Blind Source Separation of Speech in Hardware

Niall Hurley*,Naomi Harte†,Conor Fearon‡,Scott Rickard§
*‡§Dept. of Electrical and Electronic Engineering
University College Dublin, Dublin, Ireland
* Email: niall.hurley@ee.ucd.ie
‡Email: conor.fearon@ee.ucd.ie
§Email: scott.rickard@ucd.ie
†Dept. of Electrical and Computer Engineering, McMaster University, Ontario, Canada

***Abstract*— This paper presents preliminary work on a hardware implementation of a source separation algorithm employing time-frequency masking methods. DUET (Degenerate Unmixing Estimation Technique) has previously been shown to achieve excellent source separation in real time in software. The current work is a move towards a hardware realization of DUET that will allow integration of the algorithm into consumer devices. Initial stages involve investigating the performance of DUET when implemented in fixed-point arithmetic and a consideration of algorithmic changes to make DUET more amenable to implementation on a DSP processor. Performance is compared for floating-point and fixed-point implementations. A Weighted K-means clustering algorithm is presented as an alternative to gradient descent methods for peak tracking and demonstrated to achieve excellent performance without adversely affecting computational load. Preliminary performance figures are given for an implementation on a TMS320VC5510 DSK.**

I. Introduction

A. Problem Definition

DUET is a blind source separation technique capable of the separation of N sources from 2 mixtures. The N sources can be defined as $s_1(t)$, $s_2(t)...s_N(t)$. Let $x_1(t)$ and $x_2(t)$ be the mixtures derived from these sources as:

$$x_1(t) = \sum_{j=1}^{N} s_j(t), \tag{1}$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t-\delta_j), \tag{2}$$

where δ_j is the arrival delay between sensors resulting from the angle of arrival, and a_j is a relative attenuation factor corresponding to the ratio of the attenuations of the paths between sources and sensors. Yilmaz and Rickard [3] have previously presented the theory in detail. However, the result of interest in this paper is the fact that by presuming anechoic conditions and that the source signals are approximately w-disjoint orthogonal, only one source is active at any time-frequency point. By using parameter estimation techniques to estimate delay and attenuation values, it is then possible to construct a time-frequency mask M_j that will isolate source j from the mixtures, by locating corresponding peaks in a two-dimensional histogram.

B. Gradient Descent Search

As outlined by Rickard et al. [2], in order to achieve real-time operation of DUET, a gradient search technique is used for mixing parameter estimation over time. Given initial estimates of the delay and attenuation parameters a cost function $J(\tau)$ can be derived as

$$J(\tau) = \min_{a_1,\delta_1,\ldots,a_N,\delta_N} \sum_{\omega} -\frac{1}{\lambda} \ln(e^{-\lambda\rho_1} + \cdots + e^{-\lambda\rho_N}), \tag{3}$$

where

$$\rho(a_j,\delta_j,\omega,\tau) \doteq \frac{1}{1+a_j^2} |X_1(\omega,\tau)a_j e^{-i\omega\delta_j} - X_2(\omega,\tau)|^2. \tag{4}$$

Given that the number of sources being searched for is known, it is possible to derive updates for the amplitude and delay values $(a_j[k], \delta_j[k])$ from the current frame $\tau_k = k\tau_\Delta$ as:

$$a_j[k] = a_j[k-1] - \beta\alpha_j[k]\frac{\partial J(\tau_k)}{\partial a_j}, \tag{5}$$

$$\delta_j[k] = \delta_j[k-1] - \beta\alpha_j[k]\frac{\partial J(\tau_k)}{\partial \delta_j}, \tag{6}$$

where β is a learning rate constant and $\alpha_j[k]$ is a time and mixing parameter dependent learning rate for time index k for estimate j.

C. K-Means Clustering

Initial evaluation of the computational load in DUET revealed that the main part of the effort lay in the evaluation of gradients used in the parameter updates as described in equations [5,6]. Furthermore, as the results section will demonstrate, problems were encountered with the performance of the fixed-point gradient descent. Hence, it was considered worthwhile to investigate alternative methods of tracking the peak values.

The K-means algorithm is a classic technique employed in data clustering problems [6]. The algorithm efficiently partitions the points of a data matrix into K clusters. It achieves this by minimizing (in a least mean squares sense) the sum of distances from each point to its nearest cluster center. Each iteration starts by reassigning points to their nearest cluster center. Each cluster center is then recalculated as the mean

0-7803-9333-3/05/$20.00 ©2005 IEEE

of all points which have been assigned to it. This process is repeated until the cluster centers converge according to the chosen criterion.

