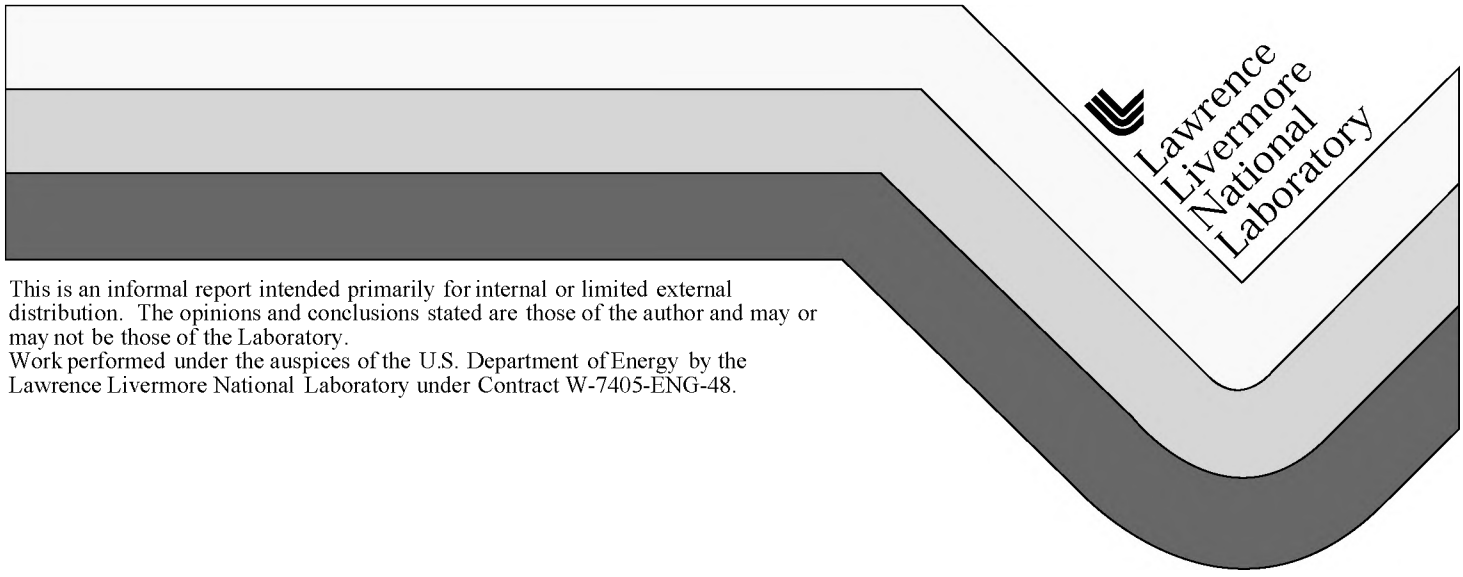
UCRL-ID-134634 Rev 1

Classification of Heart Valve Sounds From Experiments in an Anechoic Water Tank

Gregory A. Clark
Michael C. Axelrod
David D. Scott

June 1, 1999



Executive Summary

In vivo studies in both sheep and humans were plagued by a number of problems including movement artifacts, biological noise, low signal-to-noise ratio (SNR), chest-wall reverberation, and limited bandwidth recordings as discussed by [1]. To overcome these problems it was decided to record heart valve sounds under controlled conditions deep in an anechoic water tank, free from reverberation noise, including surface reflections. Experiments were conducted in a deep water tank at the Transdec facility in San Diego, which satisfies these requirements. The Transdec measurements are free of reverberations, but not totally free of acoustic and electrical noise. We used a high quality hydrophone together with a wide-band data acquisition system [2]. We recorded sounds from 100 repetitions of the opening-closing cycles on each of 50 different heart valves, including 21 SLS valves and 29 intact valves. The power spectrum of the opening and closing phases of each cycle were calculated and outlier spectra removed as described by Candy [2].

In this report, we discuss the results of our classification of the heart valve sound measurements. The goal of this classification task was to apply the fundamental classification algorithms developed for the clinical data in 1994 and 1996 to the measurements from the anechoic water tank. From the beginning of this project, LLNL's responsibility has been to process and classify the heart valve opening sounds. For this experiment, however, we processed both the opening sounds and closing sounds for comparison purposes. The results of this experiment show that the classifier did not perform well. We believe this is because of low signal-to-noise ratio and excessive variability in signal power from beat-to-beat for a given valve. The results of the classification work is summarized as follows:

- For the opening sounds, the classifier failed to classify better than chance.
- Noise canceling applied to the opening sounds resulted in an increase in the estimate of the probability of correct classification to 57.8%. However, this improvement is clearly not significant enough to recommend the classification of opening sounds for use with clinical data.
- For the closing sounds, the probability of correct classification was 83%. We believe that the closing sounds worked better than the opening sounds for this experiment because the closing sounds have a better SNR (signal-to-noise ratio).

