

# A Methodology for Evaluating Detection Performance of Ultrasonic Array Imaging Algorithms for Coarse Grained Materials

A. Van Pamel, C. R. Brett, and M. J. S. Lowe.

*Abstract*— Improving the ultrasound inspection capability for coarse grain metals remains of longstanding interest and is expected to become increasingly important for next generation electricity power plants. Conventional ultrasonic A, B, and C-scans have been found to suffer from strong background noise due to grain scattering which can severely limit the detection of defects. However, in recent years, array probes and Full Matrix Capture (FMC) imaging algorithms have unlocked exciting possibilities for improvements. In order to progress and compare these algorithms we must rely on robust methodologies to quantify their performance. This article proposes such a methodology to evaluate the detection performance of imaging algorithms. For illustration, the methodology is applied to some example data using three FMC imaging algorithms; Total Focusing Method (TFM), Phase Coherent Imaging (PCI), and Decomposition of the Time Reversal Operator with Multiple Scattering Filter (DORT MSF). However it is important to note that this is solely to illustrate the methodology; this article does not attempt the broader investigation of different cases that would be needed to compare the performance of these algorithms in general. The methodology considers the statistics of detection, presenting the detection performance as Probability of Detection (POD) and probability of False Alarm (PFA). A test sample of coarse grained INCONEL 625, manufactured to represent materials used for future power plant components and containing some simple artificial defects, is used to illustrate the method on the candidate algorithms. The data is captured in pulse-echo mode using 64 element array probes at centre-frequencies of 1MHz and 5MHz. In this particular case, it turns out that all three algorithms are shown to perform very similarly when comparing their flaw detection capabilities.

## I. INTRODUCTION

Coarse granular microstructure has long hindered ultrasonic Non-Destructive Evaluation (NDE) [1-3]. When the wavelength of the probing wave becomes dimensionally similar to that of the grains, increased attenuation [4], incoherent noise [5], and possibly anisotropic effects [6] all serve to obscure targets of interest. Yet these materials are commonplace in a variety of applications, including the electricity generation industry where their future uptake is likely to grow. Namely, improved creep strength derived from the large microstructure would permit next generation power plants to operate at increased temperatures enabling higher efficiencies [7]. The success of this evolution however will depend on technical confidence, including the capability to perform reliable NDE.

In recent years, ultrasonic arrays have been widely implemented across industry favoured for their speed of inspection and ease of interpretation over conventional NDE [8,9]. The ability to beam steer by phase-delaying elements means that one array can reproduce an inspection which would otherwise require multiple monolithic transducers. Once ensuing advances in instrumentation facilitated the acquisition of all send-receive time trace

combinations available to the array (termed Full Matrix Capture (FMC) [10]), the additional information captured in this way enabled arrays to surpass their monolithic counterparts by creating synthetically focused images [11].

Since its onset, numerous algorithms are available to post-process FMC data into an image, including wave-field extrapolation (IWEX) [12], the wavenumber algorithm [13], and what is regarded in the NDE community as the most established algorithm: the Total Focusing Method (TFM) [8,10]. Advanced imaging algorithms are currently emerging in attempts to further improve inspection of difficult materials such as coarse grained metals. Two approaches which hold promise are algorithms to recover performance losses due to anisotropy e.g. auto-focusing [14] and deviation correction [15], and de-noising algorithms to reduce image speckle caused by structural noise from grain scattering. Examples of the latter category include; Phase Coherent Imaging (PCI) [16,17], Decomposition for the Time Reversal Operator (DORT) [18] with a Multiple Scattering Filter [19-21] (MSF), Spatial Coherent Imaging [22] (SCI), and Spatially Averaged Sub-Array Correlation Imaging [23] (SASACI).

In light of the growing interest by industry, the ability to distinguish the performance of these algorithms is becoming increasingly relevant and necessary for progress. Answering this question however first requires a widely accepted definition of performance and a robust method of measuring it. Various methodologies have previously been adopted to compare imaging algorithms [17,21-24], where Signal-to-Noise Ratio (SNR) enhancement is commonly used to convey relative merits. As there exists no widely accepted method, to our knowledge, there remains an opportunity to define an objective method and find a basis for future comparisons.

This article proposes such a methodology to evaluate, from an NDE perspective, the detection performance of different imaging algorithms. Here, Section 2 discusses the importance to consider both detection and false alarm rates whereby Receiver Operating Characteristic (ROC) [25] analysis is introduced as a tool to quantify detection performance. By way of example, three candidate algorithms (introduced in Section 3); TFM, PCI, and DORT MSF, are tested on a coarse grained power plant material in Section 6. Prior to the experimental results, the proposed approach is proven using simulated data (see Section 5) where complete knowledge exists of the inspected medium. Details on data collection for each method, experiment and simulation, are presented in Section 4.

## II. IMAGING PERFORMANCE DEFINITION

NDE usually involves two aspects: flaw detection and flaw characterization; both involve some uncertainty and the latter is here assumed to also include sizing. Whilst imaging performance for NDE applications can be defined by these two indicators, the challenge lies in how to quantify them. Various methodologies have been adopted to compare imaging algorithms [17, 21-24]; an alternative method is reasoned below by considering detection and characterisation.

### A. Flaw Detection

Detection is purely related to signal and noise amplitudes, commonly packaged as Signal-to-Noise Ratio (SNR). Measuring SNR in cases where defects are not detected is difficult when noise dominates signal. In a research context however when the performance of algorithms is being investigated, knowledge of the defects, whether they are simulated or machined, often exists. This additional information available as a result of relying on artificial defects can be exploited to enable comparisons. Thus as a simplification, and applying this to the example of an ultrasound image, it is commonplace [11,21,24] to assume that the pixel intensities corresponding to the predicted flaw-response location contribute as signal (although it contains signal plus noise) whilst all remaining pixels are classified as noise. The appropriate area can be determined by predicting the flaw response through simulation or by observing a well detected case.

Whilst SNR is also a useful indicator of imaging performance, it is valuable to note that there is no benefit in maximising it once detection is achievable. This is illustrated in Figure 1 where all algorithms image the flaw, and therefore achieve equal detection performance, albeit at different SNRs and perhaps characterising ability. As a logical example, once detection is achieved it is trivial to increase SNR, as merely squaring the pixel values in an image will double its SNR [8]; the difficulty lies in actually improving detection rates.

Once signal and noise values are calculated, determining whether a target is detected can be achieved by setting a threshold. Using a selected threshold for detection, however, such as the accepted SNR of 6dB, can be lenient towards non-linear algorithms and does not allow a trivial calculation to determine false alarms.

False alarms are important as detection should also incorporate reliability; in addition to defects which are detected or missed, the number of false calls each algorithm generates has to be recorded. Without this consideration, an algorithm which detects only the defects, and one which classifies everything as a defect both maintain a perfect

detection record. To calculate false alarms, the truthful outcome of any inspection must be known which once again requires the presumption that knowledge of the defects (true-positives) exists and thereby also the potential false-positives.

