Leveraging Run-Time Feedback for Efficient ASR Acceleration

Reza Yazdani, Jose-Maria Arnau, Antonio González
Department of Computer Architecture
Universitat Politécnica de Catalunya (UPC), Barcelona, Spain
{*ryazdani, jarnau, antonio*}*@ac.upc.edu*

***Abstract*—In this work, we propose Locality-AWare-Scheme (LAWS) for an Automatic Speech Recognition (ASR) accelerator in order to significantly reduce its energy consumption and memory requirements, by leveraging the locality among consecutive segments of the speech signal. LAWS diminishes ASR's workload by up to 60% by removing most of the off-chip accesses during the ASR's decoding process. We furthermore improve LAWS's effectiveness by selectively adapting the amount of ASR's workload, based on run-time feedback. In particular, we exploit the fact that the confidence of the ASR system varies along the recognition process. When confidence is high, the ASR system can be more restrictive and reduce the amount of work.**

The end design provides a saving of 87% in memory requests, 2.3x reduction in energy consumption, and a speedup of 2.1x with respect to the state-of-the-art ASR accelerator.

***Keywords*-Automatic Speech Recognition (ASR), Viterbi Search, Hardware Accelerator, Memory-Efficient**

I. Introduction

After achieving human parity in speech recognition [1], a main focus of Automatic Speech Recognition (ASR) systems is turning towards mobile, wearables and IoT devices. Such a high degree of accuracy comes at the expense of huge memory requirements and computational cost in terms of performance and energy [2], which is unaffordable in most of these devices. Several accelerators have been recently proposed to target some challenges of ASR [2], [3], [4]. However, the main challenges of high energy-consumption and memory-bandwidth requirements still remain when deploying ASR in small form-factor battery-operated devices.

A state-of-the-art speech recognition accelerator requires an average memory throughput of 16 Gb/s [2] in order to run ASR on different speech corpora. On the other hand, IoT and wearable devices normally use low-power memory technologies with limited throughput, such as NAND flash memories [5]. As reported by Micron, these memory systems can achieve a maximum bandwidth of 6 Gb/s [6]. As a result, the ASR's memory management needs significant improvement in order to be deployed on these devices. Furthermore, each DRAM memory access consumes nearly three orders of magnitude more energy than a typical computation or on-chip memory accesses. Therefore, high memory requirement is also the main energy bottleneck for ASR systems.

Each step of an ASR system processes one frame of typically 10 ms of the speech signal. For each frame, it expands multiple nodes in a search graph, called hypotheses, to decode the speech by finding the most likely path on this graph. ASR systems normally use a beam in order to control the search space. For each frame, it keeps track of the best hypothesis and those hypotheses whose score (likelihood) is lower than the best score minus the beam are discarded.

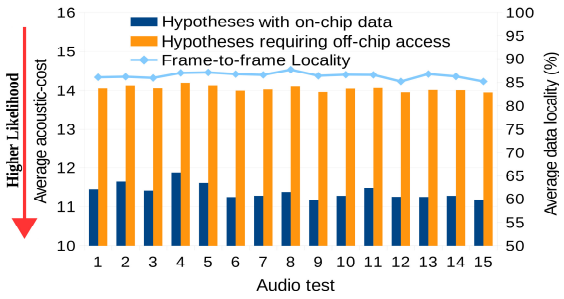

Figure 1. The bars show the average acoustic-cost of hypotheses grouped based on their data availability on-chip. We also shows the percentage of states and arcs that are the same in two consecutive frames (data locality).

Figure 1 shows the average acoustic-scores, i.e. log-space probability, of the different hypotheses, for 15 audio tests selected from different speakers of Librispeech corpus [7]. In addition, we show the average frame-to-frame locality for each audio test using the UNFOLD's architecture parameters [2], which shows nearly 86% of data-locality in processing subsequent frames. As depicted, hypotheses whose data is on-chip have considerably lower scores, i.e. higher probability, than the ones explored for the first time in each frame, whose data requires an off-chip memory access.

II. Locality-Aware Scheme for ASR Acceleration

We propose a Locality-AWare Scheme (LAWS) for ASR, which dynamically adjusts the beam distance based on the hypotheses' data-locality. In other words, our scheme uses the locality and hypotheses' likelihood to decide the search strategy, unlike previous schemes that use only likelihood information. This results in a more efficient expansion on the search graph in terms of accuracy and energy consumption with negligible impact on accuracy. By applying our techniques on a state-of-the-art ASR accelerator, we obtain 2.1x speedup and 2.3x energy savings with a small area overhead of less than 0.1% in total accelerator's design. In addition, by removing most of the memory activity due to discarding

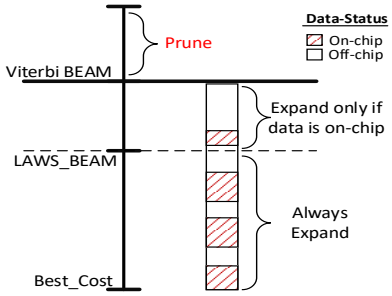

Figure 2. Viterbi expansion under Locality-AWare Scheme.

