Noname manuscript No.
(will be inserted by the editor)

Automated classification of transient contamination in stationary acoustic data

Christopher J. Bahr · Todd Schultz

the date of receipt and acceptance should be inserted later

Abstract An automated procedure for the classification of transient contamination of stationary acoustic data is proposed and analyzed. The procedure requires the assumption that the stationary acoustic data of interest can be modeled as a band-limited, Gaussian random process. It also requires that the transient contamination be of higher variance than the acoustic data of interest. When these assumptions are satisfied, it is a blind separation procedure, aside from the initial input specifying how to subdivide the time series of interest. No a priori threshold criterion is required. Simulation results show that for a sufficient number of blocks, the method performs well, as long as the occasional false positive or false negative is acceptable. The effectiveness of the procedure is demonstrated with an appli-

Christopher J. Bahr

Aeroacoustics Branch, NASA Langley Research Center, Hampton, Virginia, USA

Tel.: 1-757-864-9162

E-mail: christopher.j.bahr@nasa.gov

ORCID: 0000-0002-3095-4265

Todd Schultz

Boeing Test & Evaluation, The Boeing Company, Seattle, Washington, USA

cation to experimental wind tunnel acoustic test data which are contaminated by hydrodynamic gusts.

Keywords binary classification · noise contamination · unsupervised methods

Nomenclature

B = normalized signal bandwidth

K = Kullback-Leibler divergence

M = Mach number

N = number of samples in a block of data

n = sample index

P = probability distribution function

p = probability density function

Q = probability distribution function, estimate of P

q = probability density function, estimate of p

y_n = individual sample in a block of data

α = gamma distribution shape parameter

β = gamma distribution scale parameter

Γ = gamma function

γ = incomplete gamma function

ν = effective degrees of freedom for signal of block size N

σ^2 = variance of a block of data

χ_N^2 = sum of the squares of the samples in a block of data

1 Introduction

In aeroacoustic wind tunnel testing, experimentalists often seek to measure acoustic signals, which can be modeled as band-limited, stationary random processes. The unfortunate reality for some experimental setups is that the acoustic signal of interest will be measured along with some form of contamination. For example, in an open-jet and acoustically-treated wind tunnel facility, the contamination observed by a microphone may manifest as either stationary pressure fluctuations generated by facility acoustic sources, or transient pressure fluctuations generated by flow over the microphone (Soderman and Allen 2002). Stationary contamination may be mitigated through various forms of frequency domain background subtraction (Humphreys et al. 1998; Bahr and Horne 2017). However, such techniques are not appropriate for transient events.

Alternative analysis methods are required to classify and separate time domain contamination. While manual inspection of data is an option, this is usually impractical due to the large volume of data involved. Simple methods such as Chauvenet's criterion (Coleman and Steele 1999) allow for the classification of outliers in Gaussian-distributed random data. Advanced methods are available for analyzing more complicated scenarios as shown in Aggarwal (2017) and Hawkins (1980), for example. However, rejection of individual samples may not be beneficial in time series analysis, as continuous blocks of data are usually required for subsequent spectral processing. Common block properties, as discussed subsequently, do not follow a Gaussian distribution and thus do not lend themselves to analysis with basic tools. More advanced tools, to the authors' knowledge, have not been developed with this specific type of classification problem in mind wherein a priori data assumptions and thresholding parameters are minimized.

This work presents a tailored, alternative method which requires minimal input aside from the parameters to subdivide a given time series of interest into individual blocks, sized according subsequent analysis needs. The identification and separation methodology has a well-defined parameter for classifying transient data, which should be valid as long as the underlying assumptions are approximately obeyed. It is assumed that the acoustic signal of interest is a stationary, zero mean, Gaussian random process. With this assumption, the block variances can be modeled using a gamma distribution. It is subsequently assumed that the acoustic data of interest are of lower variance than transient contaminating data. Both mean- and median-based distributions are computed and compared, making the method robust to extreme values. The detailed development of this classification technique is given in the following section. Subsequent sections evaluate the classification performance with both simulated and experimental data. These are followed by recommendations developed from the results.

2 Theoretical Development

The first assumption required for this transient classification procedure is that the underlying acoustic signal is a stationary, zero mean, Gaussian random process. If the samples from the acoustic signal of interest, y, are truly Gaussian-distributed random variables with zero mean and unit variance, then the sum of the squares of a set of N samples,

$$\chi_N^2 = \sum_{n=1}^{N} y_n^2, \tag{1}$$

is a random variable, which follows a chi-square distribution with N degrees of freedom (Zelen and Severo 1972). Eq. (2) normalizes this sum to be an unbiased and consistent estimator of the block variance (Bendat and Piersol 2000),

$$\sigma^2 = \frac{\chi_N^2}{N-1} = \frac{1}{N-1}\sum_{n=1}^{N} y_n^2, \tag{2}$$

which also follows a chi-square distribution.

It is relatively easy to enforce the zero mean condition on acoustic data, either through high-pass filtering during data acquisition or mean subtraction in post-processing. However, the variance of the distribution for y is unknown, so a more general distribution is necessary to model the distribution of the block variance, σ^2. As a generalization of the chi-square distribution, the gamma distribution can be used (NIST 2013). The probability density function for a gamma distribution of the block variance (with a zero location parameter) is given by

$$p\left(\sigma^2\right) = \frac{1}{\beta\Gamma\left(\alpha\right)}\left(\frac{\sigma^2}{\beta}\right)^{\alpha-1} e^{\frac{-\sigma^2}{\beta}}, \tag{3}$$

where ν is the effective degrees of freedom discussed further below, $\alpha = \nu/2$ is the shape parameter required to relate a gamma distribution to the chi-square distribution for data of a given bandwidth, β is the scale parameter and Γ is the gamma function

$$\Gamma\left(\alpha\right) = \int_0^\infty t^{\alpha-1}e^{-t}\,\mathrm{d}t. \tag{4}$$

For $\beta = 2$, and substituting χ_N^2 for σ^2, this fully collapses to the chi-square distribution. This scale parameter allows a distribution fit to handle nonunity variance of y.

In practice, the acoustic signal is not truly random white noise, but has a finite bandwidth and correlation timescale. This normalized bandwidth, B, alters the effective degrees of freedom, ν, of the signal (Bendat and Piersol 2000). For example, a block of 8192 samples of a signal, which is truly random, has a spectrum of white noise and a bandwidth of 100%, so $\nu = N = 8192$. If the signal passes through an ideal lowpass filter set to 50% of the Nyquist frequency for the sampling rate, then $B = 0.5$ and the effective number of degrees of freedom is $\nu = B \times N = 4096$. This fractional, normalized bandwidth can be estimated through a simple procedure. First, the one-sided power spectral density of the signal must be computed. This function of frequency, $G_{yy}(f)$, must then be normalized such that its peak is unity,

$$G_{yy,\mathrm{norm}}(f) = \frac{G_{yy}(f)}{\max\left[G_{yy}(f)\right]}. \tag{5}$$

The average of this normalized spectral density is then computed by integrating across the measurement bandwidth and normalizing by the integration range,

$$B = \frac{1}{f_{\max}} \int_0^{f_{\max}} G_{yy,\mathrm{norm}}(f)\,\mathrm{d}f. \tag{6}$$

The normalized bandwidth can take a wide range of values. With the experimental data discussed later in this work, for example, it is found to be on the order of 10^{-3} for data acquired at a high sampling rate but dominated by low frequency spectral content.