The use of a histogram space has proved to be very powerful in DUET [1]. Rather than searching for peaks in the entire amplitude-delay space, the data is placed into a bounded histogram with a finite number of bins. The limitation of this method for real-time operation is the use of a two-pass approach. In the current work, K-means clustering is used to allow peak tracking in histogram space in real-time. A weighted version of the K-means algorithm is performed on the histogram. For each frame of data, the (weighted) histogram is updated with the powers of the corresponding time-frequency points. The histogram bin centers are passed to K-means and each point is weighted by the height of that histogram bin. In this way, peak-tracking updates are calculated using information from all previous frames. As the results section demonstrates, this method yields very accurate peak estimates.

II. REAL-TIME OPERATION

A. System Overview

The diagram in Figure 1 shows a block diagram of the system used to implement DUET in real time.

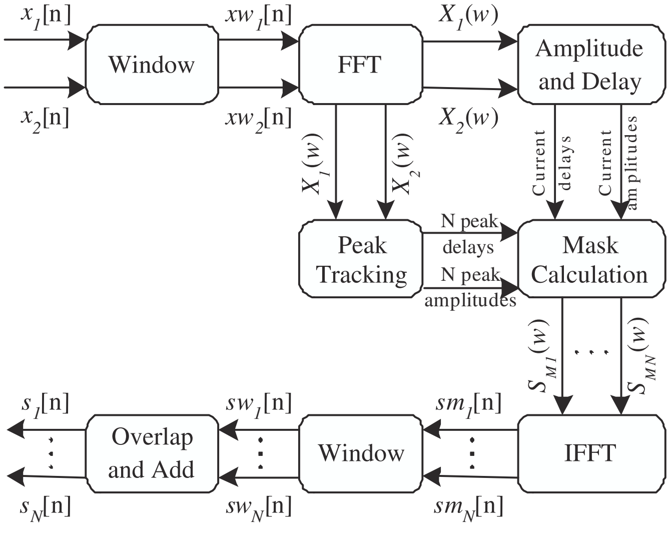


Fig. 1. DUET System - Block Diagram

With data input at 8KHz, data is processed in frames of 512 samples with a 75% overlap of successive frames. After windowing with a hamming window, the frame is transformed to the frequency domain via a 512-point FFT. Instantaneous delay and amplitude values for this frame are calculated for each time-frequency point as outlined by Rickard et al.[3]. Updated parameter estimates for the actual amplitude and delay values are obtained using the gradient update as explained in equations [5, 6] above.

The masking operation involves calculating, for each time-frequency point, which of the N peak amplitude and delay values each point is closest too. This is done using a simple Euclidean distance measure. In this way each time-frequency point is only assigned to a single source. This “winner take all” scenario greatly reduces computational complexity and, as reported previously in [2], has little perceivable impact on performance. The N masks are used to derive a time frequency representation of each of the sources as

$$S_{Mj}(\omega,\tau) = M_j(\omega,\tau)X_1(\omega,\tau). \tag{7}$$

An inverse FFT on each of the N masked signals, followed by windowing and overlap-and-add yields a new frame of each of the N source signals.

B. Fixed-Point Migration

The aim of this work is to demonstrate the feasibility of incorporating the DUET algorithm into high-end consumer devices. Considering costs, it was hence appropriate that the algorithm be targeted at a fixed-point rather than floating-point processor. The TI C5510 family was chosen as the target processor and the TMS320VC5510 DSK represented a low-cost development platform for developing the algorithm. This device is a 16-bit processor, operating at 200MHz and capable of delivering up to 400 MIPs. The chip has 160K 16-Bit On-Chip RAM and a dual MAC. Full details of the chip and DSK are available online at the TI website [4] and Spectrum Digital homepage [5].

A fixed-point implementation equivalent to the floating-point system was carried out in C, as an intermediate step to allow full system testing of the fixed-point migration. Within this system a number of simplifications were made to speed up development. Any functional blocks such as trigonometric functions, FFT, log were not written as full fixed-point libraries as these would be integrated when targeting the board. For system evaluation, inputs and outputs of these functions were appropriately quantized. This gives a slightly more advantageous performance than the final system but was considered appropriate for development purposes. All other values were stored as 16-bits. For the port to the TMS320VC5510 DSK, the free signal processing libraries supplied by TI were used.

III. EXPERIMENTS

This section outlines a number of experiments designed to test the fixed-point performance of DUET. Performance is compared to the original floating-point algorithm. Results on the use of the Weighted K-means algorithm for peak tracking are also given.

A. Fixed-Point Performance of Gradient Descent

The difference in peak estimates for amplitude and delay were compared for the floating-point and fixed-point implementations of DUET. The table below details the average percentage error in the Gradient Descent fixed-point estimate, referenced to the floating-point value obtained for a mixture of two sources.