This leaves two challenges: how to set a threshold and how to include reliability. One approach, which has enjoyed success in a variety of fields concerned with detection problems, is to rely on Receiver Operating Characteristic (ROC) analysis. The ROC method was originally developed for RADAR [25] to select an optimum operating point at which to distinguish enemy targets from allies and clutter. Later the medical research community adopted the approach [26] to interpret diagnostic accuracy and it subsequently received plentiful attention leading to further developments of analytical tools to interpret ROCs. ROC analysis demonstrates the performance of a classifier, or for instance an imaging algorithm, by calculating true-positive and false-positive rates, or in NDE terminology: Probability of Detection (POD) and Probability of False Alarm (PFA). This is an attractive approach as it circumvents the need for setting an arbitrary decision threshold by instead intentionally varying the decision threshold and observing the change in outcomes.

The ROC method is also well equipped to operate in a low SNR environment, where defects are on the verge of detection. In order to compare imaging algorithms, particularly for NDE of difficult materials, it is precisely these conditions that are of interest. Comparing two different images in terms of their SNR alone, especially when they are both low, does not often allow a meaningful comparison. For instance, they may require their results to be such that their SNRs are situated on opposing sides of the, for example 6dB, threshold before being able to distinguish their detection performance. It must be noted however that it is not suggested to use ROC analysis to redefine detection standards and select a new detection threshold; it is merely suggested as a useful tool for evaluation and comparison of candidate algorithms on cases of interest.

### *B. Flaw Characterisation*

Once flaws are detected, their characterisation becomes important. To truly assess this aspect requires different types and sizes of defects which is usually problematic for experimental studies. This is also true for the sample that we will discuss later in this article, where the defects are relatively large when compared to a typical wavelength used in inspection. Inevitably, an investigation of this would quickly conclude that lower inspection frequencies are superior

as grain scattering effects are reduced. It is therefore important to acknowledge that higher frequencies are useful to detect smaller defects.

A previously adopted approach has introduced a resolution parameter [10] which measures the area of pixels which are within a 6dB range of the peak intensity signal-pixel. Measuring this however, commands detection with relatively high SNR as noise-pixels will otherwise contaminate the measurement. Difficult materials exhibiting low SNR, such as coarse grained materials, are therefore ill-suited for such resolution measurements. While characterisation remains an important performance indicator, when considering inspection of difficult materials, it follows secondary (in sequence) to detection. Thus this methodology will focus on detection and methods for assessing characterisation will be left to future investigation.

### III. IMAGING ALGORITHMS

The three example algorithms considered are hereon summarised; the Total Focusing Method (TFM), Phase Coherent Imaging (PCI), and Decomposition of the Time Reversal Operator with Multiple Scattering Filter (DORT MSF). Figure 1 compares the images using the different algorithms of a relatively well-detected Side Drilled Hole (SDH) in a coarse grained sample (further described in Section 4a).

TFM [10] is named for creating an image which is synthetically focused at every pixel by using the standard sum-and-delay operation, more widely known as beamforming, described in Equation 1. The transmitting  $tx$  and receiving  $rx$  array elements are denoted  $i$  and  $j$  respectively, the  $x$ -coordinate signifies lateral position, the  $z$ -coordinate is the axial dimension,  $c_0$  is the wave speed, and the FMC matrix  $H_{ij}(t)$  of dimensions  $N \times N$  where  $N$  is the number of elements in the array. The image intensity,  $I$ , at location  $x, z$ , is given by Equation 1. The Hilbert Transform can be used to smooth the image.

$$I_{TFM} = \left| \sum_{i=1}^N \sum_{j=1}^N h_{tx,rx} \left( \frac{\sqrt{(x_i - x)^2 + z^2} + \sqrt{(x_j - x)^2 + z^2}}{c_0} \right) \right| \quad (1)$$

Phase Coherent Imaging [16,17] amplifies the contribution of phase information by multiplying the original TFM image by a weighted matrix  $C$  according to the statistical phase  $\varphi$  variability of all  $N^2$  time traces (i,j) for a particular position (x,z). When all phases are equal,  $C=I$ , maintaining pixel intensity, whereas when they approach a uniform

distribution  $C \rightarrow 0$ , reducing pixel intensities. This operation has shown promise [12] to inspect structurally noisy materials as backscatter signals are expected to exhibit greater phase variability than flaw signals. In this case  $I$  is given by:

$$\begin{aligned} I_{PCI} &= I_{TFM} C \\ C &= 1 - \sqrt{\text{var}(\cos\varphi(x, z, i, j)) + \text{var}(\sin\varphi(x, z, i, j))} \end{aligned} \quad (2)$$

DORT [18] relies on calculating the singular value decomposition of the FMC matrix  $K(T, f)$ , where  $T$  and  $f$  represent a time-window and frequency respectively, to achieve imaging of the inspected medium. In weakly scattering conditions, each  $i$ -th scatterer in the medium is related to a singular space of the  $i$ -th singular value  $\lambda_i$ . The corresponding singular vector  $V_i$  is the signal of delays to apply to the array to focus onto the scatterer. When structural noise dominates, significant singular values are no longer associated with scatterers of interest. Part of this noise is contributed to multiple scattering and can be reduced by exploiting the deterministic phase relation along the anti-diagonals of the  $K$  matrix which occurs even for an unknown distribution of disorder, in this case the scatterers inside the medium, for single scattering. This procedure is termed a Multiple Scattering Filter (MSF) [19,20] and has shown great promise for detection in coarse grained materials [21]. DORT MSF involves applying the regular DORT algorithm to the filtered  $K$  matrix,  $K_f$ . Here  $G^*$  is the complex conjugate of the Green's function for a homogenous medium. In this case  $I$  is given by:

$$I_{DORT}(T, f) = \lambda_i |V_i(T, f) G^*(T, f)| \quad (3)$$

The DORT MSF algorithm includes a detection criterion which aims to identify singular values which are associated with scatterers of interest and therefore merit back propagation. This is achieved by considering the statistical properties of the singular value distribution acquired from backscatter in a defect-free area for different frequencies  $f$  and depths related to time  $T$ . This method requires numerous FMCs of defect-free material which is known to be of similar microstructure. In applications of known, well-controlled, microstructure, this has been shown to generate significant improvement [21]. However, in circumstances where the microstructure is spatially varying, such as in this study, the backscatter of defect-free material cannot be related to that from areas which contain defects, and therefore this potential advantage cannot be realised.

#### IV. DATA COLLECTION

This article relies mainly on experimental data obtained from a coarse grained material to demonstrate the proposed methodology whilst simulated data is used to investigate its limitations. The experimental investigation provides an interesting outcome on how the algorithms perform on a real material and the simulated data allows us to generate data with a known outcome in order to contrast the evaluation results with what is expected. Table 1 summarises the parameters for each method of data collection, where both serve to acquire FMC data of imaging targets: a defect in structurally noisy material. As different imaging algorithms may exhibit varying levels of sensitivity towards defects located at large angles, the array is always placed centrally above the defect such that the most optimistic scenario is considered: the defect is located where the array has the highest sensitivity. In terms of post processing, the data is filtered about the centre-frequency of the array used. Similarly, whereas previous research [27] has shown that, for certain inspections, image SNR can be optimised by using sub-apertures of the array, this investigation uses the full aperture.

