14.4.2: Block diagram of SIMD neural network evaluator.

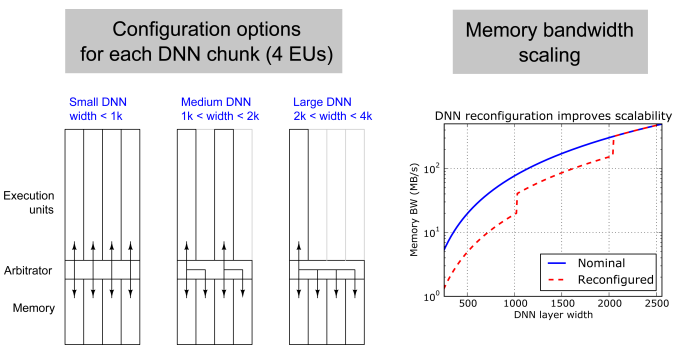
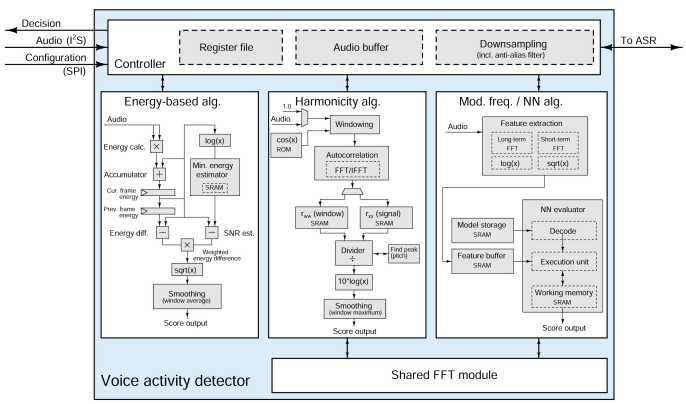


Figure 14.4.3: NN evaluator execution units are grouped into chunks that can be reconfigured for best utilization across different network sizes.



$$P_{\text{avg}} = p_{\text{VAD}} + \underbrace{[(1-p_M)D + p_{FA}(1-D)]\,p_{\text{downstream}}}_{\text{Contribution of the downstream system can be large if } p_{FA} > D}$$

Figure 14.4.4: VAD block diagram and system power model.

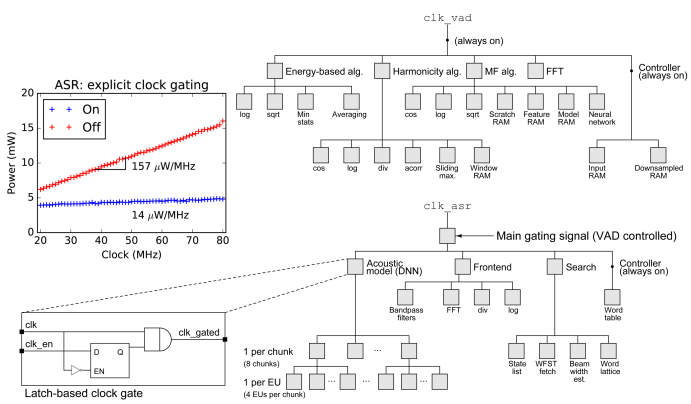


Figure 14.4.5: Explicit clock gating hierarchy: VAD clock domain (top), ASR clock domain (bottom), and measured impact on real-time ASR power versus clock frequency.

Automatic speech recognition (ASR)

Task	Vocab.	Clock	Mem. BW	WER	Power
Digits	11	3 MHz	0.11 MB/s	1.65%	172 μW
Weather	2k	23 MHz	10.1 MB/s	4.38%	4.70 mW
Food diary	7k	46 MHz	9.02 MB/s	8.57%	4.67 mW
News (1)	5k	15 MHz	4.84 MB/s	3.12%	1.78 mW
News (2)	145k	40 MHz	15.0 MB/s	8.78%	7.78 mW

Voice activity detection (VAD)

Algorithm	SNR for 10% frame level EER: White noise	Aurora2	Forklift	Power
Energy-based	−1 dB	18 dB	Fail	8.5 μW
Harmonicity	< −5 dB	5 dB	Fail	24.4 μW
Modulation frequencies	−2 dB	7 dB	1 dB	22.3 μW

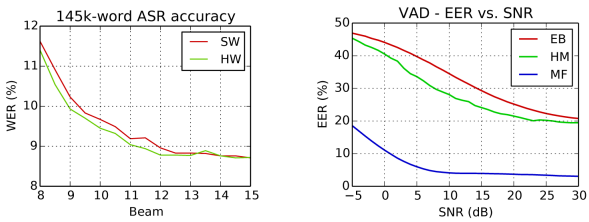
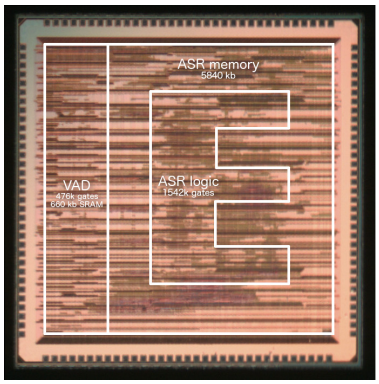


Figure 14.4.6: Summary of ASR/VAD test chip results, with highlighted results illustrated.



Specification	Value
Process	65 nm LP
Core size	3.1 × 3.1 mm
Die size	3.63 × 3.63 mm
Package	88-pin QFN
Logic gates	2088k (NAND2 equiv.)
SRAM	5.84 Mb
Supply voltage	0.60–1.20 V
Power consumption	1.8–7.8 mW (typ.)
Clock frequency	3–86 MHz
Neural network efficiency	16–56 pJ/neuron
Viterbi search efficiency	2.5–6.3 nJ/hypothesis

Figure 14.4.7: Die photo and table of key specifications.

978-1-5090-3758-2/17/$31.00