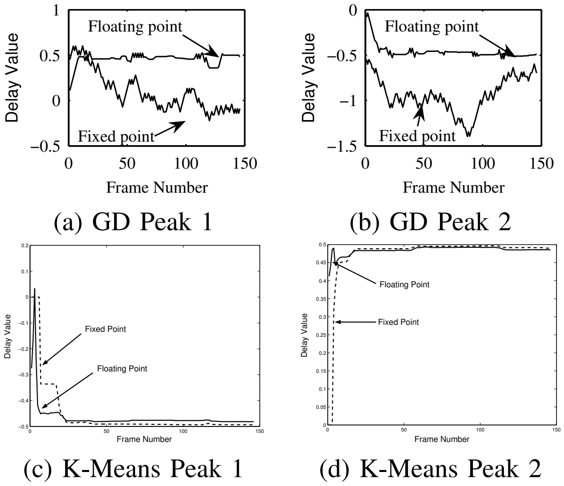
A significant error has been introduced, particularly in the delay estimate. Examining the evolution of the delay estimate

Peak	% Error
Amplitude	11.15
Delay	94.62

TABLE I

Percentage Error in Delay and Amplitude Estimates for Gradient Descent Fixed-Point System

over time in Figure 2 (a) and (b), it is clear that the delay is not converging to the true value. Closer investigation has shown that this error is largely attributable to underflow in the derivative values for the gradient update. This arises due to the large number of successive multiplies used. Even with appropriate scaling, a significant number of derivative values simply tend to zero. Clearly this method of peak tracking is problematic using fixed-point arithmetic. The use of weighted K-means alleviates this issue as is evident from Figure 2 (c) and (d).



(a) GD Peak 1 (b) GD Peak 2

(c) K-Means Peak 1 (d) K-Means Peak 2

Fig. 2. Emerging Peak Estimates for 2 Sources.

An initial port to the TMS320VC5510 DSK has yielded an upper estimate of 18 MIPS for the DUET functionality. It should be noted that this is an unoptimized port which incorporates the gradient descent peak tracking (including estimates of divide functionality) and current indications suggest a final figure of 5 MIPs as highly achievable. Work is ongoing on optimizing the performance of the algorithm in hardware. This involves migrating to proprietary libraries for trigonometric functions, the FFT and incorporating the weighted K-means method of peak tracking.

B. Performance of Weighted K-Means

As an indicator of the comparative complexity of each gradient descent and Weighted K-means for peak tracking, the number of adds and multiplies for a speech file of length N samples is shown in Table [II] for both algorithms.

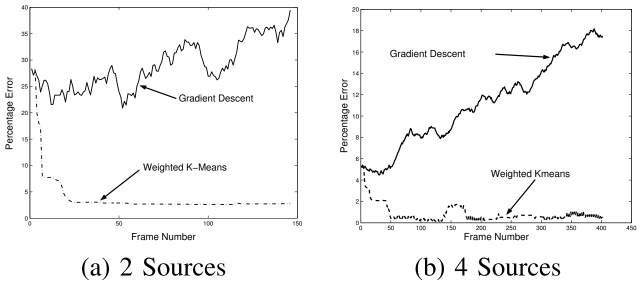
	ADDS	MULTIPLIES
Weighted K-Means (128^2 bins)	$6671N$	$4533N$
Weighted K-Means (64^2 bins)	$1555N$	$1062N$
Weighted K-Means (32^2 bins)	$318N$	$227N$
Weighted K-Means (16^2 bins)	$69N$	$53N$
Gradient Descent	$44N$	$165N$

TABLE II

No. of adds and multiplies for Weighted K-means and Gradient Descent algorithms.

Clearly, the computational complexity of Weighted K-means is highly dependent on the number of bins used in the histogram space. Reducing the number of histogram bins by a power of two reduces the number of adds and multiplies by the same factor. For the separation of only two sources, it has been found that smaller histogram spaces of 16×16 bins still yield good performance. However, a histogram space of 128×128 would be required for the case of more than 4 sources. An important advantage of Weighted K-means is that it is possible to completely eliminate the need for any divides. This is not possible with the current formulation of the gradient descent estimate updates.