##### *A. Experimental Data Collection*

The experimental data are collected from an INCONEL 625 block (kindly loaned by Siemens AG and seen in Figure 2a) which has been cut from a larger cast thick-walled cylinder representing a future high-temperature power plant component. Surface etchings indicate the grain size distribution to vary from sub-millimetre to tens of millimetres in a seemingly arbitrary spatial distribution. This variation is confirmed by conventional A, B, and C-scans where for example, the pulse-echo backwall amplitude at a depth of 70mm depth, varied significantly across a 2-D immersion scan of the side face (see Figure 2b). This behaviour is believed to be a consequence of grain size variations shaped by the manufacturing conditions and has previously been observed for other Nickel super alloys by Thompson et al. [3]. Given that the part stems from a larger power plant component, it is likely that this effect will be reproduced in future high-temperature power plant components in similar uses, thus adding further complexity to their inspection. Further to this, significant variations in wave speed ( $5700\text{-}6000\text{ms}^{-1}$ ) were found throughout the block. As this work focuses on the proposed methodology for evaluating images, rather than on the actual comparison of specific algorithms, we have not pursued any further investigation of the details of the grains. Furthermore, it is often the case in industrial context that no further information than this is available.

The sample contains three 5mm diameter Side-Drilled-Holes (SDH) found at various depths as seen in Figure 2a. FMC data are acquired using two faces of the sample, in a contact configuration with the array orientated perpendicular to the SDH, to provide five imaging targets at different depths: 50mm, 60mm, 100mm, 110mm, and 150mm. For each target, array data sets are acquired at a series of locations along the SDH's length, although in each case centrally above it, to capture multiple FMCs of the same defect surrounded by a varying microstructure. The maximum number of FMCs which can be acquired in this way is limited by the required minimum step between the measurement data sets, and the maximum available scan area, determined by the thickness of the sample. A minimum separation distance ensures that each measurement is independent, dictated by and approximately equal to the grain size [28]. In this case, the constraints permitted approximately ten independent FMC measurements per imaging target. Two 64 element array probes (manufactured by Imasonic, France), with different centre-frequencies of 1MHz and 5MHz, are employed to represent a relatively low and high inspection frequency (see Table 1). The arrays are controlled by an OPEN System (manufactured by Lecoer, France) and excited using a negative square pulse having a time duration equal to half the period at the probe centre-frequency.

### *B. Simulation of Scattering Materials*

In order to control the outcome of an inspection, an existing simulation model [24] is used to generate FMC data which are representative of those obtained from scattering materials. It assumes that image speckle associated with grain scattering can be generated by calculating the acoustic response from a random sub-wavelength distribution of point scatterers. Under a single-scattering approximation, such a response can be calculated even for a large numbers of scatterers (e.g. 5000), whereas this would be more challenging for a complete scattering model.

The grain scatterers are represented by point scatterers which behave as omnidirectional scatterers with amplitude  $S^g$ . It has been shown [24] that 5 scatterers per wavelength squared is a representative and computationally efficient density of scatterers to include. At least 15 realisations are required to produce a converged SNR at this density [24]. Realistic defects such as cracks can be introduced into the model by describing their behaviour using a scattering matrix [29], calculated for example by Finite Element models. This article however only considers a defect in the form of another point scatterer with a scattering amplitude  $S^d$ . The ratio, relative to the grain scattering amplitude  $S^g$ , is defined as  $a$  shown in Equation 4. Assuming the scatterer density is fixed, the level of backscatter can be controlled by varying  $a$ , where larger values lead to stronger scattering.



$$a = \frac{S^g}{S^d} \quad (4)$$

The frequency response  $E_{ij}(\omega)$  of a single point scatterer, for both grains and defects, is calculated using Equation 5.

$$E_{ij}(\omega) = A(\omega)B^{rx}(r_i, \omega)G^{rx}(r_i, \omega)B^{tx}(r_j, \omega)G^{tx}(r_j, \omega)S \quad (5)$$

$A(\omega)$  is the frequency spectrum of the transmitted signal and  $r$  is the distance from array element  $i$  or  $j$  to the point scatterer.  $B$  is the transducer directivity function [30] where  $rx$  and  $tx$  denote reception and transmission respectively. Similarly the Green's functions,  $G^{rx}$  and  $G^{tx}$  calculate the complex amplitude and phase received from a particular combination of receiver and transmitter element from the point scatterer for a homogenous equivalent medium. As dictated by the Born Approximation, the FMC matrix  $H$  can be obtained by super-positioning the individual contributions  $E^g$  of all  $n$  grain scatterers and that of the defect  $E^d$  as shown in Equation 6. The inverse Fourier transform of  $H(\omega)$  is used to extract the time traces and obtain the time-domain FMC matrix  $H(t)$ .

$$H_{ij}(\omega) = E_{ij}^d(\omega) + \sum_{m=1}^n E_{ij}^{g_m}(\omega) \quad (6)$$

The model simulates a 32 element 5MHz array (see Table 1) with two different scattering environments representing a well and a poor detection scenario, by setting  $a$  equal to 0.1 and 0.25 respectively. Each scenario involves 100 FMCs for different random realisations of grain scatterer positions, each realisation contains a single defect located at an axial depth of 20mm centrally below the array. Figure 3 shows one TFM image for each case.

## V. MEASURING RELIABLE DETECTION

Once the FMC data are acquired, they are post-processed to produce images for each algorithm. True to the rationale defined in Section II, ROC analysis is used to compare the performance of each algorithm. Whereas this analysis has been used in a qualitative manner where operators interpret images, the method outlined here relies on a basic form of image segmentation to quantitatively identify image objects from noise. This is in part possible due to the relative simplicity of the imaging targets considered which produce simple shapes (e.g. circular) which can be recognized by their increased intensity compared to the background. Hence there is no need for more complex pattern recognition, although this could be an avenue worth pursuing for future investigations which consider multifaceted defects.

The procedure for producing ROC plots follows these steps:

- 1) The images are converted into signal and noise datasets by adopting an image segmentation procedure. The procedure is possible because the defects are known, and is therefore only suited for a comparison methodology and not for real deployment of unknown defects. First, an area (a box) is mapped around the flaw response and the Root-Mean-Square (RMS) of the contained pixel intensities is calculated. This represents the signal amplitude and its area can be referred to as the signal box. For most experimental measurements, this will contain both signal and noise; a trade-off which increases and decreases depending on how well or poorly the object is detected. A TFM image of a simulated defect in a scattering material can be seen in Figure 4a where the outline depicts the signal box. The dimensions of the box are set to agree with the expected size of the flaw indication, determined by observing a well detected case, either experimentally or through simulation. Its shape can be modified according to the imaging target. For instance, in the case of inclined cracks, it may be more suitable to use two separated boxes which overlap with the crack tip responses.
- 2) Determining the noise dataset involves segmenting the remainder of the image (minus the signal box) into sections, here for demonstration purposes, of the same dimensions as the signal box. These can be considered as super-pixels which, similarly to the signal box, are calculated from the RMS value of the sub-pixels they contain. The super-pixel version of the TFM image in Figure 4a is shown Figure 4b. All the RMS values are then attributed to a noise dataset although any super-pixels which intersect with the signal box (the black outline in Figure 4b) are omitted from the noise distribution. (These are shown in white in Figure 4b.) The significance of the box dimensions is investigated in the following sub-section.
- 3) This procedure is repeated for all images of the defect to produce multiple images containing the same defect and different background noise. The use of numerous images, namely acquiring multiple FMC data of the same imaging target, is important to obtain two data distributions, one for the signal amplitudes and another containing noise values. Typical histograms for the distribution of a well and poorly detected case for the TFM algorithm, corresponding to average SNRs of 8dB and 3dB respectively, are shown in Figure 5. These histograms are obtained from simulated data using the model mentioned in Section III. Contrasting to real measurements, the simulated defect reflects a constant amplitude, which therefore only fluctuates according to noise in the defect box, and therefore its distribution is also similar to that of the noise. However, in

experimental measurements, like those of coarse grained samples, a variety of reasons can lead to signals fluctuating, including varying attenuation and anisotropy.

