LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-CONF-220882

DSP-Based dual-polarity mass spectrum pattern recognition for bio-detection

Vincent Riot, Keith Coffee, Eric Gard, David Fergenson, Paul Steele

April 26, 2006

Fourth IEEE Workshop on sensor array and multi-channel processing
Boston, MA, United States
July 12, 2006 through July 14, 2006

DSP-Based dual-polarity mass spectrum pattern recognition for bio-detection

Vincent Riot, Keith Coffee, Eric Gard, David Fergenson, Shubha Ramani, Paul Steele
Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94550, USA

The Bio-Aerosol Mass Spectrometry (BAMS) instrument analyzes single aerosol particles using a dual-polarity time-of-flight mass spectrometer recording simultaneously spectra of thirty to a hundred thousand points on each polarity. We describe here a real-time pattern recognition algorithm developed at Lawrence Livermore National Laboratory that has been implemented on a nine Digital Signal Processor (DSP) system from Signatec Incorporated. The algorithm first pre-processes independently the raw time-of-flight data through an adaptive baseline removal routine. The next step consists of a polarity dependent calibration to a mass-to-charge representation, reducing the data to about five hundred to a thousand channels per polarity. The last step is the identification step using a pattern recognition algorithm based on a library of known particle signatures including threat agents and background particles. The identification step includes integrating the two polarities for a final identification determination using a score-based rule tree. This algorithm, operating on multiple channels per-polarity and multiple polarities, is well suited for parallel real-time processing. It has been implemented on the PMP8A from Signatec Incorporated, which is a computer based board that can interface directly to the two one-Giga-Sample digitizers (PDA1000 from Signatec Incorporated) used to record the two polarities of time-of-flight data. By using optimized data separation, pipelining, and parallel processing across the nine DSPs it is possible to achieve a processing speed of up to a thousand particles per seconds, while maintaining the recognition rate observed on a non-real time implementation. This embedded system has allowed the BAMS technology to improve its throughput and therefore its sensitivity while maintaining a large dynamic range (number of channels and two polarities) thus maintaining the systems specificity for bio-detection.

I. Introduction

Biological warfare agents that can be released as aerosols are of major concern because they can be disseminated easily and quickly over wide areas in lethal doses [1]. The detection and identification of airborne biological particles in real-time is a capability required for an effective response to terrorist threats as well as for early medical diagnosis. The Bio-Aerosol Mass Spectrometry (BAMS) system as described in Fig. 1, makes use of a high efficiency inlet used to sample aerosol particles, a tracking and sizing region that predicts the location and speed of the particles inside the system, a 2-band fluorescence pre-selection stage and a final dual polarity (negative and positive) mass-spectrometer for single particle chemical analysis. The dual-mass spectrum recorded for each individual particle can be used as a signature. However, the data obtained from the system is simply the time-of flight arrival of various ions generated from the current particle analyzed and needs to be quickly processed in order to be associated with a given organism or toxin. The time-of flight is being recorded by one 500MHz digitizer (Model PDA1000 from Signatec Incorporated) per polarity for 16us to 32us, yielding a total of 16,000 to 32,000 samples per particle analyzed. In order to increase the sensitivity level of the BAMS instrument, we needed to increase the number of particles per second that could be processed while maintaining the specificity of the existing recognition algorithm [2].

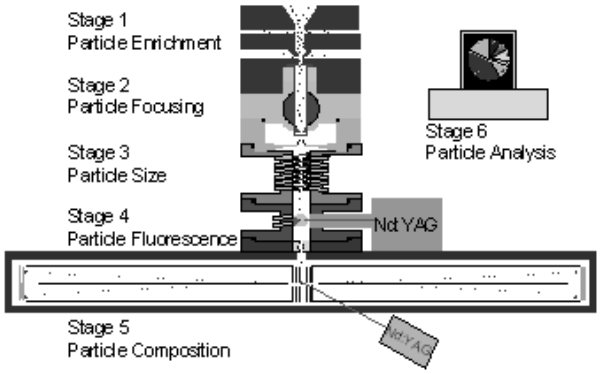

Fig. 1. Cross-sectional cartoon of the current BAMS system

We describe in this paper the implementation of the real-time algorithm processing the dual-mass spectrometer data using a combination of 9 Digital Signal Processors (DSP) (Model PMP8A from Signatec Incorporated). We will first describe the various steps involved in the identification and then focus on the parallel implementation of the algorithm.

