CATEGORY: **SIGNAL & AUDIO PROCESSING - SA03**

POSTER **P5169** | CONTACT NAME **Alexei V. Ivanov: alexei_v_ivanov@ieee.org**

GPU TECHNOLOGY CONFERENCE

Verbumware Inc.

http://verbumware.net

Verbum Sat Sapienti Est

Speech Recognition on GPUs with Open-Source Models: Faster, Better, Cheaper

Alexei V. Ivanov
Verbumware Inc., San Jose, USA
alexei_v_ivanov@ieee.org, info@verbumware.net

Abstract

•We explore the possibility of using GPUs in automated speech recognition with statistical models, prepared by the open source tool-kits. Our observations reveal that the GPU implementation is massively faster, sometimes is significantly more accurate than the original open-source reference and, remarkably, consumes less power even if compared to the modern multi-core CPU, that is fully loaded with multiple concurrent recognition jobs. This technology enables speech solution providers to efficiently up-scale their operation to the mass market.

•Our GPU-based speech recognition platform can work with models, that are created with publicly available speech tool-kits (Kaldi, HTK, Sphinx, SRI LM toolkit).

Experiment Setup

•WER (Word Error Rate) is defined as a ratio of the total number of errors (there might be errors of three kinds: substitutions of one word with another, deletions of valid words; insertions of spurious hypothesized words) to the total number of true words that shall be recognized.

•xRT (a "real time factor") is measured as a ratio of the total processing time to the processed audio length. The inverse xRT (1/xRT) reflects how much faster the system can process audio compared to the natural speaking rate, e.g. with the HTK model and TCB20ONP LM in the NOV'93 task our system works ~ 12 times faster than people can speak.

•Test materials, the "NOV'92 (5K)" and "NOV'93" evaluation sets, were chosen to illustrate system's performance over the broad range of tasks. This range spans from the easiest level (5K word vocabulary) of NOV'92 to a more difficult one (20K word vocabulary with a high OOV rate) of the official DARPA NOV'93 evaluation.

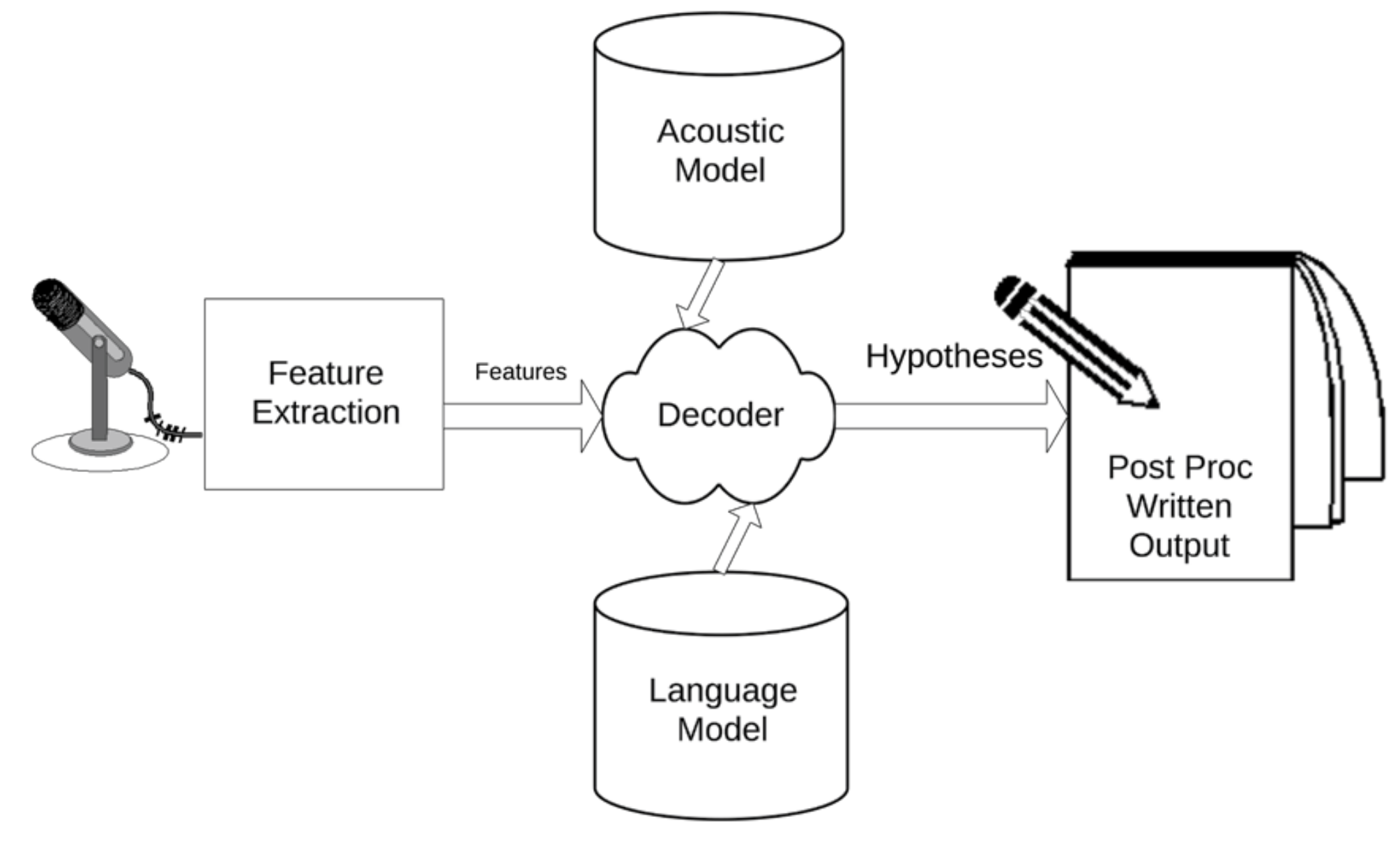
•BCB05ONP - bi-gram LM, 5K word "open" vocabulary, without verbalization of punctuation.