- 4) The ROC is calculated by establishing thresholds, varying from 0 to infinity, to the noise and signal distributions. Any data which exceeds the threshold is regarded as detected; the POD is calculated by the proportion of signal data which satisfy this criterion whilst the noise data contribute towards false alarms (PFA). The threshold at zero produces the point (1,1) in the ROC space whilst that at infinity produces (0,0), and any thresholds in between result in a combination of POD and PFA. The ROCs for the two cases shown in Figure 5 can be seen in Figure 6. The well detected case shows a POD/PFA trade-off where almost all defects can be detected without any false alarms. For the poorly detected scenario (Figure 6b), however, a significant number of false alarms arise before all defects are detected.

- 5) The previous steps are repeated for each candidate algorithm to produce a ROC plot for each imaging target.

It is important to investigate and consider the limitations of the proposed method. To enable a further discussion of ROC plots requires quantifying a parameter to describe them; the Area Under the ROC (AUC) is usually defined for this purpose [31]. Although there are drawbacks to reducing the ROC to a single index, it is a convenient description of the position of an ROC curve within the ROC space i.e. an AUC of unity denotes perfect detection and an AUC of 0.5 describes the chance line corresponding to a randomised classifier.

#### *A. Segmentation Procedure*

Calculations of signal and noise amplitudes, whether this involves windowing a time-domain signal or spatially segmenting an image, are susceptible to the settings used. In this case, the signal and noise box dimensions represent such settings and affect the outcome of the ROC curve by either lowering or increasing the measured signal and noise amplitudes. Although the method was previously demonstrated by using sensible parameters for box sizes (in this case 2mm was used), the AUC is now measured as a function of varying the size for both parameters in order to investigate their impact.

Figure 7 plots the AUC for both the well detected and poorly detected case. Both scenarios show that larger signal boxes increase false alarms, arising from the averaging effect which lowers the measured RMS amplitude. Noise boxes behave in the opposite way; small box sizes introduce peaks into the noise distribution which generate false alarms. From the well detected case, where it can be safely stated that the defect is detected, lower extremes for the

noise box dimensions lead to a relatively sensitive result, which is undesirable. In this particular case the lower limit on the box size is 2mm for both signal and noise. For future investigations and use of this methodology, the appropriate choice of box size will depend on the imaging object and resolution of the algorithm considered. However, the results suggest that the method is largely insensitive to the box size, as long as extremes are avoided, and therefore this is not considered to be critical.

The image segmentation procedure is a quantitative method to emulate the qualitative human interpretation of images. It is important to note that the detection of objects does not only depend on their intensity or SNR, but also on their size, where larger objects require less SNR to be detected. In the example here, by adopting the same box size for signal and noise, the calculation assumes only noise objects which are spatially similar to the signal can bring rise to false alarms. Such reasoning is similar to a measure of SNR which only considers noise occupying the same frequency spectrum of the signal as noise at other frequencies could easily be filtered. Thereby, the procedure adopts the most optimistic detection scenario, analogous to considering a maximum SNR, whereby imaging artifacts of different sizes to the imaging target are omitted.

### *B. Further Sampling Considerations*

A point of concern for any statistical method is the required sample size. There is no simple answer as to how many cases are required to ensure a particular statistical precision of an ROC curve [25]. It is known however that the precision does increase as more cases are included and as larger AUCs are considered. Increased precision allows smaller apparent differences between resulting curves to become meaningful in order to distinguish classifier performance. Therefore, if the analysis uses only a low number of images, only broad conclusions can be drawn between the imaging algorithms. Statistical analysis of the ROC approach can be found in [31]. In our case here, sample size does not pose a significant concern as convergence is not necessarily required to draw relative differences between algorithm performances. However, it is still beneficial to include as many independent FMC data sets as possible for each defect.

In a similar nature, curves drawn with discrete data produce staircase effects. This can make the curve appear jagged (as we will see is the case for the experimental data in Fig. 8, 9). This is investigated in [26] where the ROC plots of binned and unbinned data are compared. It is shown that the ROC plot can be affected and its approximation

becomes poor when the data becomes too coarsely binned. Once again, as more FMCs are included, the curves should begin to resemble their continuous equivalent.

## VI. IMAGING ALGORITHM EVALUATION

Here we present an example of the evaluation of the performance of imaging algorithms using the proposed methodology. We use the array measurements obtained from the Inconel block as described in Section III, and perform the imaging using the TFM, PCI, and DORT MSF algorithms.

### A. Results

Figures 8 and 9 show the ROC results for the detection of SDHs at various depths for low and high frequency respectively (1MHz and 5MHz). The points on the ROC show the step function, whereas the lines show the optimal performance of an algorithm, in ROC terminology this is named the convex-hull [32]. At low frequency (Figure 8), all candidate algorithms detected their target at all depths but the deepest (150mm). Even though some algorithms are superior in terms of SNR (e.g. Figure 1), the ROC results show their performance in this case cannot be distinguished. For the purposes of a performance evaluation, this results in a redundant outcome and more challenging detection targets are required. Higher frequencies provide this situation: the same SDHs become more difficult to detect as grain scattering increases.

At 5MHz, only the 60mm depth hole remained perfectly detected (AUROC of unity and this case was covered in Figure 6), the 150mm (Figure 9d) and 100mm (Figure 9b) depth holes were not detected at all, and the 50mm (Figure 9c) and 110mm (Figure 9a) were partially detected. When defects are not detected, the ROC curve approaches the chance line which represents a randomised classifier, shown as a dotted diagonal line in Figure 8. Whilst the higher frequency results provided the desirable scenario where targets are on the verge of detection, the results indicate that there is little performance difference between the algorithms. Table 2 summarises this by calculating the detection rates for all defects by arbitrarily setting an allowable PFA at 5% (threshold for 5% PFA shown in Figure 8).

### B. Discussion

The results show that even though some de-noising algorithms seem to increase SNR when flaws are detected, this had no impact on actual defect detection for this inspection setup as the considered algorithms, TFM, PCI, and DORT MSF, all performed very similarly. This observation holds true for both 1MHz and 5MHz array data, even though the majority of defects were detected at the lower frequency and therefore it was not possible to distinguish

algorithms. Although both PCI and DORT MSF detected slightly more defects, this was only in the order of a few percent (see Table 2) when considering the defect population for a given false-call rate (5%), and hence the difference was not statistically significant.