With the effective degrees of freedom and, thus, the shape parameter of a distribution fit derived from the signal bandwidth, the scale parameter must now be determined. An easy, if biased (Zhang 2013), estimate of β can be obtained from its maximum likelihood estimator

$$\beta = \frac{\overline{\sigma^2}}{\alpha}, \tag{7}$$

where $\overline{\sigma^2}$ is an estimate of the mean of the block variances. However, the mean of the variances is sensitive to extreme variance values, which may occur when a transient event is superimposed on the baseline Gaussian process. A statistical parameter that is less sensitive to extreme values is necessary for computing β. One such parameter is the median of the block variances. The median occurs where the probability distribution function is 0.5. The probability distribution function for the gamma distribution is given by

$$P\left(\sigma^2\right) = \frac{\gamma\left(\alpha, \frac{\sigma^2}{\beta}\right)}{\Gamma\left(\alpha\right)}, \tag{8}$$

where γ is the (non-normalized) incomplete gamma function (NIST 2013)

$$\gamma\left(\alpha, \frac{\sigma^2}{\beta}\right) = \int_0^{\frac{\sigma^2}{\beta}} t^{\alpha-1} e^{-t} \, \mathrm{d}t. \tag{9}$$

The equation for the median variance is thus

$$\frac{1}{2} = \frac{\gamma\left(\alpha, \frac{\sigma^2_{\mathrm{med}}}{\beta}\right)}{\Gamma\left(\alpha\right)}. \tag{10}$$

Software libraries exist for efficiently inverting γ for a given α, thus yielding an estimate of the median variance normalized by β. The experimental median variance can then be divided by this estimate, yielding an estimate of β. Thus, for a given shape factor α, two scale factors can be readily computed from the data. One, β_{mean}, is based on the mean of the block variances and may be significantly influenced by extreme values of block variance in the data such as may be present with transient events. The other, β_{median}, is based on the median of the block variances. Note that for large sample sizes some simplifying mathematics are possible, but avoided here to allow for small numbers of short data blocks.

Having two scale factors allows for the construction of two gamma distributions. These can be compared to gain some sense of the relative influence of extreme block variances on the data set. Numerically this can be done by evaluating the Kullback-Leibler divergence, which is one metric for comparing distributions (Cardoso 1997). The divergence K is a measure of the information lost when probability distribution Q (or density q) is used to estimate distribution P (or density p). This is expressed as

$$K\left(p \| q\right) = \int \ln\left[\frac{p\left(\sigma^2\right)}{q\left(\sigma^2\right)}\right] p\left(\sigma^2\right) \, \mathrm{d}\left(\sigma^2\right). \tag{11}$$

While, in general, this can be difficult to compute, it is greatly simplified in the case of two gamma distributions with a common α. In this case, some manipulation yields

$$K\left(p \| q\right)_{\alpha_p=\alpha_q=\alpha} = \alpha\left(\ln\beta_q - \ln\beta_p + \frac{\beta_p - \beta_q}{\beta_q}\right), \tag{12}$$

or, as used in this application,

$$K\left(p_{median} \| p_{mean}\right)_{\alpha_{median}=\alpha_{mean}=\alpha} = \alpha\left(\ln\beta_{mean} - \ln\beta_{median} + \frac{\beta_{median} - \beta_{mean}}{\beta_{mean}}\right). \tag{13}$$

The two distributions match when K is zero.

To summarize, two data distributions can be estimated. The distribution based on the block variance mean is more sensitive to blocks with high variance, such as those containing transient contamination, than the distribution based on the block variance median. A metric is constructed for comparing the two distributions. Now a procedure is proposed for determining which blocks of a given time series to retain and which to reject. The process is illustrated in Fig. 1. It should be noted here that for the number of blocks traditionally used in aeroacoustic wind tunnel testing, converged data distributions are not expected. The intent of the following procedure is to provide an automated engineering tool to locate and thus exclude blocks in the time series associated with transient events, not to accurately estimate the probability distribution of the acoustic data block variance.

First, a given microphone time record is broken into blocks of a desired number of samples, N. This value is usually dictated by the desired spectral estimation parameters. The variance of each of these blocks is computed, and the blocks are sorted by their variance, from low to high. A minimum number of blocks is selected to automatically accept as stationary. This number of blocks is taken as the lowest-variance subset of blocks from the sorted set, and should be large enough to reduce the noise in the estimate but small enough to avoid any extreme values, or contaminated blocks. Experience with simulations suggests 20% of the total block count to be a safe selection, though a lower value was successfully used with experimental data. This subset of blocks is used to compute an autospectral density, which can be used to calculate α. This can be used to compute β_{mean} and β_{median}, followed by K. The next block, in order of ascending variance, is added to the active subset of blocks and the process is repeated. This continues until all of the blocks of data have been included, producing $|K|$ as a function of the number of blocks included in the data set in order of ascending variance. The block set yielding the minimum $|K|$ is classified as stationary. Blocks excluded from this set are classified as containing significant transient contamination. They may be subsequently excluded from processing of the stationary data of interest.

3 Simulated Analysis

A simulation study is performed to measure the performance of the transient classification procedure with data representative of experimental situations and

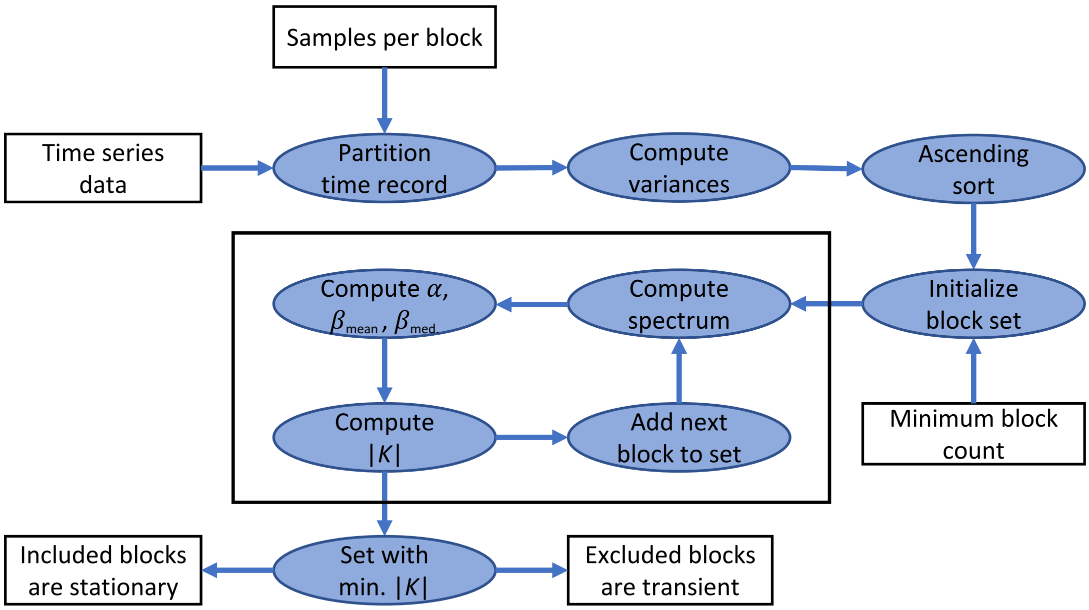


Fig. 1. Algorithm flow chart for classifying transient events.

parameter choices. The goal is to understand the performance of the procedure for a variety of situations and to gain an understanding of how the algorithm should perform for experimental data. Simulations are used as opposed to training data sets to better cover a complete range of possible situations.