II. Algorithm description

The processing algorithm can be broken into three major steps. The first step is a conditioning step used to remove any baseline offset caused by detector variations and digitizer settings. The baseline is estimated as the median value of the raw data on a given time-of-flight interval. In order to accommodate time varying baseline fluctuations, the mass vector is split into 3 chunks for each polarity and the baseline is estimated and removed on the corresponding time-period. The second step consists of pre-processing the conditioned time-of-flight data into a calibrated mass-to-charge spectrum representing the chemical composition of the particle being

analyzed. Using a calibration compound, it is possible to associate every time period of the time-of-flight data with an equivalent mass-to-charge ratio value. As the digitization speed is quite large, more than one time-of-flight sample point will contribute to a single mass-to-charge bin.

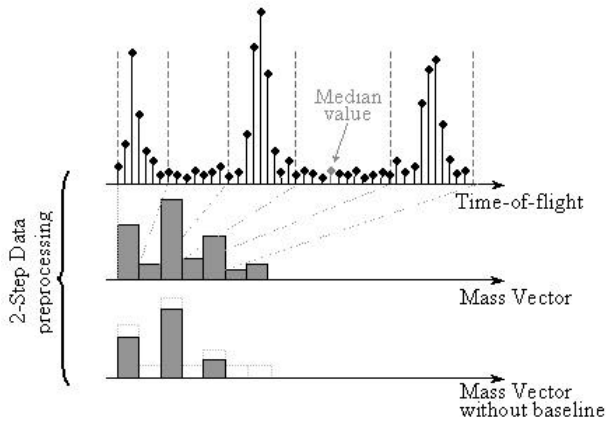

Fig. 2. Mass Spectrum time-of-flight pre-processing

The algorithm simply sums all the time-of-flight sample points corresponding to a unique mass-to-charge ratio value, thus creating what is called a mass vector. A mass vector typically has 500 to a 1000 mass-to-charge peaks per polarity. While the baseline is estimated first, it is actually removed after the calibration is performed, in order to reduce the computational complexity by minimizing the number of subtractions required (see Fig. 2).

After the two-step pre-processing, the data has been reduced to two 500 to 1000-dimensional vectors, one for each polarity of the mass spectrometer. The next and third step in the algorithm consists of associating an organism or toxin signature to the data. It has been demonstrated that individual particle signatures from time-of-flight mass spectrometry can be clustered [3] using modified Adaptive Resonance Theory techniques (ART) [4]. Using known agents, a library of typical patterns derived from the clustering is developed offline for each polarity.

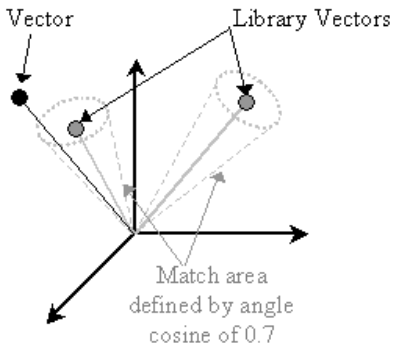

Fig. 3. Classification against a library using multi-dimensional angles

The classification step consists then of finding which pattern in the library is the closest to the multi-dimensional vector being analyzed. This step is performed by estimating the multi-dimensional angle between the vector being analyzed and the vectors in the library. This angle can be easily derived by computing the dot product between the two multi-dimensional vectors, which gives a value of the cosine of this exact angle once normalized by the Euclidian norm of each vector. A cosine value close to 1 indicates a small angle and therefore a very good match (See Fig. 3). The threshold for a possible match is set at a cosine value of 0.7 or above. The possible matches are then sorted according to their cosine value for each polarity. The final step of the classification consists of combining the two polarity results. This is done by implementing a simple score based rule tree summarized as follows:

1. Take the highest match from the sorted list from the negative polarity.
2. Starting with the highest match score from the sorted list from the positive polarity, check if the match equal to the current positive index.
    a. If YES, return the match as the final answer
    b. If NO, go to step 2, using the next highest match score from the sorted list from the positive polarity.
3. If No match has been found, go to step 2 using the next highest negative match score from the negative polarity. If the list has been exhausted, then the final answer is set to unknown.