Several possible reasons can explain the similarities in performance which were observed. Given the high levels of scattering noise, the material sample can be assumed to behave as a medium with infinite scatterers, with only one of those scatterers, the SDH, being the scatterer of interest. For phase coherence to exist, and algorithms like PCI to exploit this, the scatterer of interest would have to exhibit a greater amplitude signal such that its signal and therefore also its phase dominates the signals originating from the remaining scatterers. In addition, the DORT MSF algorithm can be expected to perform better when the grains are well characterised and the threshold is implemented as is intended by its developers [19-21]. A lack of multiple scattering effects may also hinder its capacity for improvement in this example.

Limitations of this particular comparison include the reliance on artificial defects. SDHs behave differently to real defects; signals received from cracks may be more phase incoherent which can cause algorithms such as PCI to behave undesirably as mentioned in [24]. Conversely, there may be characteristics of real defects that improve the performance of certain algorithms. Thus even though SDHs are an established benchmark, the conclusion for this specific case must be read with caution.

Given that the sample is representative of future power plant components, it is worthwhile to discuss some of the inspection findings. Overall, the array detection results show the sample to be difficult to inspect, regardless of which algorithm is used. The varying scattering behaviour throughout the block meant that detectability of defects did not necessarily decrease with depth, as the 50mm depth SDH exhibited a lower detection rate than one at 60mm. The sample confirmed that there is still a need to improve ultrasonic NDE of these materials before they can be reliably inspected. We emphasise again that these cases are presented as examples to illustrate our comparison methodology. We do not infer general conclusions about these imaging algorithms.

## VII. CONCLUSIONS

This article proposed a methodology to enable the detection performance of different imaging algorithms to be quantified and thereby allow comparisons of the performance of algorithms for candidate inspection applications and thus to progress ultrasonic NDE of difficult materials. By relying on ROC analysis and taking into consideration both

POD and PFA, the procedure aims to deliver a robust and objective evaluation. Investigation of its limitations and the impact of the required parameters suggest its settings are relatively insensitive to the results. The only main requirement is prior knowledge of the truthful outcome for an inspection, which in a research context is often the case.

Advanced de-noising ultrasonic array imaging algorithms, PCI and DORT MSF, were matched with TFM for the detection of Side-Drilled-Holes (SDH) in a sample of coarse grained power plant material, INCONEL 625. The results, applicable just to this particular case, show the algorithms to perform very similarly as no significant differences were observed in their performance. In other applications cases it may well be that there will be differences in performance, and it is proposed that this methodology would be useful to make such an assessment on arising cases of interest.

### VIII. ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Research Council (EPSRC) and the Engineering Doctorate Programme of the UK Research Centre in NDE (RCNDE). The authors acknowledge this support with gratitude. Additional thanks go to EDF R&D for the helpful discussions

### REFERENCES

- [1] S Hirsekorn, "The Scattering of ultrasonic-waves by polycrystals," *J. Acoust. Soc. Am.*, vol. 72, no. 3, pp.1021-1031, 1982.
- [2] F. Margetan, R. Thompson, I. Yalda Mooshabad, "Backscattered microstructural noise in ultrasonic toneburst inspections," *Journal of NDE*, vol. 13, no. 3, pp. 111-136, 1994.
- [3] R. Thompson, F. Margetan, P. Halidpur, L. Yu, A. Li, P. Panetta, H. Wassan, "Scattering of elastic waves in simple and complex polycrystals," *Wave Motion*, vol. 45, no. 5, pp. 655-674, 2008.
- [4] G. Ghoshal, J. A. Turner, "Numerical model of longitudinal wave scattering in polycrystals", *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 56, pp.1419-1428, 2009.
- [5] P. Panetta, F. J. Margetan, I. Yalda, R. B. Thompson, "Ultrasonic attenuation measurements in jet-engine titanium alloys" in *Review of Progress in QNDE Proceedings*, vol. 15, pp. 1525-1532, 1993.
- [6] L. Yang, S. I. Rokhlin, "Ultrasonic backscattering in cubic polycrystals with ellipsoidal grains and texture", *Journal of Nondestructive Evaluation*, vol. 32, no. 2, pp. 142-155, 2013.
- [7] G. Stein-Brzozowska, D. Florez, J. Maier, G. Scheffknecht, "Nickel-base superalloys for ultra-supercritical coal-fired power plants: Fireside corrosion. Laboratory studies and power plants exposures," *Fuel*, vol. 108, pp. 521-533, 2013.
- [8] P. D. Wilcox, "Ultrasonic arrays in NDE: beyond the B-scan," in *Rev. of Prog. in QNDE Proc.*, vol. 33, pp.33-50, 2013.
- [9] B. Drinkwater, P. Wilcox, "Ultrasonic arrays for non-destructive evaluation: A Review," *NDT&E Int.*, vol. 39, no. 7, pp. 525-541, 2006.
- [10] C. Holmes, B. W. Drinkwater, P. D. Wilcox, "Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation," *NDT & E Int.*, vol. 38, pp. 701-711, 2005.
- [11] P. Wilcox, "Array imaging of noisy materials," in *Re. Prog. QNDE*, vol. 30, pp. 890-897, 2011.
- [12] N Portzgen, D. Gisold, G. Blacquiere, "Inverse wave field extrapolation: A different NDI approach to imaging defects," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 54, no. 1, pp-118-127, 2007.
- [13] J. Hunter, B.W. Drinkwater, and P. D. Wilcox, "The wavenumber algorithm for full-matrix imaging using an ultrasonic array," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 55, no. 11, pp. 2450-5462, 2008.
- [14] J. Zhang, A. Hunter, B. W. Drinkwater, P. D. Wilcox, "Monte Carlo inversion of ultrasonic array data to map anisotropic weld properties," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 59, no. 11, pp. 2487-2497, 2012.
- [15] G. Connolly, M. Lowe, J. Temple, and S. Rokhlin, "The application of Fermat's principle for imaging anisotropic and inhomogeneous media with application to austenitic steel weld inspection" *Proceedings of the Royal Society A*, vol. 465, pp. 3401-3423, 2009.
- [16] J. Camacho and C. Fritsch, "Phase coherence imaging of grained materials," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 58, no. 5, pp. 1006-1015, 2011.
- [17] J. Camacho, M. Parrilla, and C. Fritsch, "Phase coherence imaging," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 56, no. 5, pp. 958-974, 2009.
- [18] C. Prada, S. Manneville, D. Spoliansky, M. Fink, "Decomposition of the time reversal operator: Detection and selective focusing on two scatterers," *J. Acoust. Soc. Am.*, vol. 99, no. 4, pp. 2067-2076, 1996.

