9.8 A 25mm² SoC for IoT Devices with 18ms Noise-Robust Speech-to-Text Latency via Bayesian Speech Denoising and Attention-Based Sequence-to-Sequence DNN Speech Recognition in 16nm FinFET

Thierry Tambe[1], En-Yu Yang[1], Glenn G. Ko[1], Yuji Chai[1], Coleman Hooper[1], Marco Donato[2], Paul N. Whatmough[1,3], Alexander M. Rush[4], David Brooks[1], Gu-Yeon Wei[1]

[1]Harvard University, Cambridge, MA
[2]Tufts University, Medford, MA
[3]ARM, Boston, MA
[4]Cornell University, New York, NY

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9366062

Automatic speech recognition (ASR) using deep learning is essential for user interfaces on IoT devices. However, previously published ASR chips [4-7] do not consider realistic operating conditions, which are typically noisy and may include more than one speaker. Furthermore, several of these works have implemented only small-vocabulary tasks, such as keyword-spotting (KWS), where context-blind deep neural network (DNN) algorithms are adequate. However, for large-vocabulary tasks (e.g., >100k words), the more complex bidirectional RNNs with an attention mechanism [1] provide context learning in long sequences, which improve ASR accuracy by up to 62% on the 200k-words LibriSpeech dataset, compared to a simpler unidirectional RNN (Fig. 9.8.1). Attention-based networks emphasize the most relevant parts of the source sequence during each decoding time step. In doing so, the encoder sequence is treated as a soft-addressable memory whose positions are weighted based on the state of the decoder RNN. Bidirectional RNNs learn past and future temporal information by concatenating forward and backward time steps.

This paper presents a 16nm SoC that executes a full speech-enhancing ASR pipeline in hardware, with the following key contributions: 1) unsupervised speech denoising using a Markov Source Separation Engine (MSSE) and 2) a reconfigurable accelerator (FlexASR) that demonstrates large-vocabulary sequence-to-sequence (seq2seq) ASR using bidirectional RNNs with attention. The full ASR pipeline (Fig. 9.8.1) pre-processes the incoming speech using an Arm Cortex-A53, then denoises the signal (up to 7.3dB SDR) in the MSSE accelerator, and finally accelerates a bidirectional attention-based speech-to-text model in the FlexASR accelerator. The 16nm test chip consumes 2.24mJ of energy per frame while achieving end-to-end latency of 18ms – enabling real-time throughput.

In the proposed speech-enhancing ASR pipeline, shown in Fig. 9.8.1, an always-on Arm M0 interfaces with an off-chip ADC to detect acoustic activity. The M0 autonomously monitors incoming audio amplitudes and subsequently boots the A53, MSSE, and FlexASR when the signal magnitude exceeds a threshold in order to reduce power. The dual-issue pipeline and SIMD datapath of the A53 efficiently compute the feature extraction tasks required to synthesize the spectrograms with overlapping 32ms frames. Then, the MSSE performs unsupervised real-time speech denoising by constructing and solving, from the input spectrograms, a 2D-grid probabilistic graphical model called a Markov Random Field (MRF) [2]. The MRF is solved using a Markov chain Monte Carlo method called Gibbs Sampling. The Bayesian algorithm particularly excels in a more dynamic environment, such as when sources are moving with respect to the microphones [3], which can potentially create underperforming corner cases for supervised methods where it is necessary to cover all scenarios with training data. MSSE ultimately produces a binary label corresponding to *noise* or *speech*. The A53 then convolves the *speech* label mask with the original spectrogram in order to extract the *clean* speech, which is subsequently accelerated in FlexASR using a bidirectional attention-based seq2seq DNN.

Figure 9.8.2 shows the overall SoC architecture comprising FlexASR, MSSE, an Arm Cortex-M0 microcontroller, and a dual-core Arm Cortex-A53 CPU cluster with 2MB L2$ (common in high performance embedded and mobile SoCs) connected together via AHB and AXI buses. MSSE utilizes 12 parallel Gibbs samplers to solve the spectrogram MRF. It is highly optimized for sound source separation and only supports binary label workloads, resulting in a shorter, faster (2×) pipeline and a more energy-efficient (2×) datapath compared to [2], which is a general-purpose Bayesian inference accelerator that supports 64 labels. FlexASR comprises four processing elements (PEs) and a multi-function global buffer (GB) unit, connected via broadcast and arbitrated crossbar channels.

