Towards High-Quality Speech Recognition on Low-End GPUs

Kshitij Gupta and John D. Owens
University of California, Davis

UC DAVIS
UNIVERSITY OF CALIFORNIA

Abstract

We focus on optimizing compute and memory-bandwidth-intensive GMM computations for low-end, small-form-factor devices running on GPU-like parallel processors. With special emphasis on tackling the memory bandwidth issue that is exacerbated by a lack of CPU-like caches providing temporal locality on GPU-like parallel processors, we propose modifications to three well-known GMM computation reduction techniques. We find considerable locality at the frame, CI-GMM, and mixture layers of GMM compute, and show how it can be extracted by following a chunk-based technique of processing multiple frames for every load of a GMM. On a 1,000-word, command-and-control, continuous-speech task, we are able to achieve compute and memory bandwidth savings of over 60% and 90% respectively, with some degradation in accuracy, when compared to existing GPU-based fast GMM computation techniques.

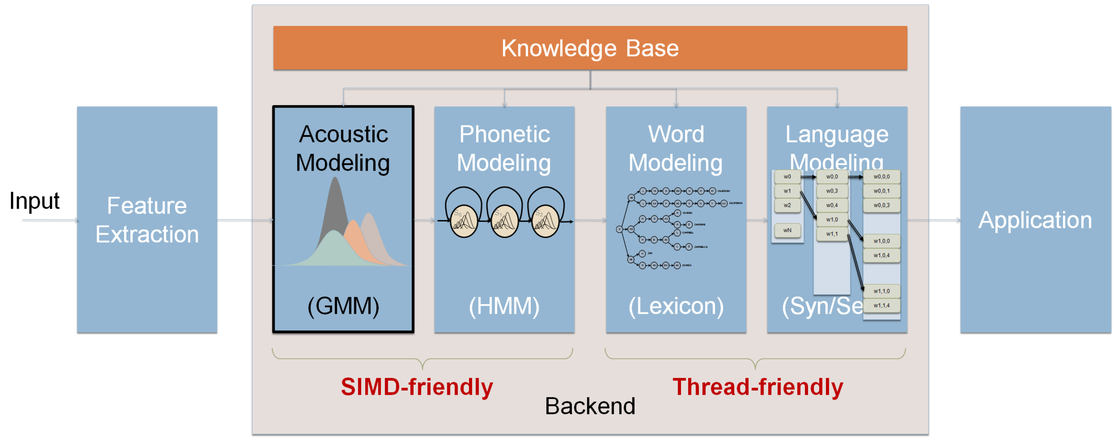
Nature of ASR Algorithms

		Frontend		Backend			
		Feature Extraction		Acoustic Modeling		Language Modeling	
Core kernels		FFT, DCT		GMM computation & HMM state traversal		Layered graph search	
Memory	Footprint	Very small	++	Medium	+	Very large	- -
	Bandwidth	Low	++	Very high	- -	Medium	+
	Access patterns	N/A		Spatial locality	+	Temporal locality	+
Compute		Very low	++	Very High	- -	Low	++
Data-structure		N/A		Dense	+	Sparse	- -
Time	System	< 1%		50-90%		10-50%	

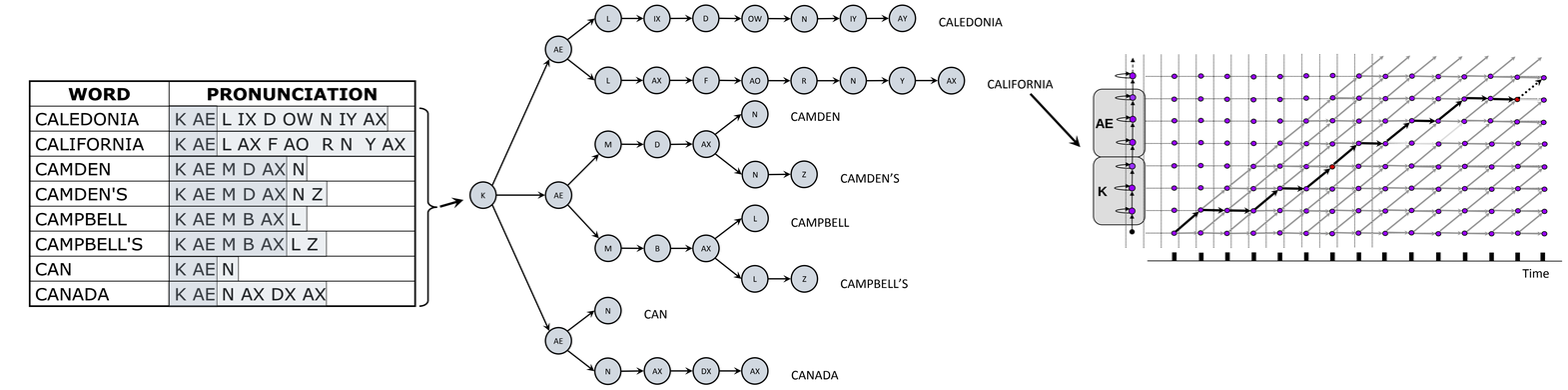
ASR Application Domains

	Server	Desktop	Embedded
	Off-line & On-line	On-line & Off-line	On-line
Real-Time constraint	N/A & Soft	Soft	Hard
Application domain	Transcription	Desktop control	Search
	Data mining	Dictation	Dictation
	Customer support	Game consoles	SMS/Chatting
	Distributed Speech Recognition	Home automation	Command & Control
		Data mining	Automotive

Speech Recognition Overview



WORD	PRONUNCIATION
CALEDONIA	K AE L IX D OW N IY AX
CALIFORNIA	K AE L AX F AO R N Y AX
CAMDEN	K AE M D AX N
CAMDEN'S	K AE M D AX N Z
CAMPBELL	K AE M B AX L
CAMPBELL'S	K AE M B AX L Z
CAN	K AE N
CANADA	K AE N AX DX AX



Results*

AML + CI-GMM

Chunk	CI State Threshold	WER	Compute Saved(%)	BW Saved(%)
1	1	3.09	46.16	46.16
4	3	3.08	60.66	82.27
4	**4**	**3.03**	**67.97**	**90.18**
8	3	3.03	47.59	90.26
8	4	2.97	54.92	91.89

AML + SVQ

Chunk	CI State Threshold	Top Mix.	WER	Compute Saved(%)	BW Sv(%)
4	4	3	4.00	69.11	93.94
4	**4**	**4**	**3.29**	**65.06**	**92.69**
8	4	3	6.21	72.77	95.58
8	4	4	4.40	67.09	94.56

AML + CI-GMM + SVQ

Chunk	Top Mixtures	WER	Compute Saved(%)	BW Saved(%)
4	3	3.57	36.61	85.53
4	**4**	**2.95**	**23.56**	**81.96**
8	3	5.48	39.76	91.50
8	4	3.92	25.50	89.41

* Kshitij Gupta, John D. Owens, "Three-Layer Optimizations for Fast GMM Computations on GPU-like Parallel Processors", in Proceedings of the Eleventh Biannual Speech Recognition and Understanding Workshop, 2009.

Summary

- Traditional fast GMM techniques map well onto GPU-like parallel architectures.
- Significant temporal locality at every stage of GMM compute exists and can be extracted without significant overhead.
- Three layers optimized:
  - Frame layer
  - CI-GMM layer
  - Mixture layer
- Savings obtained:
  - Compute: ~60%
  - Memory bandwidth: ~90%
- These savings are critical for achieving high-quality speech recognition on low-end GPU-like platforms.