Real-time multi-parametric thermal therapy monitoring: GPU versus CPU

Christopher MacLellan[1,2], David Fuentes[1,2], Florian Maier[1], Wolfgang Stefan[1], John D. Hazle[1,2], and R. Jason Stafford[1,2]

[1]*Department of Imaging Physics, University of Texas MD Anderson Cancer Center, Houston, Texas, United States,* [2]*Medical Physics Program, University of Texas Graduate School of Biomedical Sciences at Houston, Houston, Texas, United States*

Target Audience: Researchers interested in real-time multi-parametric monitoring of thermal therapies and the use of parallel computational methods for fast MR processing.

Purpose: Fast chemical-shift imaging (fCSI) using multiple gradient-echo (MGE) acquisitions has previously been developed as a robust method for temperature monitoring by direct measurement of the water proton resonance frequency.[1] Additionally, T2*, T1, and amplitude can be simultaneously estimated using multi-flip angle variations of these sequences which facilitates multi-parametric monitoring of tissue changes during therapy.[2,3] However, the bottleneck to real-time monitoring with this technique is post-processing time. In this work we compare the performance of two autoregressive moving average (ARMA) based algorithms implemented on CPU and GPU architecture on a simulated multiple flip angle multi gradient-echo acquisition (MFA-MGE) to evaluate the feasibility of real time measurement of amplitude, T1,T2*, and resonance frequency for monitoring thermal therapies.

Methods: *In silico* MFA-MGE data was generated for a mixture of methyl and water protons based on the signal equation for a spoiled gradient echo sequence in the steady state:

$$S(\alpha, TE) = \sum_{n=1}^{N} S_{0,n} \frac{\left(1-\exp\left(-\frac{TR}{T1_n}\right)\right)sin(\alpha)}{1-exp\left(-\frac{TR}{T1_n}\right)cos(\alpha)} \exp\left(-\frac{TE}{T2_n^*} - i \cdot \delta f_n \cdot \gamma \cdot B_0 \cdot TE\right) + \varepsilon_r + i\varepsilon_i$$

where the assigned tissue properties and scan parameters are given in tables 1 and 2 and $\varepsilon \sim \mathcal{N}(0, \sigma)$. Parameters were recovered in two stages. First, for each phase, the MGE signal was modeled as a sum of complex exponentials and solved for amplitude, T2*, and resonance frequency using either a Prony algorithm (CPU/GPU) alone and a Steiglitz-McBride (SM) algorithm (5 iterations) using the Prony solution as an initial condition (CPU only). Second, at each phase, the T1 value was recovered using the recovered amplitude values and nominal flip angles with a linear fit technique.[3] Computations were performed in MATLAB (v2013a, Mathworks, Natick, MA) using three different processing methods: serial CPU (Intel Xeon E5640 2.67GHz), parallel CPU (8 x Intel Xeon E5640 2.67GHz), and parallel GPU (NVIDIA GF100). The parallel toolbox was used for multi-threaded CPU calculations. GPU algorithms were implemented as CUDA kernels. As the compute kernels are expected to be memory bound on the GPU architecture, memory management, global memory access, and memory coalescence were carefully considered in the implementation. Each processing method and algorithm combination was benchmarked by increasing the ROI size. The accuracy of each parameter was evaluated for 10^4 samples over a range of SNR values ($SNR \equiv \frac{|S_{0,1}(TE=minTE)|}{\sigma}$) from 1 to 100 using both algorithms.