3.1 Performance metrics

Identification of a data block contaminated with noise is a binary classification problem where the data block is either a transient, contaminated block or a stationary, uncontaminated block. Thus, performance metrics used to evaluate binary classifiers can be used here (Ting 2010). Note that for this study, classification of a data block as a transient, along with its subsequent rejection by the method and removal of the data block from the set of interest is considered as a positive result. The associated negative result is the classification of a data block as stationary. This study considers three performance metrics: accuracy, false positive rate, and false negative rate. The *accuracy* is the fraction of test cases that are correctly classified as either a transient data block or a stationary data block. The *false positive rate* is the fraction of the total number of stationary data blocks that are incorrectly classified as transient data blocks. It provides a measure of reduction in useful, stationary data blocks due to the classification process. The *false negative rate* is the fraction of the total number of transient blocks that are misclassified as stationary data blocks and provides a measure of the contaminated data blocks that are allowed through the algorithm.

An intermediate step for computing the accuracy, false positive rate, and false negative rate is the calculation of the confusion matrix. For a binary classification problem, the confusion matrix is a two by two table containing the counts of the classifier output for true positives and true negatives on the diagonal elements and false positives and false negatives on the off-diagonal elements. Thus, the accuracy is the sum of the diagonal elements divided by the total number of data blocks, while the false positive rate and false negative rate are the off-diagonal elements divided by the total number of true or known positives or negatives, respectively.

3.2 Simulation cases

The desired measured signal and the contamination signal are modeled as independent Gaussian noise signals with different variances, with the variance of the contamination larger than the variance of the desired signal. Five parameters are studied in simulations. These are the ratio of the variance of the contamination to the variance of the signal, the total number of data blocks, the number of points N in each data block, the percentage of the data blocks contaminated, and the percentage of the points in each data block that are contaminated. For all simulation cases, the total number of data blocks is swept through values of 100, 200, 300, 400, 500, and 1,000. The remaining parameters are given in Table 1. These combinations yield a total of 132 individual simulation cases.

3.3 Simulation procedure

The simulation procedure is as follows. First, a simulation case is selected, and the case parameters are noted. Next, the noncontaminated signal is modeled as a unit variance Gaussian random signal with the number of data points per data block and the number of blocks specified for the simulation case. Next, the clean signal is divided into the desired number of blocks, with no block overlap. Then, the desired number of blocks are contaminated for the desired percentage of points (selected as the first part of the block) with additive noise specified by the variance ratio and added to the the block. The transient classification algorithm is applied to the simulated data, and the data blocks classified as transients are logged. For these simulations, the transient classification procedure automatically considers the 20% of data blocks with the lowest variance to be stationary because lower total block

Table 1. Parameter values for simulation cases. All cases sweep through six values of the total number of data blocks of 100, 200, 300, 400, 500, and 1000.

Variance ratio	**Points per data block**	**Percentage of data blocks contaminated**	**Percentage of points in each data block contaminated**
2	8192	75	100
2	8192	50	100
2	8192	25	100
2	8192	25	50
2	8192	25	25
2	8192	75	50
2	8192	75	25
2	2048	75	25
2	2048	75	100
2	2048	25	25
2	2048	25	100
3	8192	25	25
5	2048	25	100
5	2048	75	100
5	2048	25	25
5	2048	75	25
5	8192	75	25
5	8192	75	100
5	8192	25	25
5	8192	25	100
10	8192	25	25
100	8192	25	25

counts approach the minimum necessary for a reasonable autospectral estimate. The confusion matrix elements are then calculated and recorded. The process is repeated for a total of 50,000 trials of data generation for each simulation case. The individual elements of the confusion matrix are examined to ensure the mean and standard deviation have converged to within 0.1% based on the values from one iteration to the next. Finally, the mean estimate for the confusion matrix is used to compute the estimated mean accuracy, false positive rate, and false negative rate for the simulation case.

3.4 Results

Table 2 presents a statistical summary of the three performance metrics over all of the simulation cases. The accuracy ranges from 80.1% to 99.3%. However, if the number of blocks is greater than or equal to 300, which is desirable for averaging of the spectral estimate as it approaches a normalized random error of 5%, the mean accuracy is greater than 90%. This condition also further constrains the false positive rate bounds to range from 0.9% to 12.9%, and the false negative rate bounds to range from 0.0% to 2.0%, improving on the results summarized in Table 2.

Table 2. Statistical summary of performance metrics for all simulation cases.

	Accuracy (%)	**FPR (%)**	**FNR (%)**
minimum	80.1	0.9	0.0
mean	94.4	8.8	0.3
median	97.0	6.2	0.01
maximum	99.3	26.4	4.2

To assess the behavior of the algorithm, and example plot of $|K|$ as a function of block count is shown in Fig. 2. Here a global minimum is observed when 749 of the blocks are retained, which is very close to the true number of 750 uncontaminated blocks. Note that the accuracy is less than that expected for 999/1000 correct classifications due to false positives and negatives, as discussed further below. Also note that some noise is present at extremely low block counts. This points back to the comment in Section 2 to set a minimum number of retained blocks.

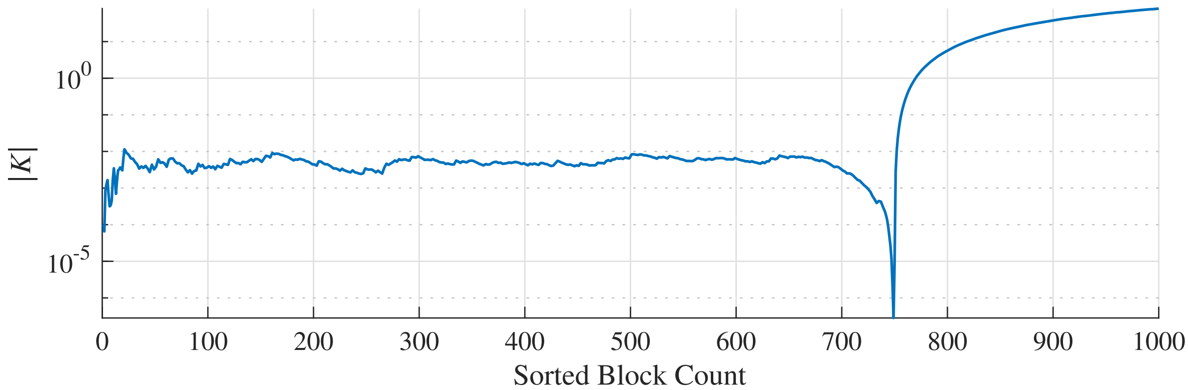


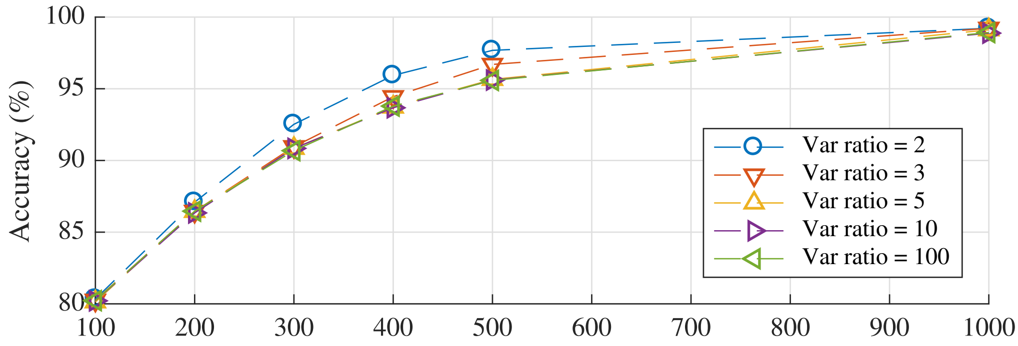
Fig. 2. Kullback-Leibler divergence as a function of included block count for a simulation with 1000 blocks of data, a variance ratio of 2, 8192 points per data block, 25% of the data blocks contaminated, and 50% of the points in each block contaminated.