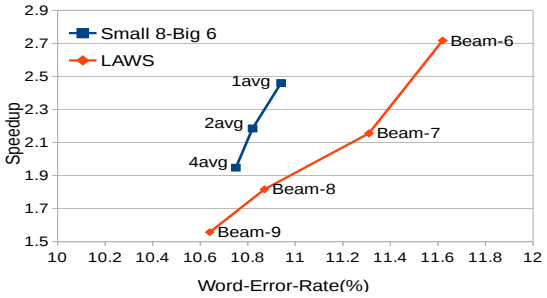

Figure 3. Speedup versus Word-Error-Rate (WER), for LAWS and $Small_8$-Big_6, compared to UNFLOD [2]. The baseline WER is 10.62%.

most of the hypotheses requiring an off-chip memory access, we save around 74% of the memory bandwidth.

Figure 2 illustrates the Viterbi search expansion for both the conventional beam search and our Locality-AWare Scheme (LAWS). In the conventional approach, the hypothesis with the best cost is identified on every frame. Hypotheses whose distance to the best hypothesis is smaller than a given threshold are explored, whereas the rest are discarded since they are very unlikely. Our scheme is different since we consider both the cost (likelihood) of the hypotheses and their temporal locality to prune the search space.

LAWS is highly effective at saving energy consumption and main memory bandwidth by significantly reducing ASR's workload. Figure 3 shows the speedup versus accuracy obtained by LAWS. As illustrated, in spite of increasing performance, WER gets worse when decreasing the LAWS_Beam beyond 9 (this WER is similar to the baseline WER). In order to handle this problem, LAWS uses a dynamic beam adaptation policy, which decides on the beam distance by taking into account the confidence of the ASR system at each frame. There are some frames for which there are many hypotheses with scores very close to the best, whereas there are fewer in other frames. When the number hypotheses close to the best is high, the ASR system has low confidence, since there are many alternatives similar to the best, whereas the confidence is high when this number is low.

After exploring different numbers of LAWS_Beam, we find out that a dual-beam selection mechanism reaches the best trade-off between performance, area, and accuracy. We refer to our technique as $Small_8$-Big_6, which selects the beams 8 and 6 for the frames with the number of hypotheses smaller and bigger than the specified threshold, respectively. Moreover, we use the dynamic-average number of explored hypotheses, in order to decide between different confidence regions. As depicted by Figure 3, $Small_8$-Big_6 obtains a speedup of 2.2x (using *2 x average* confidence threshold) with a negligible increase of 0.2% in WER, whereas by using a single beam, we lose almost 0.85% of accuracy to achieve the same performance improvement.

In addition to the performance speedup, LAWS achieves significant savings in both the energy-consumption and memory-bandwidth, due to removing most of the memory fetches required for the hypotheses whose data is off-chip. The energy-consumption decreases by 2.3x, while the bandwidth requirement gets as low as 381 MB/s, being reduced by almost 74% when compared to the baseline.

III. Conclusions

In this work, we target one of the main challenges for ASR systems for mobile, IoT and wearable devices, which is the high memory and energy requirements to perform speech recognition. We present LAWS, a scheme that obtains high benefits in workload reduction without compromising accuracy through a new approach that combines novel insights on data-locality characteristics of ASR with the statistical scores computed during the Viterbi search. LAWS removes over 87% of the off-chip memory activity, which improves performance and energy consumption by 2.1x and 2.3x, respectively.

Acknowledgment

This work has been supported by the the CoCoUnit ERC Advanced Grant of the EU's Horizon 2020 program (grant No 833057), the Spanish State Research Agency under grant TIN2016-75344-R (AEI/FEDER, EU), and the ICREA Academia program.

References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *CoRR*, 2016.

[2] R. Yazdani, J.-M. Arnau, and A. González, "Unfold: A memory-efficient speech recognizer using on-the-fly wfst composition," ser. MICRO-50 '17, 2017, pp. 69–81.

[3] R. Yazdani, A. Segura, J.-M. Arnau, and A. Gonzalez, "An ultra low-power hardware accelerator for automatic speech recognition," in *MICRO'49*, Oct 2016, pp. 1–12.

[4] R. Yazdani, A. Segura, J. M. Arnau, and A. Gonzalez, "Low-power automatic speech recognition through a mobile gpu and a viterbi accelerator," *IEEE Micro*, vol. 37, no. 1, pp. 22–29, Jan 2017.

[5] "Iot and wearable devices mean rethinking memory design," *Micron Technology*, 2014.

[6] "Nor/nand flash guide," *Micron Technology*, 2017.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, April 2015, pp. 5206–5210.