Figure 9.8.3 describes the FlexASR PE, which contains a 1MB 16-bank weight buffer and a 4KB input activation buffer, feeding sixteen 8b floating-point vector MACs. Inputs and weights are stored in 8b floating-point format, with additional support for weight clustering via 4b indexes (2× compression). An activation unit performs vector operations on the accumulated results, composing LSTM, GRU, or vanilla RNN layers. Fig. 9.8.3 also depicts the custom tiling strategy wherein 16-by-16 weight blocks are rearranged and interleaved in the weight buffer for hazard-free computation in the activation unit.

Figure 9.8.4 shows the architecture of the FlexASR GB. It collects and unifies partial activated output states from the PEs across time steps in a 1MB 16-bank buffer. It also computes the attention mechanism, mean/max pooling, and normalization, all of which are operations in modern seq2seq networks. A 16KB auxiliary buffer stores seq2seq decoder RNN outputs, attention intermediate states, and weights of the normalization layer. Fig. 9.8.4 also details the computation of the attention mechanism. The encoder and decoder states are read from the GB's unified and auxiliary buffers, respectively, before a MAC generates scores processed by a *softmax* unit to produce the attention weights. To avoid numerical instability, the *softmax* is computed by subtracting the max score from its numerator and denominator. While computing attention, FlexASR saves energy and cycles by gating and skipping MAC operations whenever decoder states are null. *Concat*, *sum*, and *average* merge modes, required for bidirectional RNN operations, are supported by striping forward and backward time steps across alternate banks in the 1MB buffer. This enables seamless concatenation of the bidirectional activations.

Figure 9.8.5 demonstrates the accuracy and performance benefits of the proposed speech-enhancing ASR pipeline by comparing four inference scenarios using the LibriSpeech dataset: (A) with noiseless audio of the speaker, (B) with noise mixed with the speaker's voice in a simulated room environment at a signal-to-noise ratio (SNR) of 0.90dB, (C) with a much larger ASR model (22MB) trained with a noise-corrupted LibriSpeech dataset in order to accommodate noisy inputs, and (D) with the proposed pipeline using Bayesian speech enhancement to separate the noise from the speaker's voice. By denoising incoming signals prior to speech recognition, MSSE allows FlexASR to store a much smaller ASR model (1/6×, i.e., 3.5MB), which obviates the very inefficient strategy of scaling up the DNN model (C) in order to achieve noise robustness. Notably, the proposed pipeline achieves 4.3× lower latency, and 7× energy improvement compared to the similarly-accurate "Noisy+Big" case (C), which requires significant off-chip data movements because the upsized ASR model cannot fit on-chip. Furthermore, the proposed ASR pipeline delivers 3× accuracy improvement over the unseparated noise case (B) and is within 1% of the clean-input-baseline case (A). Fig. 9.8.5 also shows that commercial edge platforms fail to provide real-time performance as their per-frame latencies exceed the 32ms frame length, despite substantial energy expenditures.

Figure 9.8.6 shows that voltage scaling on FlexASR and MSSE produces efficiency ranges of 2.6-7.8TFLOPs/W and 4.33-17.6GS/s/W, respectively. The per-frame, end-to-end latency varies from 45ms to 15ms as the SoC voltage scales from 0.55-1.0V, while consuming 19-227mW. Compared to other speech processing chips (Fig. 9.8.6), this is the first work to demonstrate on-chip support for denoised, large-vocabulary ASR using state-of-the-art bidirectional attention-based speech recognition that enables substantial WER benefits – all together, while demonstrating competitive 18ms per-frame end-to-end latency. The annotated die and physical layout photos along with the chip summary are shown in Fig 9.8.7.

Acknowledgement:
This work is supported in part by JUMP ADA, DARPA CRAFT and DSSoC programs, NSF Awards 1704834 and 1718160, Intel Corp., and Arm Inc. We thank B. Khailany, R. Venkatesan, B. Keller, and Y. Shao (Nvidia); and U. Gupta, L. Pentecost, and V. Reddi (Harvard); and S. Garg (Mentor) for helpful discussions.