3.4.1 Number of data blocks and variance ratio

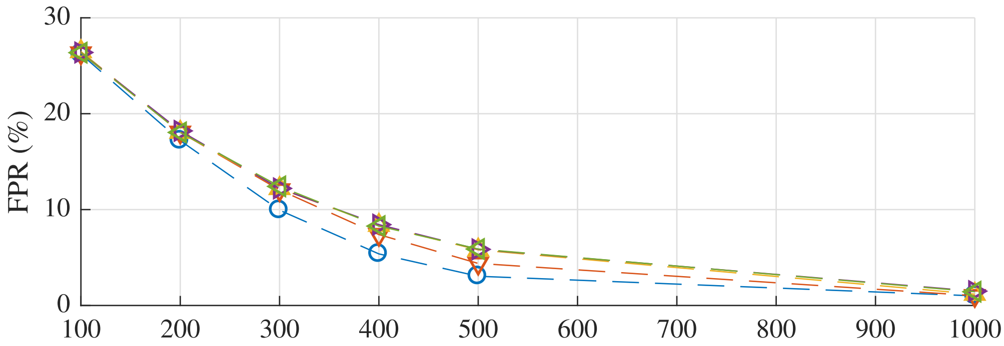
The variation in the performance of the algorithm is studied as a function of the total number of data blocks and contamination to the signal variance ratio. Here, the number of data points per block was held to $N = 8,192$ points, the percent of contaminated blocks to 25%, and the percent of each contaminated block perturbed to 25%. This resulted in 30 simulation scenarios selected from the 132 total cases. The results, as plotted in Fig. 3, show that all performance metrics converge as a function of variance ratio when the ratio is greater than five. The accuracy and the false positive rate improve as the total number of data blocks increases. The false negative rate shows more variation, but the values are below 0.14% for all 30 scenarios. These rates correspond to total false negative counts of zero, one, or, at worst, two misclassified blocks.

3.4.2 Percent of contaminated block perturbed

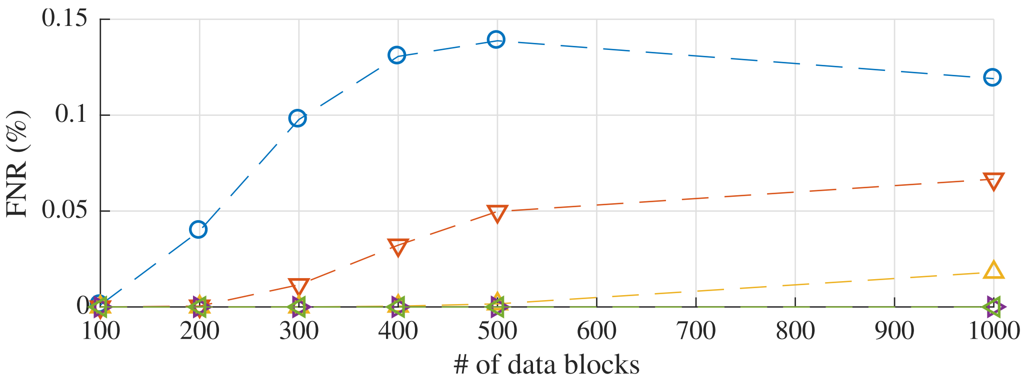
In the actual experiments analyzed in a subsequent section, transient gust contamination occurs sporadically and for short durations. Thus, for any data block that is impacted, only a portion of that block may be contaminated. Understanding how sensitive the performance metrics are to the percentage of any given data block that is perturbed is critical to assessing the robustness of the method. This simulation subset held the variance ratio to 2 (the most challenging value in the simulation study), the number of data points per block to $N = 8,192$ points, and the percentage of contaminated blocks to 25%. This resulted in 18 simulation scenarios selected from the 132 total cases. The results, as plotted in Fig. 4, show that the accuracy and the false positive rate are minimally affected by the percentage of the contaminated data block that is perturbed, especially when compared to the impact from the total number of data blocks. The magnitudes of the correlation coefficients between the accuracy and percentage of the data block contaminated, and between the false positive rate and the percentage of the data block contaminated are less than 0.1, confirming the lack of a linear relationship as seen in



(a) accuracy



(b) FPR



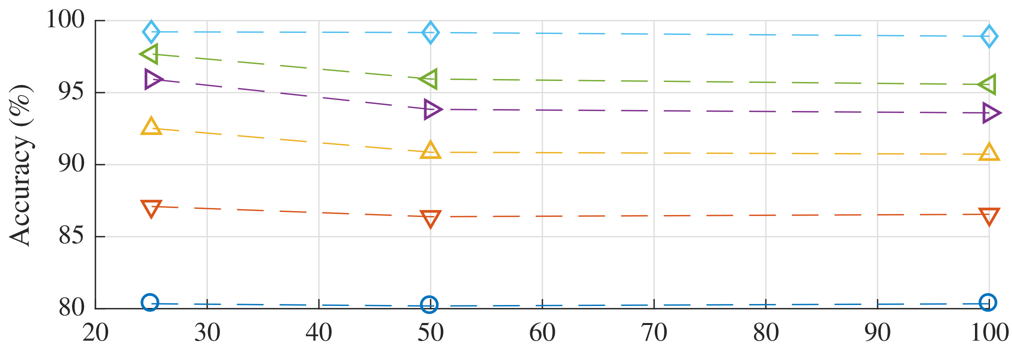
(c) FNR

Fig. 3. Performance metrics varying the total number of data blocks and the contamination to signal variance ratio. The number of data points per block is held to to $N = 8,192$ points, the percentage of contaminated blocks to 25%, and the percent of each contaminated data block perturbed to 25%.

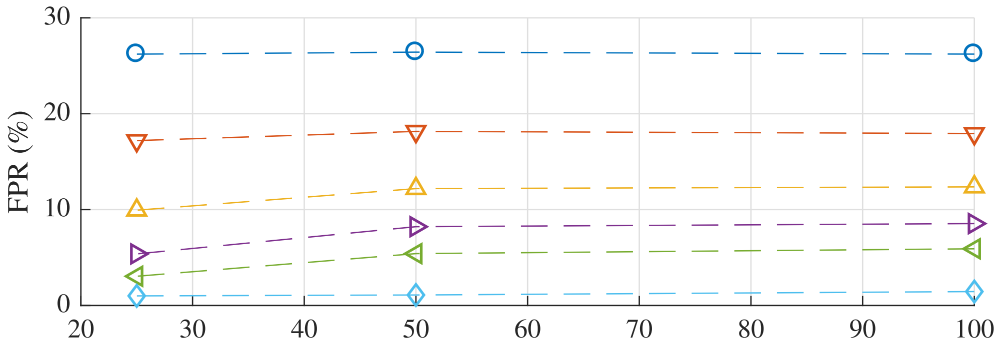
Fig. 4. However, the false negative rate does show a functional dependence on the percentage of the data block contaminated. This has a correlation coefficient of -0.25 (p-value of 0.004). Thus, as the percentage of the data block that is contaminated increases, the method can more easily identify data blocks that have been contaminated. However, the maximum false negative rate is still only 0.14%.