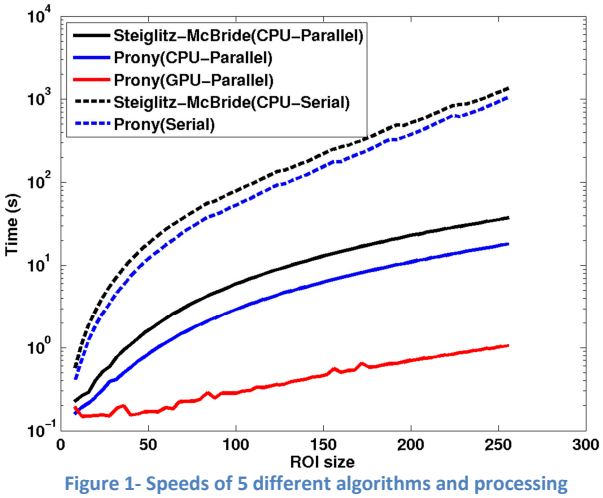

Figure 1- Speeds of 5 different algorithms and processing methods as a function of ROI size.

	Water protons	Methyl protons
S_0	1000	1500
T2*	35	15
T1	500	260
Center Freq. (δf)	0 ppm	3.5 ppm

Table 1- *In Silico* phantom properties.

Matrix size	8x8 – 256x256
Minimum echo time	1.6 ms
Echo spacing	1.6 ms
Number of echoes	16
Flip angles (α)	$20^0/30^0$
Repetition time (TR)	50 ms
Center frequency (γB_0)	128 MHz
Signal to noise ratio	1-100

Table 2- Simulated pulse sequence parameters.

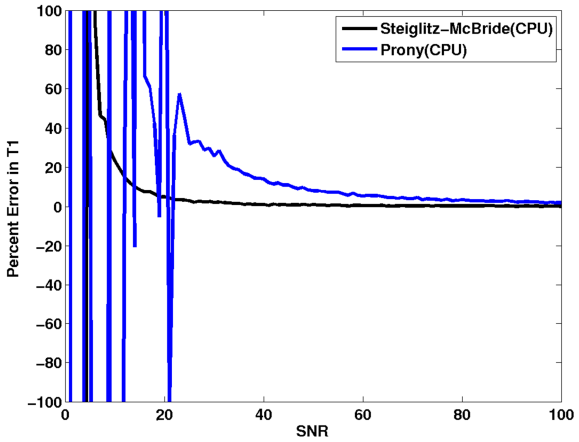

Figure 2- Accuracy of recovered T1 values as a function of SNR (T1=260ms)

Results: The benchmarking results are shown in figure 1 for each algorithm and processing method. Parallel processing on CPU is approximately one order of magnitude faster compared to serial computation for both algorithms. The SM algorithm takes 1.5-2 times longer than the Prony algorithm for both serial and parallel computation on the CPU. The Prony algorithm on GPU is approximately one and two orders of magnitude faster than the parallel and serial Prony algorithms, respectively. The increased accuracy of the SM algorithm was substantial at low SNRs with T1 estimation being the least stable. The accuracy of T1 estimation for methyl protons is shown in figure 2.

Discussion: Benchmarking results show that parallel CPU and GPU computation can reduce the time needed for estimation of multiple parameters by 1-2 orders of magnitude. At ROI sizes currently used for thermal therapy monitoring (<100x100) the Steiglitz-McBride algorithm can be run on 8 CPUs within the time it takes to acquire one image (≈5s). GPU architecture is an order of magnitude faster and could potentially be used to process entire images or permit the use of accelerated acquisitions using techniques such as parallel imaging or compressed sensing. Examination of the accuracy of both algorithms shows that the Steiglitz-McBride algorithm is necessary for real time monitoring at low SNRs and should be implemented on GPU architecture in the future.

Conclusion: Post processing on parallel CPU and GPU architectures makes multi-parametric monitoring of thermal therapies, and other amenable post-processing approaches, in near real-time feasible. While GPU architecture provides an order of magnitude speedup over parallel CPU, the Steiglitz-McBride algorithms should be implemented to maximize the accuracy of the parameter recovery at low SNR.

References: **1.** Taylor et al., *Med Phys.,* 35(2), 2008; **2.** Taylor et al., *NMR Biomed.,* 24(10), 2011; **3.** Todd et al., *Magn. Reson. Med.,* 69(1), 2013