- [19] A. Aubry, A. Derode, "Detection and imaging in a random medium: A matrix method to overcome multiple scattering and aberration," *J. Appl. Phys.*, vol. 106, no. 4, pp. 044903-044903-19, 2009.
- [20] A. Aubry, A. Derode, "Random matrix theory applied to acoustic backscattering and imaging in complex media," *Phys. Rev. Lett.*, vol. 102, no. 8, pp. 84301, 2009.
- [21] S. Shahjahan, "Improvement of flaw detection with ultrasonic array probes in multiple scattering polycrystalline materials by means of a random matrix approach: experimental and simulation studies" Ph.D. dissertation, Université Paris-Diderot, Paris, 2013.
- [22] G. E Trahey, S. W. Smith, O. T. Von Ram, "Speckle pattern correlation with lateral aperture translation: Experimental results and implications for spatial compounding," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 33, no. 3, pp. 257-264, 1986.
- [23] T. Lardner, M. Li, R. Gongzhang, A. Gachagan, "A new speckle noise suppression technique using cross-correlation of array sub-apertures in ultrasonic NDE of coarse grain materials," in *Rev. of Prog. in QNDE Proc.*, vol. 33, pp. 865-871, 2013.
- [24] J. Zhang, B. W. Drinkwater, P. D. Wilcox, "Comparison of ultrasonic array imaging algorithms for non-destructive evaluation," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 60, no. 8, pp. 1732-1745, 2013.
- [25] C. E. Metz, "Basic principles of ROC analysis", *Seminars in Nuclear Medicine*, vol. 8, no.4, pp. 283-297, 1978.
- [26] M. H. Zweig, G. Campbell, "Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clin. Chem.*, vol. 39, no. 4, pp. 561-577, 1993.
- [27] C. Li, D. Pain, P. D. Wilcox, B. W. Drinkwater, "Imaging composite material using ultrasonic arrays," *NDT & E Int.*, vol. 53, pp. 8-17, 2012.
- [28] L. Yu, R. B. Thompson, Margetan F. J, "The spatial correlation of backscattered ultrasonic grain noise: theory and experimental validation." *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 57, no. 2, pp.363-378, 2010.
- [29] J. Zhang, B. Drinkwater, and P. Wilcox "Defect characterization using an ultrasonic array to measure the scattering coefficient matrix", *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.*, vol. 55, no.10, pp. 2254-2265, 2008.
- [30] F. G. Miller and H. Pursey, "The field and radiation impedance of mechanical radiators on the free surface of a semi-infinite isotropic solid," *Proc. R. Soc. Lond.*, vol. 223, no.15, pp. 521-541, 1954.
- [31] J. A. Hanley, B. J. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.
- [32] T. Fawcet, "An introduction to ROC analysis," *Pat. Rec. Lett.*, vol. 27, pp. 861-874, 2006.



## LIST OF FIGURES

Fig.1. Images (40dB dynamic range) for (a) TFM (b) PCI (c) DORT MSF of a 60mm depth Side Drilled Hole (SDH) in Inconel 625 using a 5MHz 64 element array located centrally at (0,0).

Fig. 2. (a) Inconel 625 step wedge and (b) the C-scan amplitude image of the step wedge and an ultrasonically clean steel cylinder, taken with a 2.25MHz focused probe. The C-scan setup is a double through-transmission: the measured amplitude is that of the reflection from the back wall; the thickness is 70mm. The amplitude variation seen in the step wedge can be compared to the steel cylinder for scale. Lighter and darker shading represent high and low amplitude respectively.

Fig. 3. TFM image obtained from simulated data. Both images contain a point source at (20,0) with a relatively high (a)  $a=1/10$  and low (b)  $a=1/4$  scattering amplitude to simulate a well and poorly detected scenario respectively.

Fig. 4. (a) Typical TFM image of a defect (outlined in white) in a scattering medium. (b) The processed version of the TFM image in (a) where the super-pixels and their RMS value has been calculated. The single super-pixel, holding the signal RMS value which corresponds to the defect, is outlined in black. The omitted noise super-pixels which neighbour the signal pixel are shown in white.

Fig. 5. Examples of the signal and noise distributions for (a) a well detected case with an average SNR of 8dB and (b) a poorly detected case with an average SNR of 3dB. The distributions were measured from simulated data imaged using TFM.

Fig. 6. Examples of ROC curves corresponding to the (a) well detected case and (b) poorly detected case of Figure 5 for the TFM algorithm.

Fig. 7. Influence of box size on ROC curve for the well detected case (a) and poorly detected case (b) of Figure 5. AUC is the Area Under the Receiver Operator Characteristics curve where 1 represents perfect detection; and 0.5 a random classifier.

Fig. 8. Low frequency (1MHz 64 element array) ROC results for (a) the SDH at 60mm depth which here exemplifies perfect detection (all defects are detected without raising false alarms). The diagonal dotted line represents a random classifier or chance line. The threshold for a 5% PFA is also illustrated by the dashed vertical line. (b) ROC result for 150mm depth where the detection record is no longer perfect, but the candidate algorithms produce similar results.

Fig. 9. High frequency (5MHz 64 element array) ROC curves for (a) 110mm, (b) 100mm, (c) 50mm, and (d) 150mm depth SDHs. The detection rate can be seen to vary, where the defects at 100mm and 150mm were particularly difficult to detect.

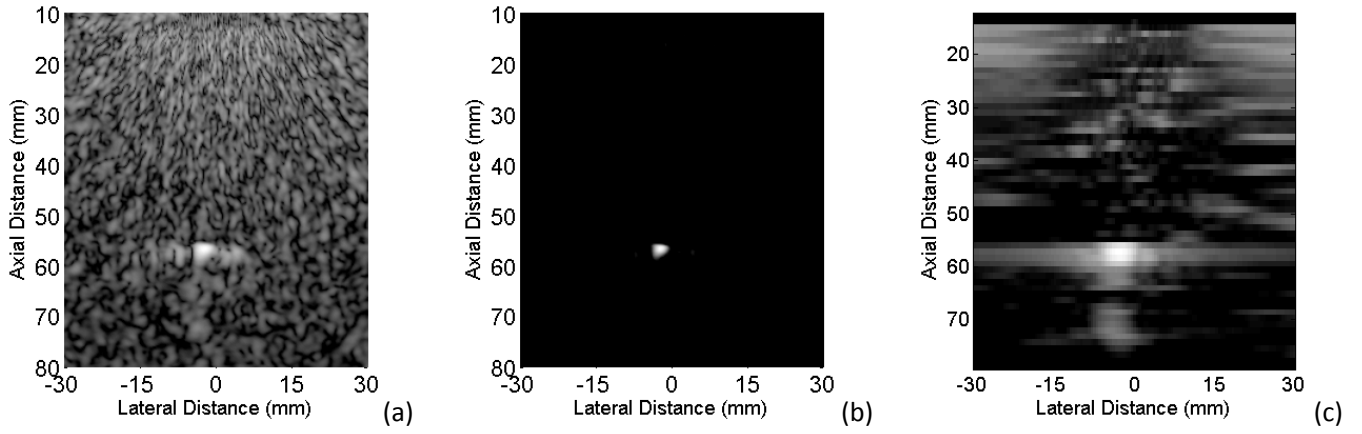


Fig.1. Images (40dB dynamic range) for (a) TFM (b) PCI (c) DORT MSF of a 60mm depth Side Drilled Hole (SDH) in Inconel 625 using a 5MHz 64 element array located centrally at (0,0).