3.4.3 Percent of data blocks that are contaminated

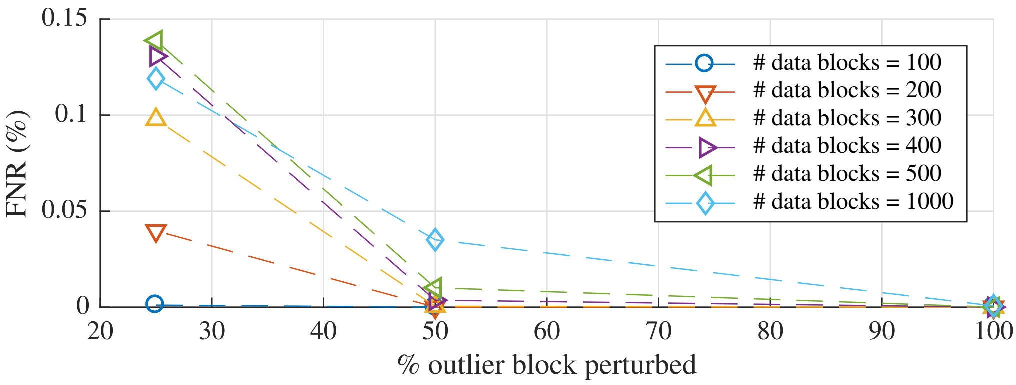
The variation in the performance of the classification algorithm is studied as a function of the percentage of data blocks that are contaminated. This simulation subset held the variance ratio to 2, the number of data points per block to $N = 8,192$ points, and the percent of each contaminated block perturbed to 25%, resulting in 18 simulation scenarios selected from the 132 total cases. The results, as plotted in Fig. 5, show that the accuracy and false positive rate improve with an increasing percentage of transient blocks in the total data set, whereas the false negative rate worsens. The values of all three performance metrics as a function of the percentage of contaminated blocks present in the total data set are also impacted by the total number of data blocks. However, when there is a total of 1,000 data blocks, the variation in the performance metrics as a function of the percentage of contaminated data present is minimal. With at least 300 total blocks, as might be recommended, the variation is greatly reduced. Note that a critical value of the percentage of contaminated blocks appears to exist between 50% and 75% where the behavior of the performance metrics changes.



(a) accuracy

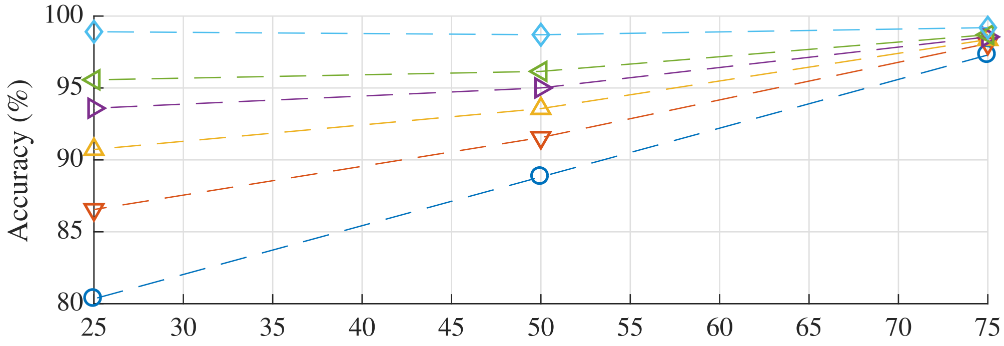


(b) FPR

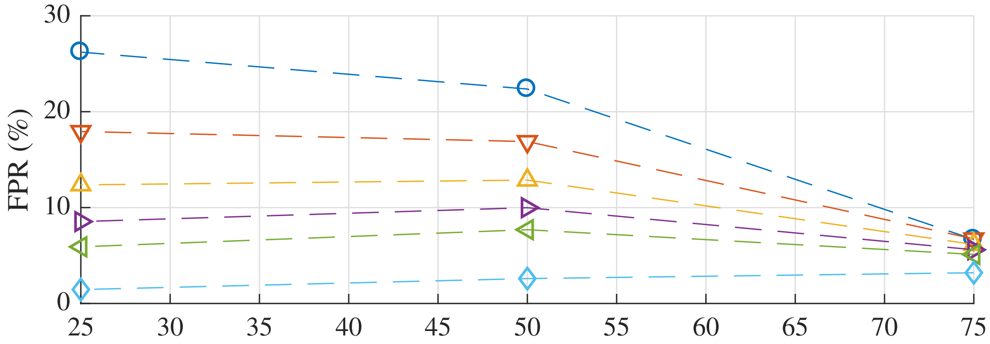


(c) FNR

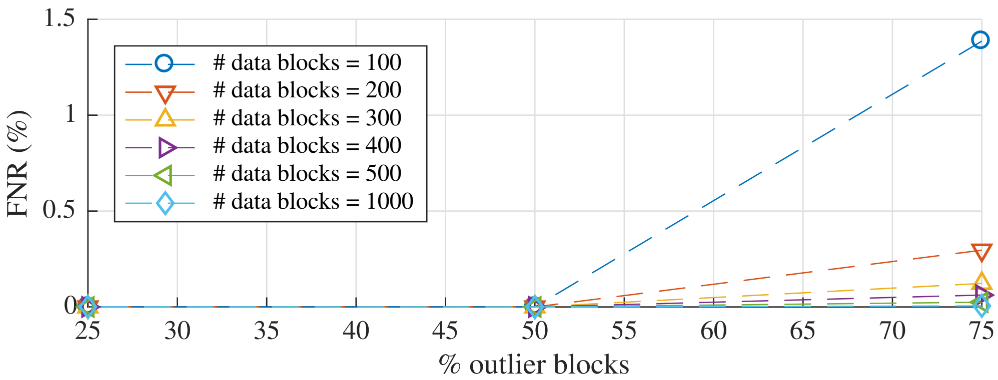
Fig. 4. Performance metrics varying the percentage of the contaminated block that is perturbed from the contamination signal while holding the variance ratio to 2, the number of data points per block to $N = 8,192$ points, and percentage of contaminated blocks to 25%.



(a) accuracy



(b) FPR



(c) FNR

Fig. 5. Performance metrics varying the percent of data blocks that are contaminated while holding the variance ratio to 2, the number of data points per block to $N = 8,192$ points, and percentage of the data points in the data block perturbed by contamination to 25%.

4 Experimental Results

The transient classification procedure is applied to an advanced aircraft noise study conducted at the NASA Langley Research Center's 14- by 22-Foot Subsonic Tunnel (Heath et al. 2016). A photograph of an example test configuration from this study is shown in Fig. 6, where a hybrid wing body model is installed inverted in the facility test section. As shown in the photograph, microphones are installed on sideline traversing towers, as well as a truss and array panel located above the facility test section.

The NASA Langley 14- by 22-Foot Subsonic Tunnel is, by design, an aerodynamic wind tunnel, which can operate in an open test section configuration. While significant acoustic improvements have been applied to the facility, measurement microphones are, under some installation configurations, close enough to the open-jet shear layer that hydrodynamic gusts may contaminate the out-of-flow acoustic measurements. This was primarily observed when microphones were at the far-downstream end of the test section, although occasional gust impingement was seen at other measurement stations.

An extreme example of gust impingement from the airframe noise component of the test is shown in Fig. 7. The plotted data are for an acquisition where one of the speakers embedded in the model body was driven with a random noise signal that was bandpass filtered to span a frequency range of 4 kHz to 16 kHz. The speaker data are used rather than the model's isolated airframe noise data as, in the frequency domain, these data provide a more clear visual representation of the influence of transient contamination over a limited bandwidth due to more distinct spectral structures.