Because the DSP implementation is limited by memory space the algorithm uses the rules on only the first three best matches of each polarity.

III. FIXED-POINT COMPUTATION

The DSPs used in the implementation of the algorithm have fixed point arithmetic and have limitations in terms of computations. Multiplications operate on two 16 bits (or less) numbers and produce a 32 bit number while additions operate on two 32 bits numbers and produce a 32-bit number. An overflow check bit is available in case the result would be more than 32-bits wide. In order to accommodate for fixed point operations, several precision measures have been taken in order to ensure appropriate results.

Each time-of-flight data point is represented with 8 bits coming from the digitizer. The various calibrations obtained for conversion to mass-to-charge ratio tend to show that a single mass-to-charge bin will not exceed more than 512 time-of-flight sample points. This yields values for the mass-to-charge ratios that can be represented with 17 bits. In order to compute the final vector Euclidian norm we then have to truncate those values by 5 bits (equivalent to dividing by 32). This reduces the final precision but is requiring only 12 bits for each individual mass-to-charge ratio. The Euclidian norm for a 500 length mass vector would then require 33 bits and therefore could overflow. However, in practice no mass vector will have all its values set to the maximum height and if it did, the data would be considered corrupted, so the overflow process will serve to flag unrealistic signatures.

The mass vector in the signature library is set so that each mass-to-charge ratio is represented with 8 bits and that the Euclidian norm is set to 256. This is only possible if no vector needs to have a single non-zero mass-to-charge value. In practice we have never encountered such a case. The dot product of a library mass vector with the current particle mass vector will yield a number that can use up to 29 bits. The dot product result can be shifted by 8 bit as the norm of the library vector is 256. We now have a 21 bit number representing the Euclidian norm of the current mass vector multiplied by the cosine of the multi-dimension angle between the two mass vectors. Because we do not have easy access to the square root function in a DSP, we will estimate the square of the cosine instead. As the square function is an increasing monotonous bijection we can then have equivalence if we use the 0.7 threshold squared instead. We therefore square the dot product number shifted by 8 bits to the right. Even though, it seems that we could overflow (21bits squared could yield up to a 42bits numbers), we know it cannot be greater than the norm of the current vector (33bits and only not corrupted if within 32 bits) as the cosine is less than one. Dividing this number with the previously 32 bit norm value would only give a 0 or 1 result as they are all represented in the same fixed point representation. In order to go around this issue, the norm is modified to loose 8 bits precision by being shifted right by 8 bits before being divided to the squared normalized dot product. The result is then an approximation of the cosine squared with 8 bit precision where 256 would be one. The next step is then to compare the result to 0.7 squared or 0.49, which corresponds to a value of 125 in 8 bit fixed point representation. The sorting of the different matches can be done on the squared cosine values the same as it would be done with the simple cosine values.

IV. Parallel implementation

In order to achieve maximum performance in terms of speed, the algorithm described has been implemented on a highly parallel platform containing 9 DSPs, each of them having eight arithmetic and logic units (TMS32C6201 from Texas Instrument). The PMP8A from Signatec Incorporated contains two independent groups of four processing DSPs and a ninth DSP acting as a master, responsible for initiating data transfer between the various interfaces and the processing DSPs. In addition, the board contains one large memory per group of four DSPs.