•BCB05CNP - bi-gram LM, 5K word "closed" vocabulary, without verbalization of punctuation.

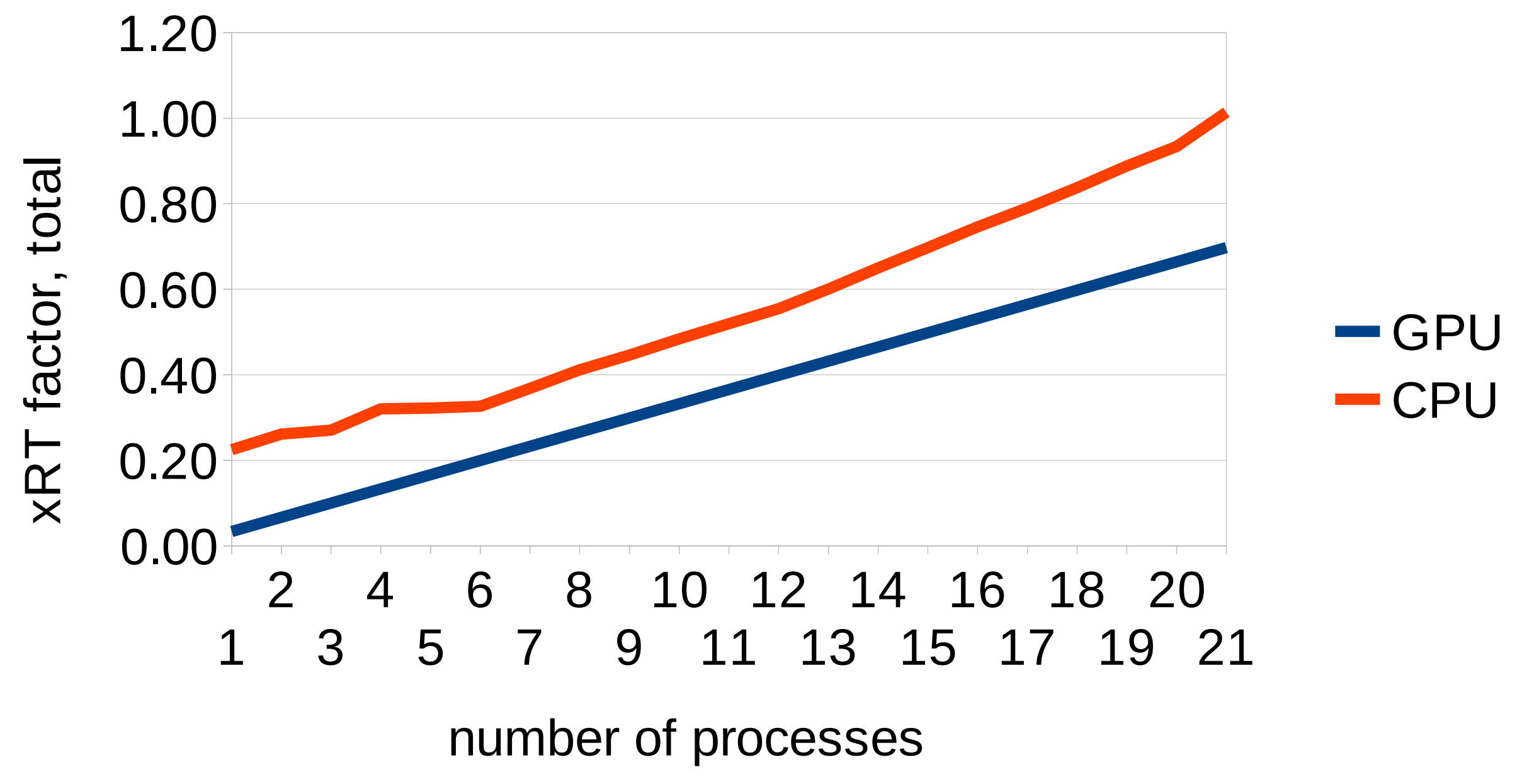
•TCB20ONP - tri-gram LM, 20K word "open" vocabulary, without verbalization of punctuation.