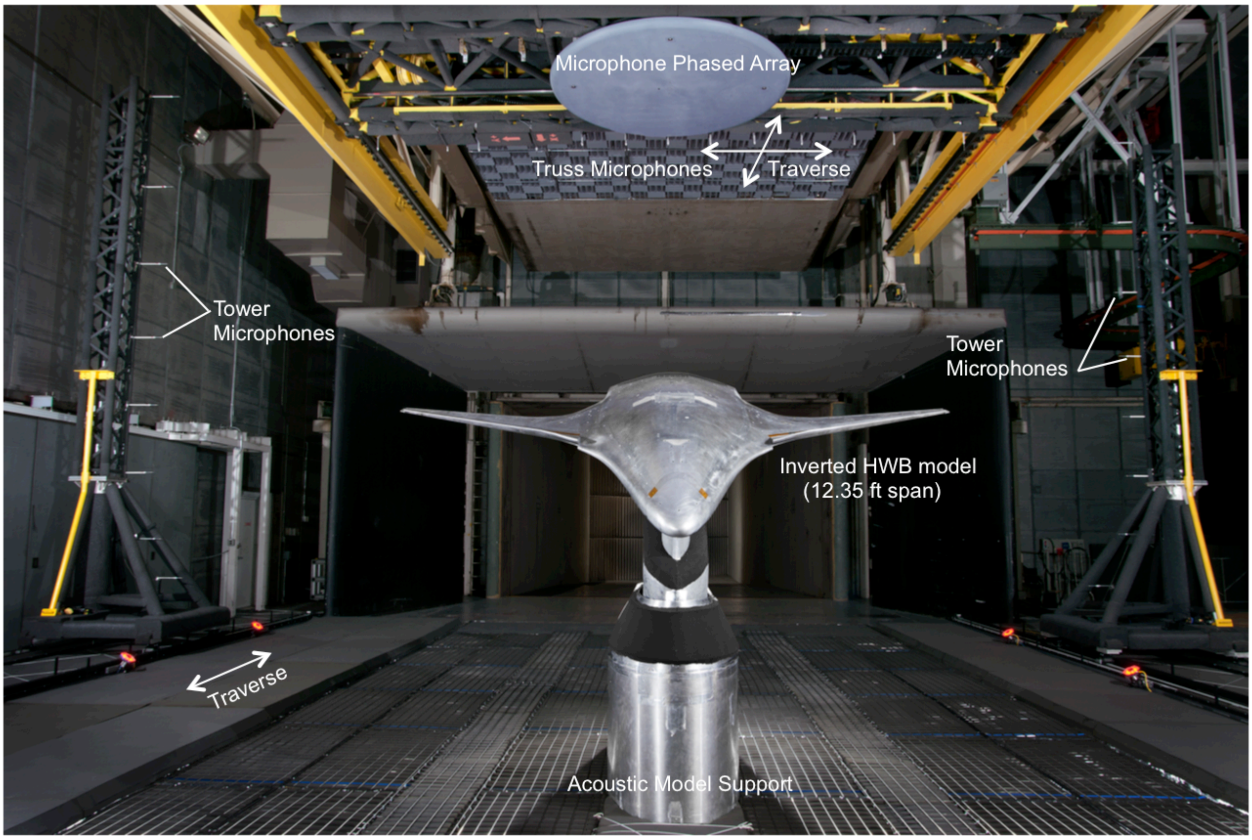


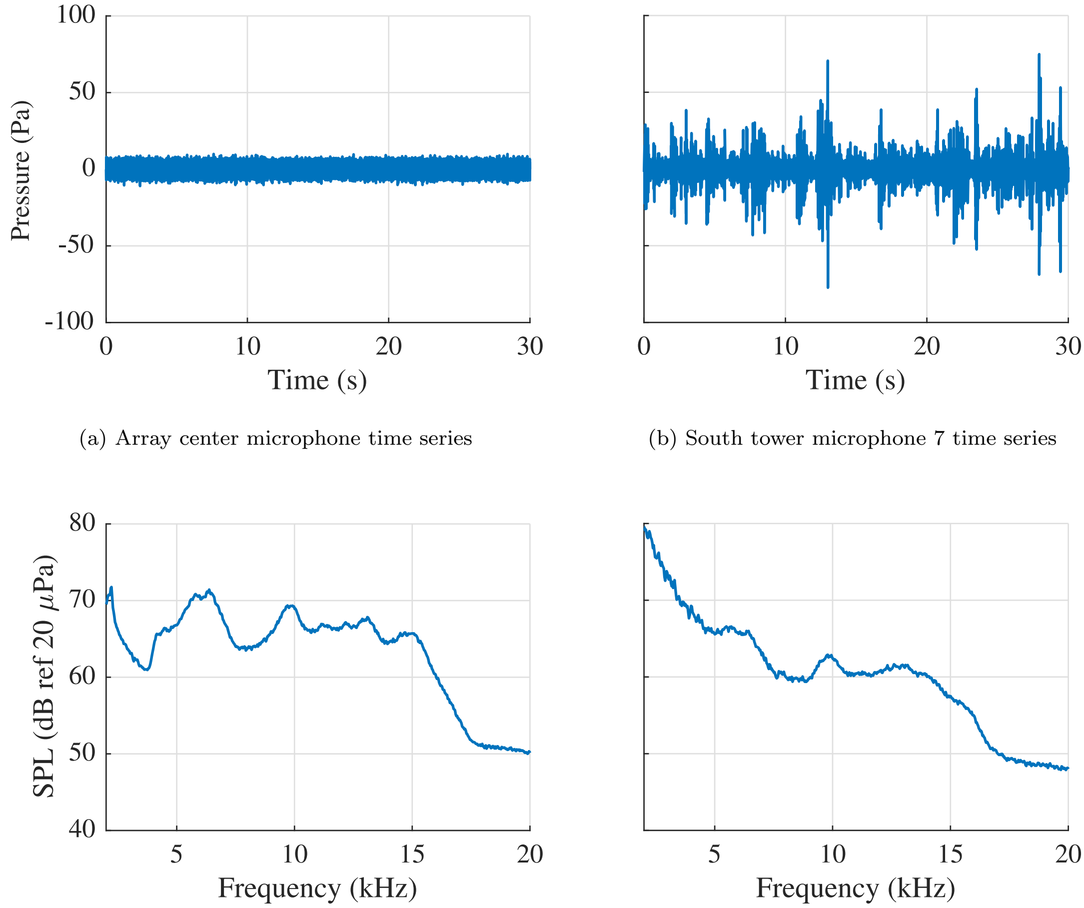
Fig. 6. Example arrangement of a hybrid wing body model, phased array and tower traverses installed in the NASA Langley 14- by 22-Foot Subsonic Tunnel.

The hybrid wing body model was pitched to an angle of attack of 14.5°, and the test section Mach number was $M = 0.23$. The acoustic measurement hardware was traversed to the far-downstream end of the test section. As shown by the time series in Fig. 7a, the array center microphone signal appears as might be expected for a stationary, band-limited random signal. The south tower microphone, located in the upper-right-hand corner of the picture in Fig. 6, clearly experiences extreme transient bursts as shown in Fig. 7b. The corresponding autospectra are shown in Figs. 7c and 7d. While the array center microphone spectrum shows the low frequency content of the signal at 4 kHz, the south tower microphone spectrum is masked by the low frequency content of the burst. Note that at this stage of processing, two clean signals would not overlay due to differences in propagation distance between the source and each microphone, along with the speaker directivity. Also, this test is a prime example of why an automated classification method is desirable. The contamination in the data is clear and could readily be separated manually. However, roughly a quarter of a million time series records were generated during the test. Manual inspection of such a volume of data is unreasonable.

For these data, the procedure developed for transient classification is applied by breaking the microphone time series into 920 blocks of desired length $N = 8192$ points. This corresponds to the baseline processing parameters used in the test for spectral analysis (Bahr et al. 2014). The minimum number of accepted blocks is set to 100 based on observation of the spectral convergence. A histogram of the south tower microphone data block variances is shown with respect to the left axis in Fig. 8, with the 16 most energetic blocks removed from the plot. Even without these blocks, which would extend the plot abscissa beyond a variance of 500 Pa^2, this histogram shows a long, thin tail in the direction of large variance values. The corresponding probability density functions for the median- and mean-based models are shown with respect to the right axis in the figure. Note that the bandwidth parameter B for these data is extremely low, generally between 4×10^{-3} and 5×10^{-3} depending on the included blocks. This is due to the high levels of low frequency data, below the speaker operating range.