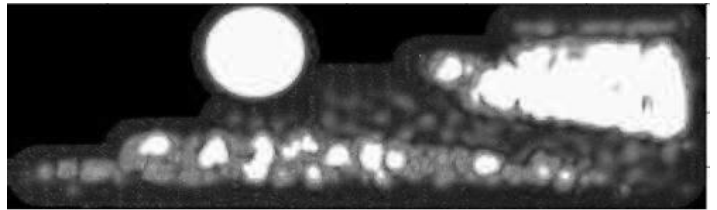
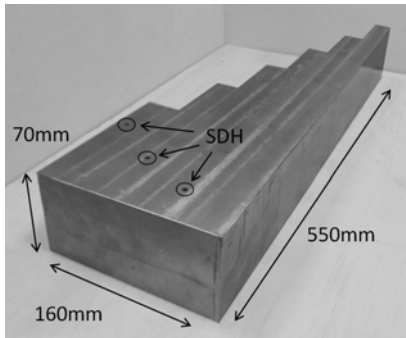
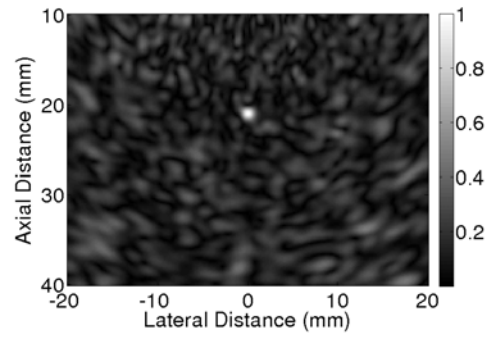
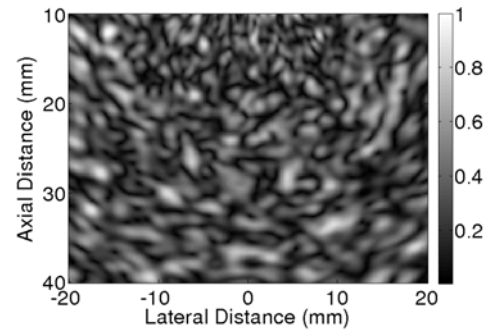


Fig. 2. (a) Inconel 625 step wedge and (b) the C-scan amplitude image of the step wedge and an ultrasonically clean steel cylinder, taken with a 2.25MHz focused probe. The C-scan setup is a double through-transmission: the measured amplitude is that of the reflection from the back wall; the thickness is 70mm. The amplitude variation seen in the step wedge can be compared to the steel cylinder for scale. Lighter and darker shading represent high and low amplitude respectively.



(a)



(b)

Fig. 3. TFM image obtained from simulated data. Both images contain a point source at (20,0) with a relatively high (a)  $a=1/10$  and low (b)  $a=1/4$  scattering amplitude to simulate a well and poorly detected scenario respectively.

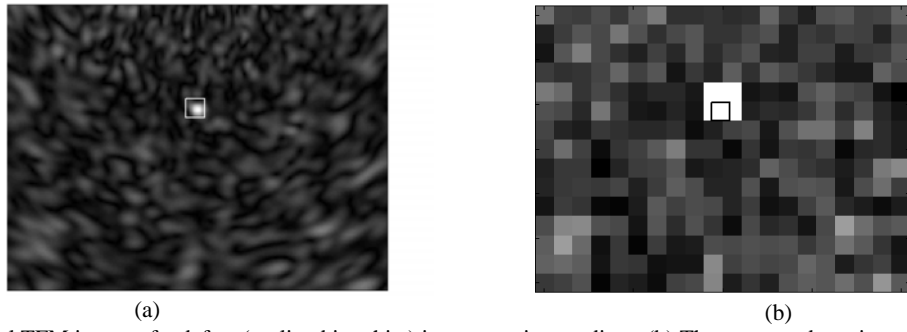


Fig. 4. (a) Typical TFM image of a defect (outlined in white) in a scattering medium. (b) The processed version of the TFM image in (a) where the super-pixels and their RMS value has been calculated. The single super-pixel which corresponds to the defect is outlined in black. The area of omitted noise super-pixels which neighbours the signal pixel is shown in white.

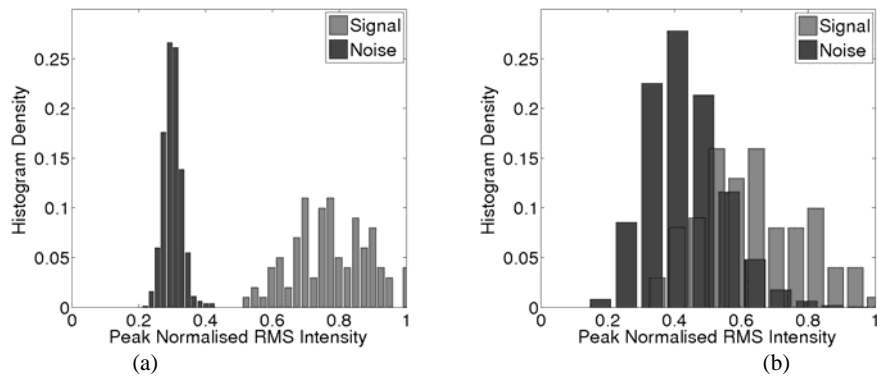


Fig. 5. Examples of the signal and noise distributions for (a) a well detected case with an average SNR of 8dB and (b) a poorly detected case with an average SNR of 3dB. The distributions were measured from simulated data imaged using TFM.

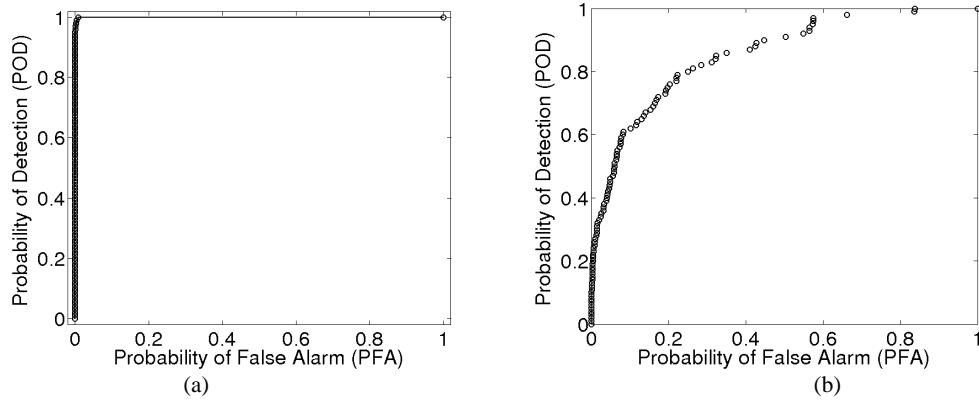


Fig. 6. Examples of ROC curves corresponding to the (a) well detected case and (b) poorly detected case of Figure 5 for the TFM algorithm.