•Measurements were done with the following systems:
 1. The TITAN-equipped server system:
- CPU: Intel Core i7-4930K CPU @3.40GHz;
- GPU: Nvidia GeForce GTX TITAN BLACK 6144MB@980 Mhz // CUDA 6.5;
- Host RAM Total: 32888512 kB (32 GB);
- Operating System: Ubuntu 14.04.

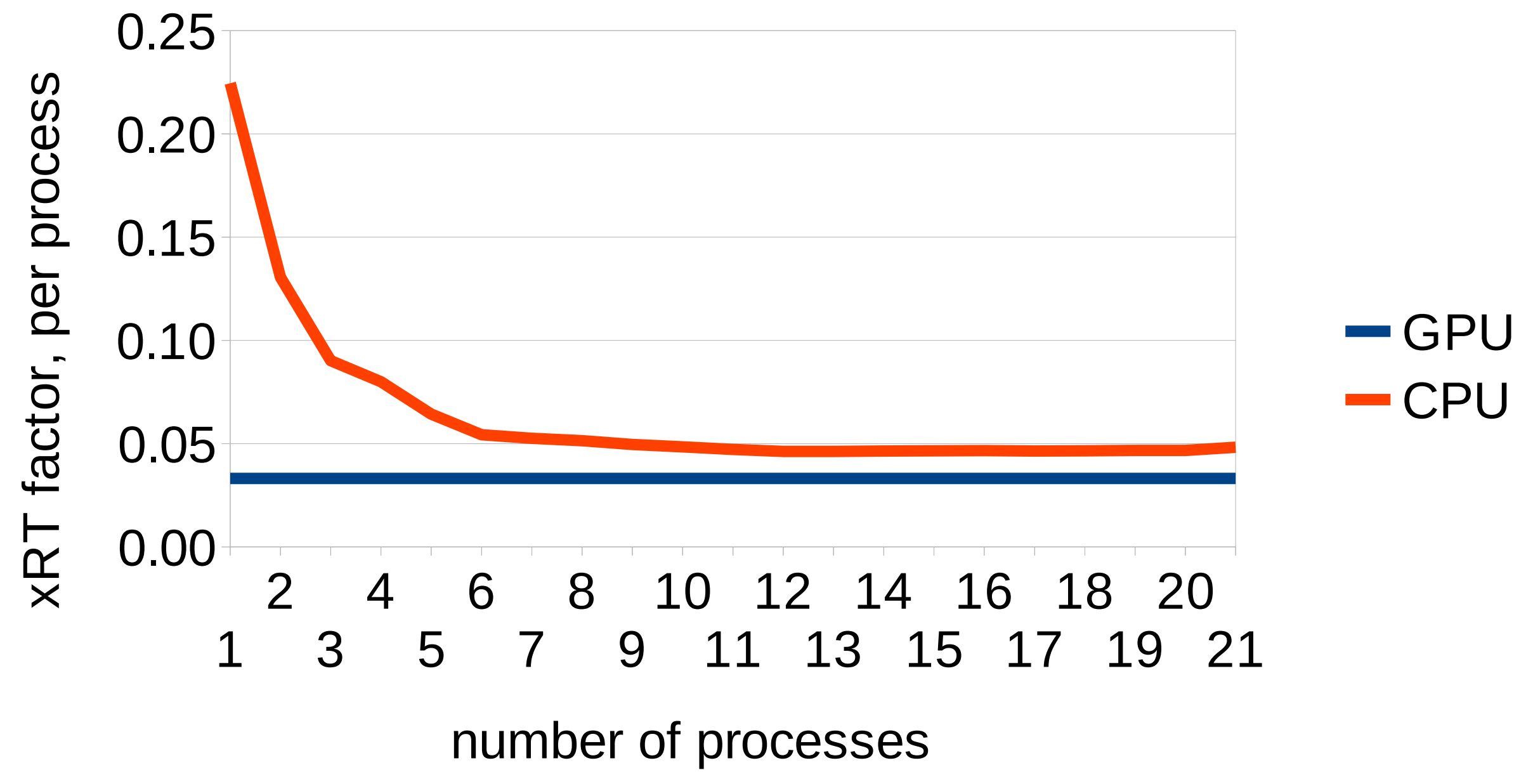
•2. The mobile system:
- NVIDIA Tegra K1 32-bit Processing Unit with
- GK20A GPU @ 852 MHz
- ARMv7 Processor rev 3 (v7l) 4+1 CPU Cores
- RAM Total: 1788136 kB (1.7 Gb)
- Operating System: Linux for Tegra v.19.2