Of the 920 input blocks, 567 are rejected. The computed $|K|$ as a function of block count used to separate the blocks is plotted in Fig. 9, showing an obvious minimum as it did with the simulated data in the previous section. The almost-monotone increase at higher block counts is similar in behavior to the $|K|$ plot for simulated data, and appears to be associated with data convergence. The series of local minima near the global minimum are not easily explained, as the baseline variances, variance means, and variance medians are all reasonably smooth. It is only when differences and ratios of these parameters are computed that more jagged features appear.

It should be noted that while 567/920 blocks is a large portion of the data to reject, this microphone acquisition is from a location normally outside of the bounds of reasonable acoustic measurement positions in the facility. The histogram of the remaining block variances is shown in Fig. 10, along with the median- and mean-based probability density function estimates for the retained block set. As expected, the probability density functions overlay for the minimum value of $|K|$, indicating near-total agreement between the mean and median models and that all the data blocks provide useful information to the statistics. The outliers, as modeled, have been eliminated. The output of the procedure is shown in Figs. 11a and 11b. Visually, the technique has identified and removed the obvious contamination from the time series. In the spectral analysis, the 4 kHz content of the signal is now visible, with a reduction of up to 10 dB in the microphone autospectrum at lower frequencies. Higher frequencies are unaffected.



(a) Array center microphone time series

(b) South tower microphone 7 time series

(c) Array center microphone autospectrum

(d) South tower microphone 7 autospectrum

Fig. 7. Example data contamination by hydrodynamic impingement. The two compared microphones observed a calibration signal with an output band of 4 kHz to 16 kHz, emitted by one of the model embedded speakers. The hybrid wing body model was at an angle of attack of 14.5°, and the test section Mach number was $M = 0.23$. Acoustic hardware were at the far downstream traverse location. Spectral binwidths are 30.5 Hz.

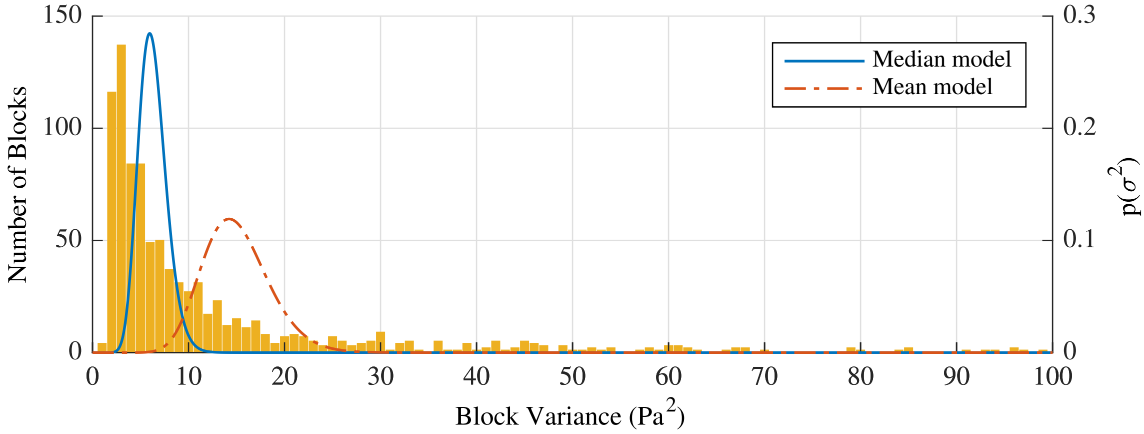


Fig. 8. Histogram of block variances from the south tower time series in Fig. 7 excluding the 16 most energetic blocks, and modeled data probability density functions.

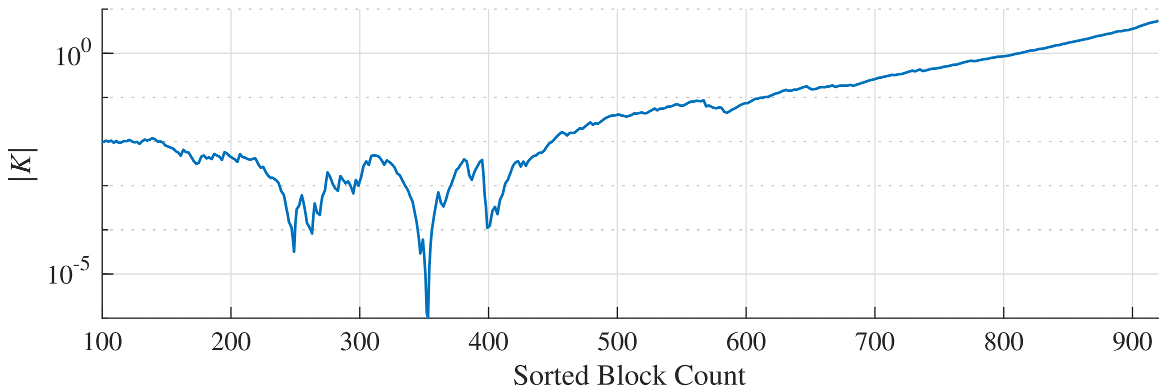


Fig. 9. Kullback-Leibler divergence as a function of included block count for the south tower time series data.

