FAST AND POWER EFFICIENT HARDWARE-ACCELERATED CLOUD-BASED ASR FOR REMOTE DIALOG APPLICATIONS

Alexei V. Ivanov[†][‡], Patrick L. Lange[†] and David Suendermann-Oeft[†]

[†] Educational Testing Service R&D, 90 New Montgomery St, # 1500, San Francisco, CA, USA
[‡] Verbumware Inc., San Jose, CA, USA

ABSTRACT

We demonstrate a GPU-based implementation of an automated speech recognition system that is massively faster, sometimes significantly more accurate and more power-efficient than a modern CPU-based open-source reference. This technology enables speech solution providers to efficiently up-scale their operation to the consumer market. Our GPU-based speech recognition platform supports statistical models, created with publicly available speech toolkits. The platform is cloud-based and capable of supporting online spoken interaction with remote interlocutors as a mass service.

Index Terms— hardware-accelerated speech recognition; dialog interaction with remote interlocutors.

1. INTRODUCTION

Apart from the accuracy, a successful mass automated speech recognition (ASR) service requires excellent processing speed and energy efficiency. For such use cases as serving a spoken dialog system (SDS) and mining user's audio data for specific keywords, the processing speed is even more important [1].

Processing speedup is achievable via committing larger areas of the die for solving a single task. In the multi-core CPU programming model that is accomplished by construction of multi-threaded programs that are sharing common data. Graphical processing units (GPUs) allow for an easier processing resource management. The GPU chip lacks extensive control logic making it potentially more efficient. The downside is increase of the programming effort.

Early attempts at parallelization of speech recognition with CPUs [2] and GPUs [3, 4] produced encouraging results. There are hybrid realizations involving multi-stage processing via rescoring [5]. In contrast our system is a completely GPU-based speech recognizer [6] that is capable of outperforming the CPU implementation even in the full computational occupancy case [7].
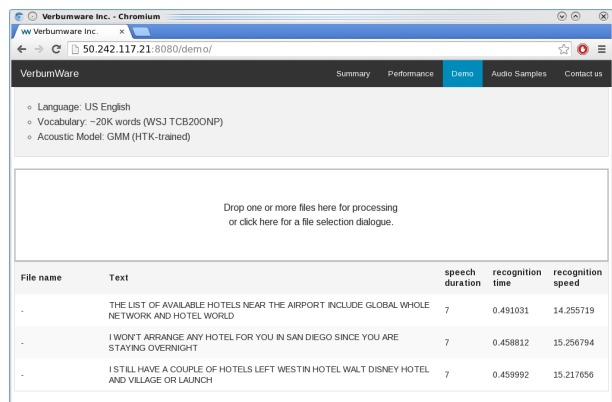
In the following sections we summarize the observed results and outline the system demonstration. The demonstration is performed via accessing the ASR web interface in a general purpose browser (see Fig.1 for a screen shot). Sample audio files are dragged onto the designated area of the web page for subsequent local playback and recognition by the remote cloud-based ASR server.

2. SYSTEM PERFORMANCE

The "NOV'92" and "NOV'93" evaluation sets of the Wall Street Journal spoken corpus are chosen to illustrate system performance on a range of tasks: 5K word vocabulary of NOV'92 and 20K word vocabulary with a high OOV rate of the DARPA NOV'93 evaluation. We have used the following language models: "bcb05onp" – bi-gram, 5K word "open" vocabulary, "bcb05cnp" – bi-gram, 5K word "closed" vocabulary, "tcb20onp" – tri-gram, 20K word "open" vocabulary.

The measurements are done with the following systems: an Intel Core i7-4930K server system equipped with TITAN X; a mobile system with NVIDIA Tegra K1 processing unit. The TITAN-equipped system is available for public evaluation[1].

Fig. 1. Speech recognition web interface.

Table 1 summarizes recognition results of the deep neural network (DNN) acoustic model trained with the Kaldi toolkit. The GPU-enabled engine accuracy is approximately equal to that of the open-source baseline. There is a small fluctuation of the actual Word Error Rate (WER) due to the arithmetic implementation differences. For a single-channel recognition the TITAN-enabled engine is ≈ 7 times faster than the reference. This is important in tasks like serving ASR to an SDS

---

[1]http://verbumware.org:8080/demo/

Task / LM	bcb05onp	bcb05cnp	bcb05onp	bcb05cnp	tcb20onp	bcb05onp	bcb05cnp	tcb20onp
NOV'92 WER	5.66%	2.30%	5.66%	2.30%	1.85%	5.77%	2.19%	1.63%
NOV'92 1/xRT	2.15	2.14	30.58	30.49	27.47	5.08	5.26	4.54
NOV'93 WER	18.22%	19.99%	18.22%	19.99%	7.77%	18.13%	20.19%	7.63%
NOV'93 1/xRT	2.15	2.15	30.12	30.21	26.67	4.33	4.20	3.90
Power, W/RTchan.	3.6		9			15		
Hardware	Tegra K1 (32 bit)		GeForce GTX TITAN BLACK			i7-4930K @3.4GHz		
	Hardware-accelerated					nnet-latgen-faster		

Table 1. Performance of the hardware accelerated version of the DNN-HMM system vs Nnet-latgen-faster decoder baseline.