ASR System Structure



Cumulative xRT



Specific xRT

Verbumware DNN-WFST System vs Kaldi Nnet-latgen-faster

TASKS\LMs	BCB05ONP	BCB05CNP	BCB05ONP	BCB05CNP	TCB20ONP	BCB05ONP	BCB05CNP	TCB20ONP
NOV'92 (5K) WER	**5.66%**	2.30%	**5.66%**	2.30%	1.85%	5.77%	2.19%	1.63%
NOV'92 (5K) xRT	0.4647	0.4683	**0.0327**	**0.0328**	**0.0364**	0.1967	0.1900	0.2203
NOV'93 WER	18.22%	**19.99%**	18.22%	**19.99%**	7.77%	18.13%	20.19%	7.63%
NOV'93 xRT	0.4658	0.4651	**0.0332**	**0.0331**	**0.0375**	0.2309	0.2382	0.2562
Power/RTchan.	**~3.6W**		**~9 W**			~ 15 W		
Hardware	**Tegra K1 (32 bit)**		**GeForce GTX TITAN BLACK**			**i7-4930K @3.40GHz**		
	Verbumware DNN-HMM System					**Nnet-latgen-faster**		

DNN-WFST Kaldi

- Let's consider a combination of the deep neural network (DNN) acoustic model trained with Kaldi toolkit and the standard n-gram language model, supplied in the distribution of the WSJ speech corpus. Comparative analysis of the performance of our GPU implementation against the open-source reference lets us draw the following conclusions:

- Accuracy of our GPU-enabled engine is approximately equal to that of the reference implementation. There is a small fluctuation of the actual Word Error Rate (WER) due to the differences in arithmetic implementation.
- For the single-channel recognition the TITAN-enabled engine is significantly (~7 times) faster than the reference implementation. This is important in tasks like serving ASR to a Spoken Dialogue System (SDS) or media-mining for specific spoken events.
- Our implementation of speech recognition in the mobile device (NVIDIA Tegra K1) enables twice faster than real-time processing without any degradation of accuracy.
- Our GPU-enabled engine allows unprecedented energy efficiency of speech recognition. The value of 15W per one RT channel for i7-4930K was estimated while the CPU was fully loaded with 12 concurrent recognition jobs. This configuration is the most power efficient manner of CPU utilization. Our TITAN-enabled server does better (~9 W per one RT channel) while maintaining its processing speed. The Tegra-based solution is several times more power efficient (~3.6 W per one RT channel).
- 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.

Verbumware GMM-HMM System vs HTK HDcecode

TASKS\LMS	BCB05ONP	BCB05CNP	TCB20ONP	BCB05ONP	BCB05CNP	TCB20ONP
NOV'92 WER	**9.27%**	**5.47%**	**4.50%**	9.30%	5.59%	5.64%
NOV'92 xRT	**0.0697**	**0.0700**	**0.0756**	3.3023	3.3291	3.9413
NOV'93 WER	28.05%	26.72%	**11.66%**	27.88%	26.46%	13.37%
NOV'93 xRT	**0.0724**	**0.0722**	**0.0775**	4.0167	4.0631	5.2329
	Verbumware GMM-HMM System			**HDecode**		

GMM-HMM HDecode

- A classical Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) system behaves in a similar manner. Comparing our GPU implementation to the HTK tool-kit reference we see, that:
- With the bi-gram LMs accuracy is roughly the same.
- With the tri-gram LM (TCB20ONP) accuracy is significantly better. Apparently, unlike our implementation, the HTK HDecode speech recognition decoder contains an error, that does not allow it to attain the maximum possible accuracy. It is worth noting that with such accuracy level our engine would have won the DARPA challenge but unlike the actual winner, our system in not gender-dependent.
- Our speech recognition system is massively faster (from 50 to 67 times).
- Our recognition speed with the optimal parameter set does not depend much on the total search network size. This fact indicates optimality of the search implementation.

Acknowledgments

We would like to express our gratitude to NVIDIA that supported this work and supplied us with the Jetson TK1 evaluation board within the framework of “The Tegra® K1 CUDA® Vision Challenge”.