The two polarity time-of-flight spectra are processed independently until the final step. In addition, the digitized data is generated in two different digitizers. In order to parallelize the pre-processing, each polarity time-of-flight data is therefore sent to each group of four processing DSPs as shown in Fig. 4. For each polarity, the adaptive baseline removal can be split into three independent processes that can be run in parallel. The data is therefore transferred by contiguous blocks into three of the processing DSPs in a group, where the two pre-processing steps occur. The mass vector pieces are then reassembled in the fourth DSP of the group for the final dot product matching step. In order to improve the overall speed of the pre-processing, it is possible to optimize the size of the various blocks sent to the first three DSPs of the group. The first DSP will obviously have more processing time than the second DSP itself having more processing time than the third, since the memory (RAM) can only be accessed by one DSP at a time. The following equation can be solved for optimum processing speed:

N_1 : number of samples transferred on DSP1

N_2 : number of samples transferred on DSP2

N_3 : number of samples transferred on DSP3

N : number of samples in time - of - flight data

P : processing rate (sample per second)

T : transfer rate (sample per second)

$$\begin{cases} N_1 + N_2 + N_3 = N \\ \frac{N_1}{P} = \frac{N_2}{T} + \frac{N_2}{P} \\ \frac{N_2}{P} = \frac{N_3}{T} + \frac{N_3}{P} \end{cases} \quad (1)$$

In this application, the transfer rate is two hundred times larger than the processing rate which makes the data length different from each other by about 100 samples.

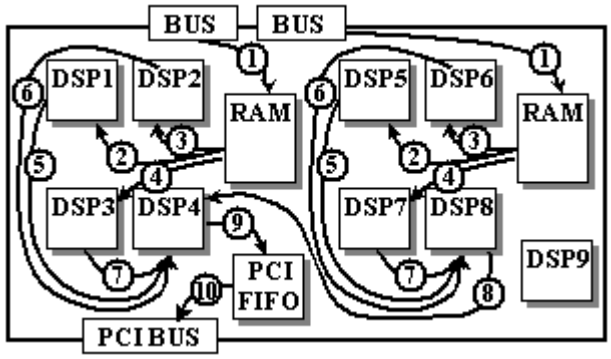

Fig. 4. Parallel processing data flow

The final score based rule combining step is done in the fourth DSP of the first group. It receives the sorted data from the other polarity and performs the rule tree algorithm on the three best matches of each polarity. The result is then sent to the computer through a PCI interface. For diagnostics and further offline analysis, the data stored in the RAM blocks can also be exported through the PCI interface, but it obviously reduces the overall speed performances.

The embedded processing has been successfully implemented on the PMP8A DSP board and run at 30 particles per seconds with better than 80 percent recognition rate for most agents, matching the offline equivalent recognition algorithm performance in specificity but exceeding the processing by a factor of 5. Theoretically the system could

process up to 1000 particles per seconds but currently we do not have an ionization laser able to operate at this speed.

IV. Conclusion

We have demonstrated speed improvement in multi-dimensional data processing for bio-security application using highly parallel implementation of an existing algorithm. Digital signal processors have allowed us to move the speed bottleneck caused by long processing time to instrumental limitations such as laser repetition rate. Furthermore as technology in those areas improves, we will be able to maintain higher signature analysis rates. We have also demonstrated the ability to use multi-channel signatures in order to reach high specificity levels in bio-detection in a real-time and standalone system.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48 – UCRL-CONF-220882. This project was also funded in part by the Defense Advanced Research Projects Agency (DARPA).

References

[1] Walt, D. R. & Franz, D. R. “Biological Warfare Detection”, *Analytical Chemistry* 72, 730 A-746 A (2001).

[2] David P. Fergenson et al., “Reagentless Detection and Classification of Individual Bioaerosol Particles in Seconds”, *Analytical Chemistry*., 76 (2), 373 -378, 2004

[3] Xin-Hua Song, Philip K. Hopke, David P. Fergenson and Kimberly A. Prather , “Classification of Single Particles Analyzed by ATOFMS Using an Artificial Neural Network, ART-2A”, *Analytical Chemistry*, 71,860-865, 1999

[4] Carpenter, G.A., Grossberg, S., Rosen, D.B., "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, 4, 493-504, 1991