References:
[1] D. Badhanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR*, 2015.
[2] G. Ko et al., "A 3mm² Programmable Bayesian Inference Accelerator for Unsupervised Machine Perception using Parallel Gibbs Sampling in 16nm," *IEEE Symp. VLSI Circuits*, 2020.
[3] M. Kim et al., "Stereophonic Spectrogram Segmentation Using Markov Random Fields," *IEEE MLSP*, 2012.
[4] M. Price et al., "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks," *IEEE JSSC*, vol. 53, no. 1, pp. 66-75, Jan. 2018.
[5] R. Guo et al., "A 5.1pJ/Neuron 127.3us/Inference RNN-based Speech Recognition Processor using 16 Computing-in-Memory SRAM Macros in 65nm CMOS," *IEEE Symp. VLSI Circuits*, pp. C120-C121, 2019.
[6] J. Giraldo et al., "18μW SoC for Near-Microphone Keyword Spotting and Speaker Verification," *IEEE Symp. VLSI Circuits*, pp. C52-C53, 2019.
[7] S. Yin et al., "A 141 μW, 2.46 pJ/Neuron Binarized Convolutional Neural Network based Self-Learning Speech Recognition Processor in 28nm CMOS," *IEEE Symp. VLSI Circuits*, pp. 139-140, 2018.

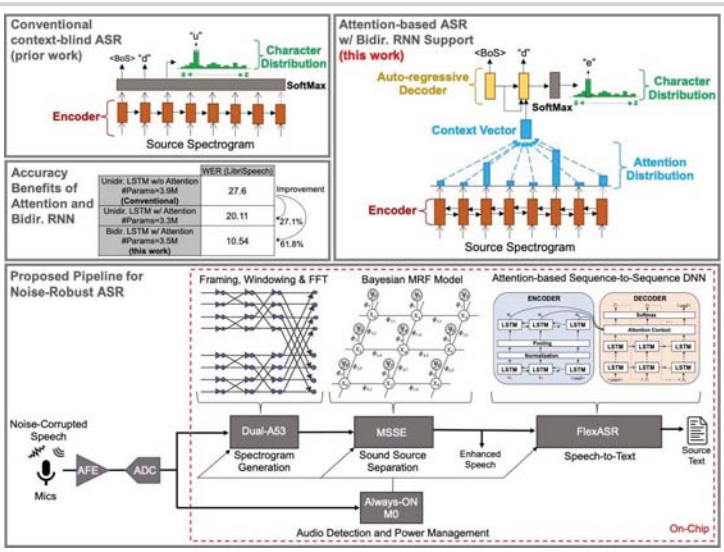


Figure 9.8.1: Context-blind ASR vs. attention-based bidirectional ASR producing significant WER improvements. Noise-isolating ASR is accelerated on-chip using Bayesian Gibbs Sampling for denoising and attention-based DNNs.

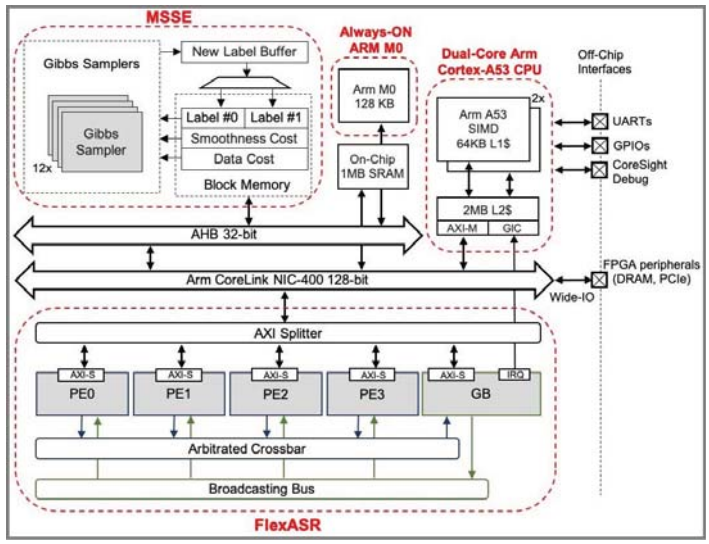


Figure 9.8.2: 16nm SoC architecture, highlighting a dual-core Arm Cortex-A53, the always-on Arm M0, the MRF source separation engine (MSSE), and attention-based ASR accelerator (FlexASR).