Task / LM	bcb05onp	bcb05cnp	tcb20onp	bcb05onp	bcb05cnp	tcb20onp
NOV'92 WER	9.27%	5.47%	4.50%	9.30%	5.59%	5.64%
NOV'92 1/xRT	16.67	16.76	15.25	0.30	0.30	0.25
NOV'93 WER	28.05%	26.72%	11.66%	27.88%	26.46%	13.37%
NOV'93 1/xRT	16.49	16.58	15.04	0.25	0.25	0.19
Power, W/RTchan.	20			260		
Hardware	GeForce GTX TITAN X			i7-4930K @3.4GHz		
	Hardware-accelerated			HTK HDecode		

Table 2. Performance of the hardware accelerated version of the GMM-HMM system vs HTK HDecode.

or media-mining for specific spoken events. The implementation in the mobile device (NVIDIA Tegra K1) enables twice faster than real-time (RT) processing without any degradation in accuracy. The GPU-enabled engine allows unprecedented energy efficiency. The value of 15W per one RT channel was estimated while the CPU was fully loaded with 12 concurrent recognition jobs. That is the most power efficient manner of CPU utilization [7]. The TITAN-enabled server does better (9 W per one RT channel) while maintaining its processing speed. The Tegra-based solution is several times more efficient (3.6 W per one RT channel). The power per one RT channel was estimated as system's average cumulative consumption adjusted by the processing realtime factor and the number of concurrent processes. Power consumption and recognition speed of the GPU-based solution are linearly proportional to the system's load. On the contrary, the CPU consumes much more energy (per channel) when operating at the maximum pace, i.e. working on a single channel.

A Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) system behaves similarly. Comparison of the GPU implementation to the HTK toolkit reference is given in Table 2. The bi-gram LM accuracy is roughly the same. With the tri-gram LM (tcb20onp) accuracy is significantly better. The hardware-accelerated speech recognition system is massively faster (from 55 to 80 times). The hardware-accelerated recognition speed with the optimal parameter set does not depend much on the total search network size, which suggests efficiency of the search implementation.

3. CONCLUSIONS

The experiment presented in this paper has proven our conjecture about the possibility to build a faster ASR via committing larger areas of the die for solving a single speech recognition task. The GPU utilization (according to the nvidia-smi utility) is about 85 %.

The resulting system is suitable for faster than RT serving remote SDS applications from the cloud. Alternatively, the same system can be used for achieving RT processing speed with more complex statistical models.

4. ACKNOWLEDGMENTS

5. REFERENCES

[1] P.L. Lange, A.V. Ivanov, D. Suendermann-Oeft, V. Ramanarayanan, Y. Qian, and J.Tao, "Designing a cloud-based and open-source speech recognition server for dialog applications," in *submitted to IEEE ICASSP*, Shanghai, China, 2016.

[2] F. Brugnara, "A multithreaded implementation of Viterbi decoding on recursive transition networks," in *Proceedings of INTERSPEECH'2011*.

[3] J. Chong, Y. Yi, A. Faria, N.R. Satish, and K. Keutzer, "Data-parallel large vocabulary continuous speech recognition on graphics processors," *Electrical Engineering and Computer Sciences University of California at Berkeley, Technical Report No. 69*, 2008.

[4] K. You, J. Chong, Y. Yi, E. Gonina, C.J. Hughes, Y.-K. Chen, W. Sung, and K. Keutzer, "Parallel scalability in speech recognition," *IEEE Signal Processing Magazine*, pp. 124–135, November 2009.

[5] J. Kim and I. Lane, "Accelerating large vocabulary continuous speech recognition on heterogeneous CPU-GPU platforms," in *Proceedings of ICASSP'2014*. IEEE, 2014, pp. 3291–3295.

[6] "A.V. Ivanov and F. Brugnara", "Making it fast and reliable: Speech recognition with GPUs by sequential utilization of available knowledge sources," in *Proceedings of the GPU Technology Conference (GTC'2014)*. 2014, NVIDIA Corp.

[7] A.V. Ivanov, "Speech recognition on GPUs with open-source models: faster, better, cheaper," in *Proceedings of the GPU Technology Conference (GTC'2015)*. 2015, NVIDIA Corp.