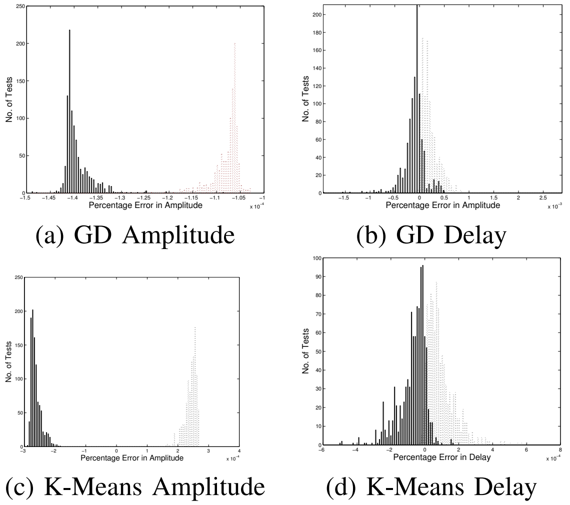
The error in peak estimates for a two-source and four-source mixture at each frame is shown in Figure 3. The Weighted K-means clearly outperforms Gradient Descent in terms of accuracy for both a two-source mixture and a four-source mixture.



(a) 2 Sources (b) 4 Sources

Fig. 3. Total percentage error plots for Weighted K-means and Gradient Descent.

Further tests, using one thousand mixtures of the TIMIT database, demonstrates the accuracy of Weighted K-Means as compared to Gradient Descent Table [III]. The performance of Gradient Descent drastically degrades when applied to fixed-point arithmetic, as underflow issues dominate to such a degree as to make comparisons meaningless. As can be seen in Table [IV], the error performance achieved by Weighted K-Means is far better than that of Gradient Descent, indicated by the lower mean error and smaller variance value. Added to this, Figure 4 shows histograms of the percentage errors in terms of amplitude and delay. From these it is obvious that the algorithms optimise accuracy by minimising delay error, implying that a much higher resolution is required in the amplitude direction. The delay error histograms Figure 4(b) and (d) could possibly be improved by using a higher-order all-pass filter to obtain a fractional delay.



(a) GD Amplitude (b) GD Delay

(c) K-Means Amplitude (d) K-Means Delay

Fig. 4. Percentage Error in Amplitude and Delay for the two algorithms obtained from one thousand tests.

	Peak 1	Peak 2
Weighted K-Means (Floating point)	0.8625%	0.8650%
Gradient Descent (Floating point)	1.5650%	1.7700%
Weighted K-Means (Fixed point)	2.1625%	2.5075%
Gradient Descent (Fixed point)	43.2425%	66.835%

TABLE III

THE WEIGHTED K-MEANS AND GRADIENT DESCENT ALGORITHMS PERCENTAGE PEAK ERRORS.

IV. CONCLUSION

This paper has presented initial work on the migration of the DUET algorithm to a fixed-point implementation in hardware. Significant problems were encountered in migrating to a fixed-point implementation of DUET incorporating Gradient Descent peak tracking. Weighted K-means clustering is shown to outperform Gradient Descent for amplitude and delay peak tracking without significant adverse effects on computational load. Work is ongoing on the fixed-point implementation to integrate Weighted K-means clustering and to optimize performance on the DSP. Other algorithmic enhancements currently being considered include the exploitation of the properties of speech and improvements in performance in echoic conditions.

	Mean		Variance	
	Peak1	Peak2	Peak1	Peak2
Weighted K-Means (Amplitude)	.0859	.0790	$.3067\times10^{-4}$	$.3995\times10^{-4}$
Weighted K-Means (Delay)	-.0226	.0301	$.7175\times10^{-3}$	$.7441\times10^{-3}$
Gradient Descent (Amplitude)	-1.1415	-.8837	$.5627\times10^{-3}$	$.5049\times10^{-3}$
Gradient Descent (Delay)	-.0318	.0658	.0049	.0040

TABLE IV

MEAN AND VARIANCE OF PEAK ESTIMATES FOR 1000 TESTS (FLOATING POINT).

ACKNOWLEDGMENT

The authors would like to thank...
Work funded by the Enterprise Ireland Proof of Concept fund: Contract PC/2004/347.

REFERENCES

[1] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures”, IEEE conference on Acoustics, Speech, and Signal Processing (ICASSP2000), Vol 5, pp 2985–2988, Istanbul, Turkey, June 2000.

[2] S. Rickard, R. Balan, and J. Rosca, “Real-Time Time-Frequency Based Blind Source Separation”, 3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA2001), San Diego, CA, December 9-12, 2001

[3] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking”, IEEE Transactions on Signal Processing, Vol. 52(7), pp 1830–1847, July 2004.

[4] “TMS320C5000(tm) Platform Overview”, http://dspvillage.ti.com/

[5] “DSP Starter Kit for the TMS320VC5510”, http://www.spectrumdigital.com/

[6] G.A.F. Seber, “Multivariate Observations”, Wiley, New York, 1984.