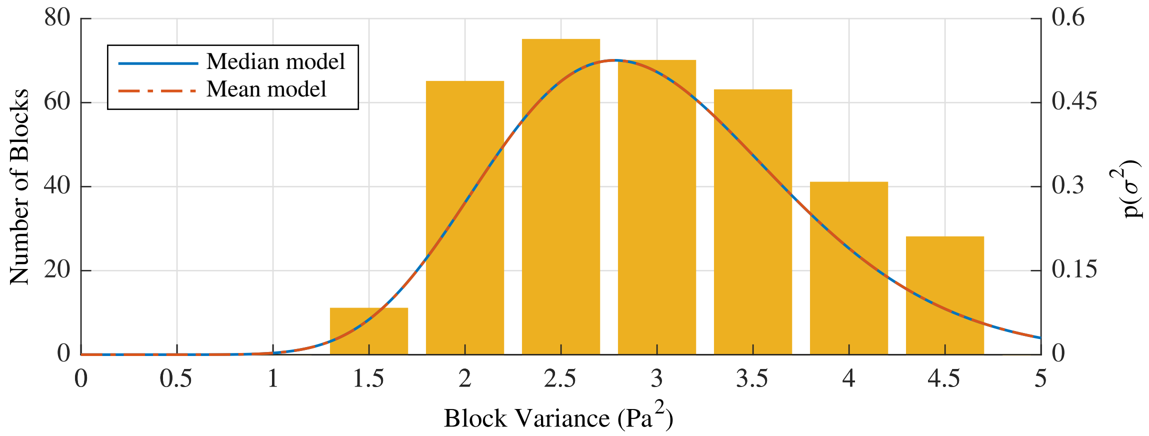
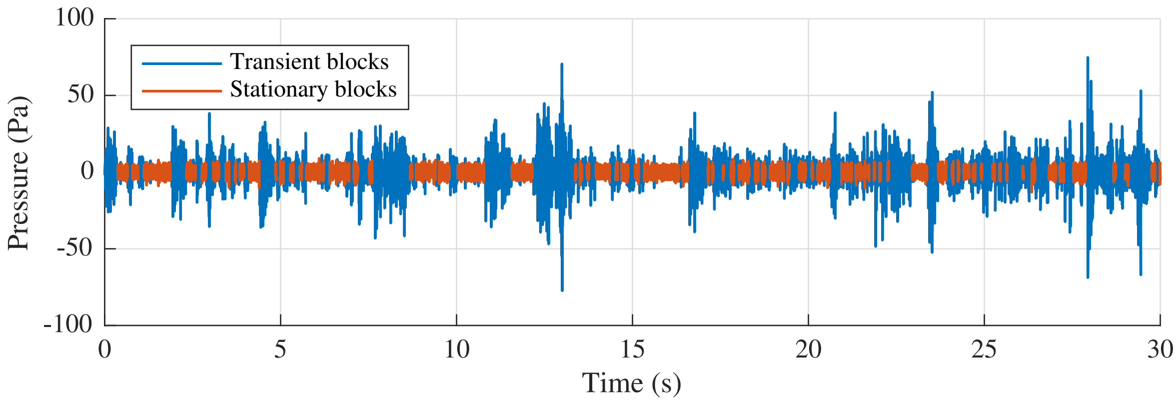
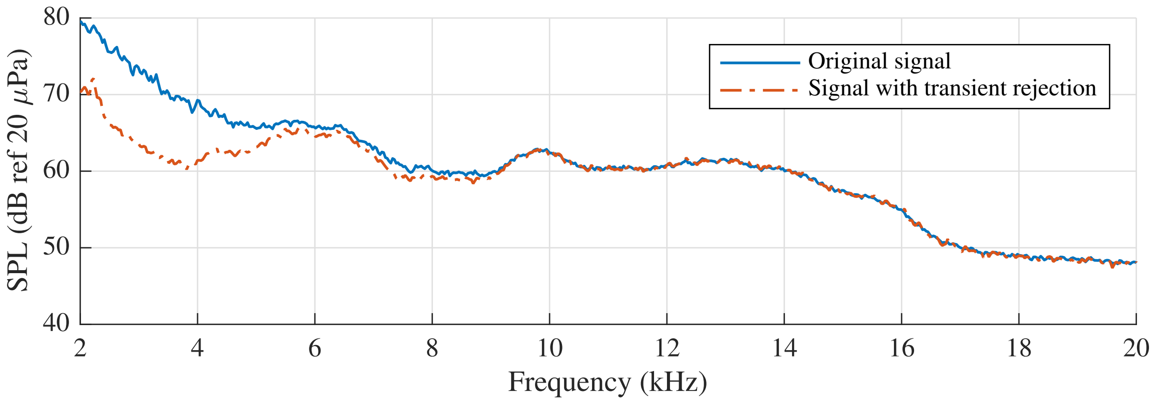


Fig. 10. Post-classification histogram of south tower time series data from Fig. 8, along with post-rejection models (mean model almost completely overlays the median model).



(a) Blocks accepted and rejected by the algorithm



(b) Effect of transient block rejection on microphone autospectrum

Fig. 11. Results of transient rejection algorithm when applied to the south tower time series data from Fig. 7b. Data blocks are plotted as a function of time. The shift in the estimated data autospectrum is shown. Spectral binwidths are 30.5 Hz.

5 Summary & Conclusions

An automated method for classifying transient data segments that contaminate stationary acoustic data is presented. The method requires two assumptions. First, it treats the underlying stationary signal of interest as having Gaussian random characteristics. Second, it assumes that contaminated segments of data will have higher variance than clean segments of data. Under these assumptions, it is an unsupervised method that performs binary classification: either a data block is contaminated by a transient signal or it is clean.

An extensive set of simulations covering a broad range of conditions shows that the technique has a high degree of accuracy as long as at least 300 data blocks are used, though 500 may be preferable. The FPR may still be greater than 5% under some of the simulated circumstances. However, falsely classifying a few blocks of stationary data as transient and discarding them is not problematic. Wind tunnel time is expensive, so data records have a practical duration limit based on cost. Regardless, standard spectral estimation techniques will still perform well if a few extra blocks are discarded while hundreds are retained. Simulations suggest the technique has a very low FNR for the parameter space explored, so misclassifying enough transient data as stationary to noticeably contaminate a spectral estimate is unlikely.

Experimental results from a worst-case scenario in an aeroacoustic wind tunnel test show that, visually, the method succeeds in separating contaminated blocks from the baseline signal of interest. Spectral estimation of the signal both before and after the application of the technique shows up to a 10 dB improvement in signal-to-noise ratio due to the removal of contamination. Features in the acoustic spectrum that are masked in the baseline data set are revealed once the transient blocks are removed.

Acknowledgements The authors would like to acknowledge the support provided by the 14- by 22- Foot Subsonic Tunnel team, by colleagues in the Aeroacoustics and Advanced Sensing & Optical Measurements branches at the NASA Langley Research Center, and by colleagues in the Aerodynamics, Noise, and Propulsion Laboratory in the Boeing Test & Evaluation organization at The Boeing Company. The hybrid wing body test was funded by the NASA Environmentally Responsible Aviation Project.

References

Aggarwal CC (2017) Outlier Analysis, 2nd edn. Springer, New York, NY

Bahr CJ, Horne WC (2017) Subspace-based background subtraction applied to aeroacoustic wind tunnel testing. International Journal of Aeroacoustics 16(4-5):299–325

Bahr CJ, Brooks TF, Humphreys WM, Spalt TB, Stead DJ (2014) Acoustic Data Processing and Transient Signal Analysis for the Hybrid Wing Body 14- by 22-Foot Subsonic Wind Tunnel Test. 20$^{\text{th}}$ AIAA/CEAS Aeroacoustics Conference, AIAA Aviation 2014, Atlanta, GA, 2014, AIAA 2014-2345

Bendat JS, Piersol AG (2000) Random Data Analysis and Measurement Procedures. John Wiley & Sons, Inc., New York, NY

Cardoso J (1997) Infomax and maximum likelihood for blind source separation. IEEE Signal Processing Letters 4(4):112–114

Coleman HW, Steele WG (1999) Experimentation and Uncertainty Analysis for Engineers, 2nd edn. John Wiley & Sons, Inc., New York, NY

Hawkins D (1980) Identification of Outliers. Springer, New York, NY

Heath SL, Brooks TF, Hutcheson FV, Doty MJ, Bahr CJ, Hoad D, Becker LE, Humphreys WM, Burley CL, Stead DJ, Pope DS, Spalt TB, Kuchta DH, Plassman GE, Moen JA (2016) NASA Hybrid Wing Body Aircraft Aeroacoustic Test Documentation Report. Tech. Rep. NASA TM-2016-219185

Humphreys WM, Brooks TF, Hunter WW, Meadows KR (1998) Design and Use of Microphone Directional Arrays for Aeroacoustic Measurements. 36^{th} AIAA Aerospace Sciences Meeting & Exhibit, Reno, NV, 1998, AIAA-98-0471

NIST (2013) NIST/SEMATECH e-Handbook of Statistical Methods, chap 1.3.6.6.11. Gamma Distribution. URL `http://www.itl.nist.gov/div898/handbook/`

Soderman PT, Allen CS (2002) Microphone measurements in and out of airstream. In: Mueller TJ (ed) Aeroacoustic Measurements, Springer-Verlag, Berlin, Heidelberg, New York, chap 1, pp 22–24

Ting KM (2010) Confusion matrix. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning, Springer US, Boston, MA, p 209

Zelen M, Severo NC (1972) Probability functions. In: Abramowitz M, Stegun IA (eds) Handbook of Mathematical Functions, Dover Publications, Inc., New York, chap 26, pp 940–943

Zhang J (2013) Reducing bias of the maximum likelihood estimator of shape parameter for the gamma distribution. Computational Statistics 28(4):1715–1724