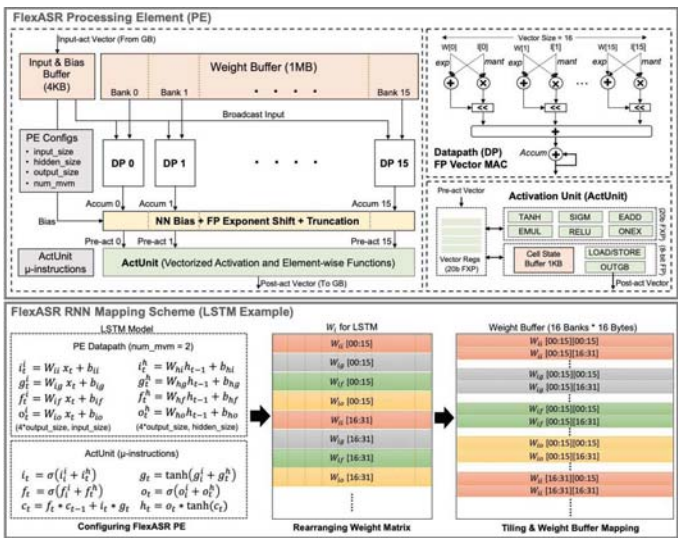


Figure 9.8.3: FlexASR PE with floating-point datapath. RNN weight tiles ($W_{i/h}$) are rearranged and interleaved in the weight buffer for hazard-free computation in the activation unit.

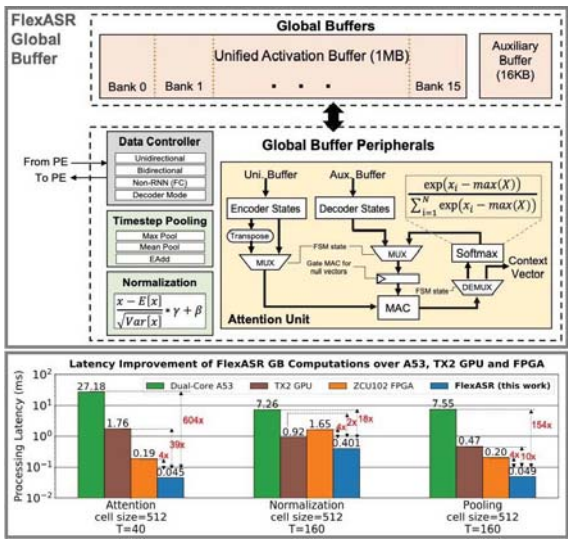


Figure 9.8.4: FlexASR multi-function global buffer (GB) with attention datapath. FlexASR GB computes the attention mechanism with a 4-to-604× speedup over commercial platforms, including CPU, GPU and FPGA.

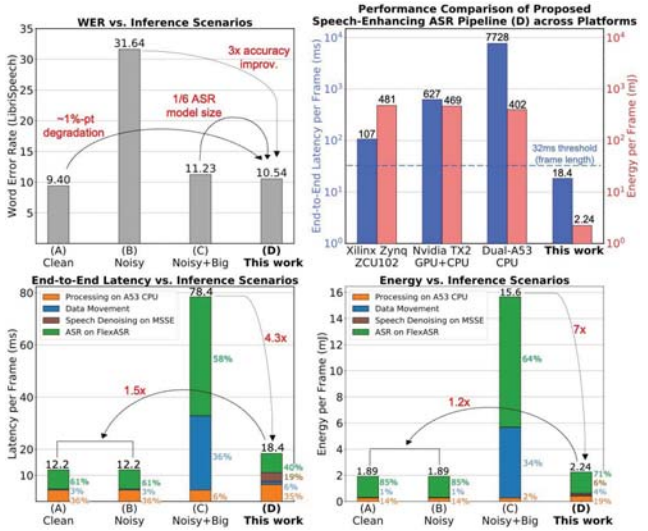
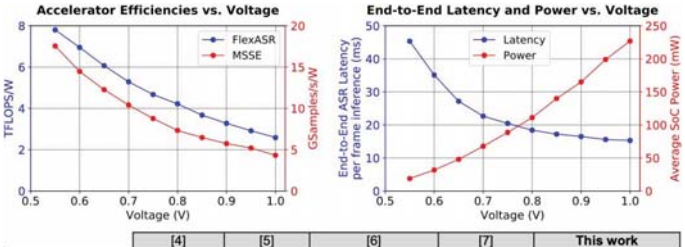