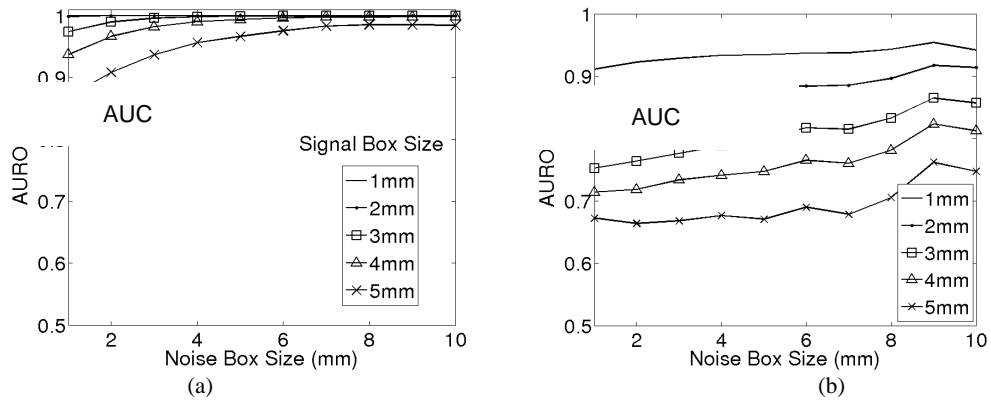


Fig. 7. Influence of box size on ROC curve for the well detected case (a) and poorly detected case (b) of Figure 5. AUC is the Area Under the Receiver Operator Characteristics curve where 1 represents perfect detection; and 0.5 a random classifier.



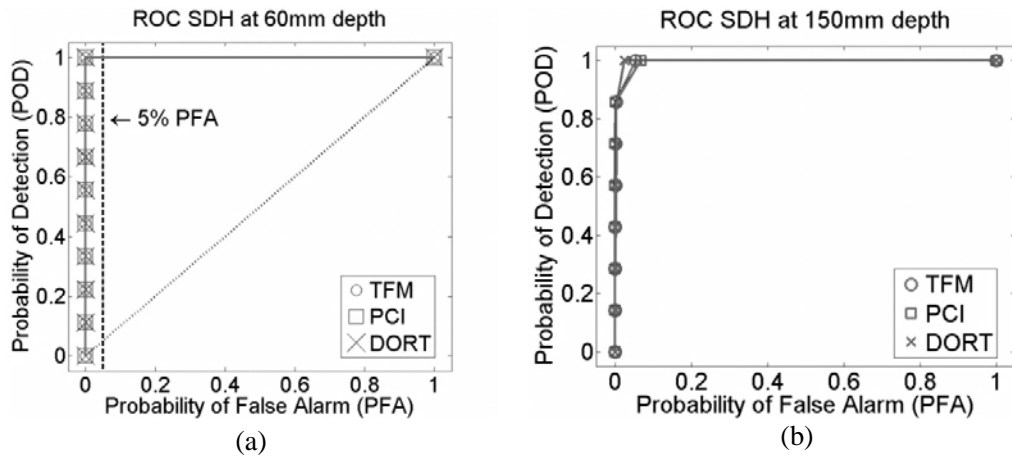


Fig. 8. Low frequency (1MHz 64 element array) ROC results for (a) the SDH at 60mm depth which here exemplifies perfect detection (all defects are detected without raising false alarms). The diagonal dotted line represents a random classifier or chance line. The threshold for a 5% PFA is also illustrated by the dashed vertical line. (b) ROC result for 150mm depth where the detection record is no longer perfect, but the candidate algorithms produce similar results.

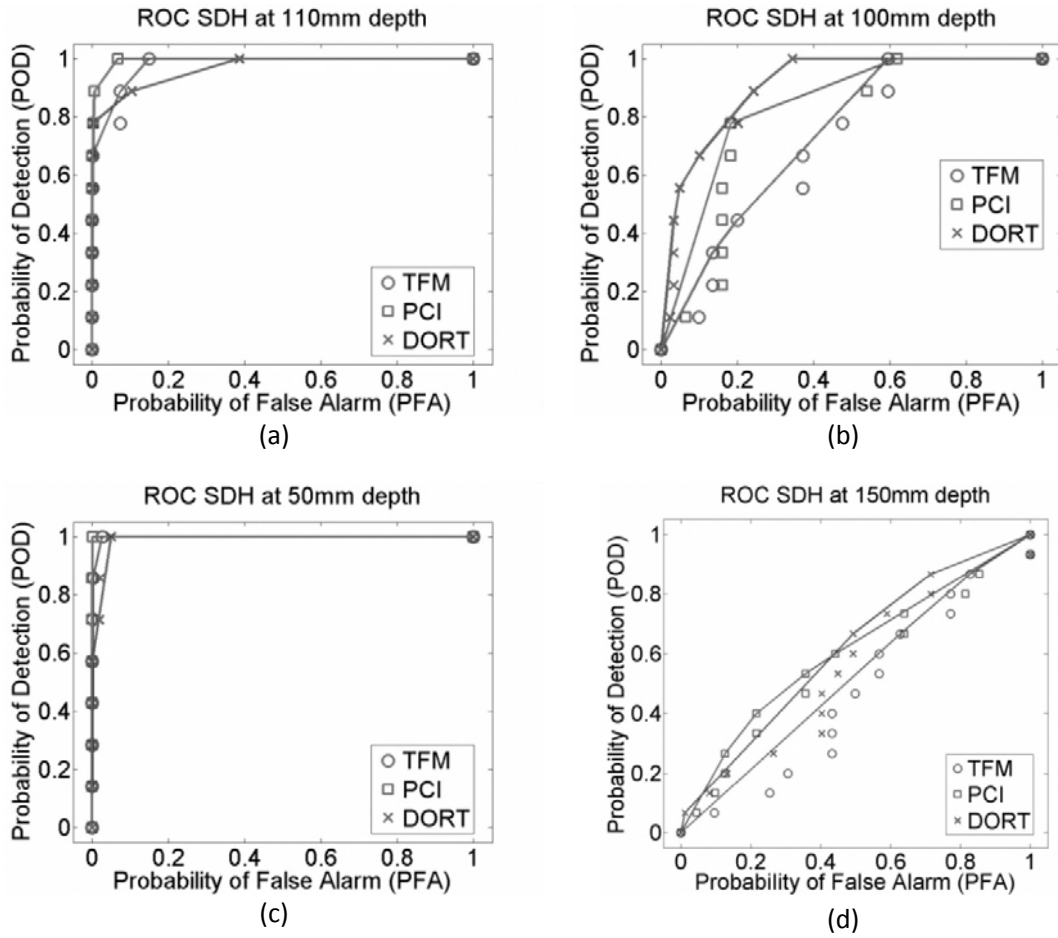


Fig. 9. High frequency (5MHz 64 element array) ROC curves for (a) 110mm, (b) 100mm, (c) 50mm, and (d) 150mm depth SDHs. The detection rates can be seen to vary, where the defects at 100mm and 150mm were particularly difficult to detect.

TABLE I. SUMMARY OF SIMULATION AND EXPERIMENT PARAMETERS.

Parameter	Simulation	Experiment: high freq. array	Experiment: low freq. array
Number of elements	32	64	64
Element pitch	0.6mm	0.6mm	2.1mm
Element width	0.5mm	0.5mm	1.6mm
Element length	15mm	15mm	20mm
Centre-frequency	5MHz	5MHz	1MHz
Bandwidth (-6dB)	64%	65%	50%
FMCs per defect	100	~10	~10

TABLE 2 . SUMMARY OF RESULTS: PORTION OF DEFECTS DETECTED BY EACH ALGORITHM FOR A SET 5% PFA.

Imaging Algorithm	Low Frequency	High Frequency
TFM	100%	51%
PCI	100%	55%
DORT MSF	100%	54%