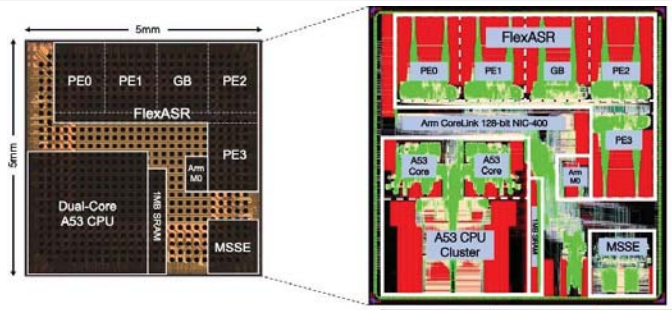
Figure 9.8.5: Summary of measurement results for ASR inference with (A) clean input audio, (B) noisy input audio, (C) noisy input audio using a 6× larger ASR model, and (D) this work - noisy input audio with Bayesian denoising.



	[4]	[5]	[6]	[7]	This work
Technology	65 nm	65 nm	65 nm	28 nm	16 nm
Core Dimension	9.6 mm²	6.2 mm²	2.6 mm²	1.3 mm²	21.8 mm²
Application	ASR	KWS	KWS	ASR	Speech Denoising, ASR
Algorithm	HMM	RNN	RNN	CNN	Bayesian MRF + Attention-based RNN
On-Chip Speech Denoising	No	No	No	No	Yes (7.3 dB SDR)
Dataset (Vocabulary Size)	News 2 (145k words)	Smart Home (11 words)	GSCD (30 words)	TIMIT (6k words)	LibriSpeech (200k words)
Datatype	4-12b FxP	1b FxP	4b/8b FxP	1b FxP	Denoising: 32b FxP ASR: 8b FP
Total SRAM	730 KB	18 KB	105 KB	52 KB	9.8 MB
Supply Voltage	0.6 V – 1.2 V	0.9 V – 1.1 V	0.6 V – 1.2V	0.57 V – 0.9 V	0.55 V – 1.0 V
Frequency	3 – 86 MHz	5 – 75 MHz	250 kHz – 12.5 MHz	2.5 – 50 MHz	130 – 775 MHz
Latency per Frame	/	0.127 ms	16 ms	0.5 ms – 25 ms	15 – 45 ms
Power	7.78 mW @ 0.9V/40MHz	26 mW @ 0.9V/75MHz	18.3 uW @ 0.6V/250 KHz	1.42 mW @ 0.58V/20MHz	111 mW @ 0.8V/Fmax *

*: Average SoC power measured while running FlexASR, MSSE, and A53 at their max frequencies of 440MHz, 544MHz, and 643MHz, respectively, at 0.8 V.

Figure 9.8.6: Accelerator performance and end-to-end ASR latency vs. scaled voltage, and comparison table.



Technology	TSMC 16nm FFC
Die Area	25mm^2
Total SRAM	9.8MB
Gate Count	11M
Clock Domains	6
Power Domains	5
Supply Voltage	0.55 – 1V
Packaging	Flip-chip BGA-672

	Dual-A53	MSSE	FlexASR
Workloads	FFT, Application Logic	Markov Random Field	Attention-based seq2seq models
Data Type	FP64	FxP32	FP8
Area	6.21mm^2	1.31mm^2	8.84mm^2
SRAM	2.41MB	0.103MB	5.03MB
Voltage	0.55 - 1V	0.55 - 1V	0.55 - 1V
Frequency	354 – 775MHz	287 – 651MHz	130 – 573MHz
Power @ Fmax/0.8V	50.4mW	42.2mW	214mW
Power Eff.	-	4.33 – 17.6 Gsamples/s/W	2.6 – 7.8 TFLOPS/W

Figure 9.8.7: Annotated die photo, physical layout, chip summary, and breakdown of the key components that implement the full speech-enhancing ASR pipeline via a combination of software and custom hardware.

978-1-7281-9549-0/21/$31.00 ©2021 IEEE