Several issues having to do with the experiments make classification difficult.

- We observe excessive beat-to-beat variation in the sound signal energy for both opening and closing sounds. The valves are apparently not being excited consistently and with enough force, leading to excessive SNR.

- We observe excessive valve-to-valve variation in the signal sound energy for both opening and closing sounds. We cannot be sure what causes this variation. Possible causes could include variability in the physical characteristics of the valves, inconsistent excitations, and environmental perturbations.

The resolution of these data issues and the improvement of classification results involves three main recommended approaches: (1) Improvements in experiment design to obtain increased SNR and reduced variability. (2) Signal processing R&D to use a nonstationary signal model rather than a stationary one. The signal is transient and therefore nonstationary, so new features using this model could result in significant benefit. (3) Classification algorithm R&D, including work in advanced feature analysis.

Table of Contents

Introduction 6

Classification Results 9

Appendix I: Data Acquisition and Signal Extraction 19

Appendix II: Signal Processing 21

Appendix III: Feature Analysis and Classification 22

Appendix IV: Analysis of Sound Data Variability 41

References 49

Introduction

Between 1979 and 1986 about 86,000 patients worldwide received Bjork-Shiley prosthetic heart valve implants using the Convexo-Concave (BSCC) tilting disc design. As illustrated in Fig. 1, the valve consists of a metal flange and a free-floating polymer occluder disc held between two metal struts. The outlet strut is welded to the flange in contrast to the inlet strut which is integral to it. By opening and closing, the disc alternately allows or restricts blood flow. The weak point in this design is fatigue failure at the outlet-strut weld. In a small number of valves the outlet strut fractures, resulting in a condition known as *single-leg separation* (SLS). By 1990 at least 600 failures occurred resulting in at least 400 deaths The actual failure rate is likely much larger [3].

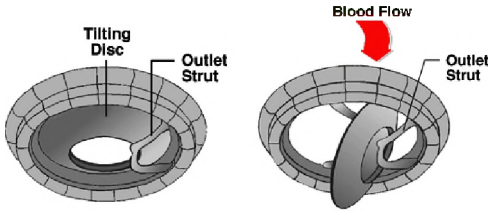


Fig. 1. Bjork-Shiley Convexo-Concave (BSCC)
Prosthetic Heart Valve: Disc Occluder (tilted) and Outlet Strut

Because surgical replacement of heart valves is inherently risky, there is considerable interest in developing minimally invasive techniques to identify reliably those patients with SLS valves. Research has concentrated on two techniques: radiographic imaging and acoustical signal analysis of heart valve sounds. Radiography has drawbacks. First, it requires exposure of the patient to ionizing radiation; Second, it is extremely difficult to achieve sufficient image resolution to detect the separated leg. We believe the acoustic approach holds promise because it seems reasonable to expect that an SLS valve would emit a different sound than an intact valve when the disc occluder strikes the struts on opening or closing. One analogy (albeit imperfect) is to a cracked bell. A bell with a crack would have different vibrational modes than a similar bell without a crack. Detection would follow from a frequency analysis of the bell's ring. This idea has motivated us to study the frequency spectra of opening and closing valve sounds, and to find a feature or set of spectral features that would discriminate intact from SLS valves.

In vivo studies in both sheep and humans were plagued by a number of problems including movement artifacts, biological noise, low signal-to-noise ratio (SNR), chest-wall reverberation,

and limited bandwidth recordings as discussed by [1]. To overcome these problems it was decided to record heart valve sounds under controlled conditions deep in an anechoic water tank, free from reverberation noise. The main goal of this experiment was to obtain measurements of "pure" heart valve sounds free of the scattering effects of the body. We used a high quality hydrophone together with a wide-band data acquisition system [2]. We recorded sounds from 100 repetitions of the opening-closing cycles on each of 50 different heart valves, including both SLS and intact types. The power spectrum of the opening and closing phases of each cycle were calculated and outlier spectra removed as described by Candy [2].

In this report we discuss the results of our classification analysis of the heart valve sound measurements from the anechoic water tank at Transdec in San Diego [2]. Our overall approach is depicted in block diagram form in Fig. 2. The *data acquisition* step yields time series data containing opening and closing valve sounds as well as noise and other transient events. The *signal extraction* step separates the opening from the closing valve sounds. The *signal processing* step yields an ensemble of closing and opening spectra for each valve. The *feature extraction* step transforms these spectra into useful features for the classifier. The most salient features are chosen during the *feature selection* step and this parsimonious set of features is use by the *classifier*. The classifier can then identify a new valve as either intact or SLS. The performance of the classifier is then assessed by calculating the rate of correct classification we would expect on a new set of valves of unknown condition.

The data acquisition, signal extraction and signal processing steps were discussed in a another report [2]. This report focuses on the feature extraction, feature selection and classification steps. We analyze the experimental variability and interpret the results. The appendices contain discussions of the theory behind our calculations.

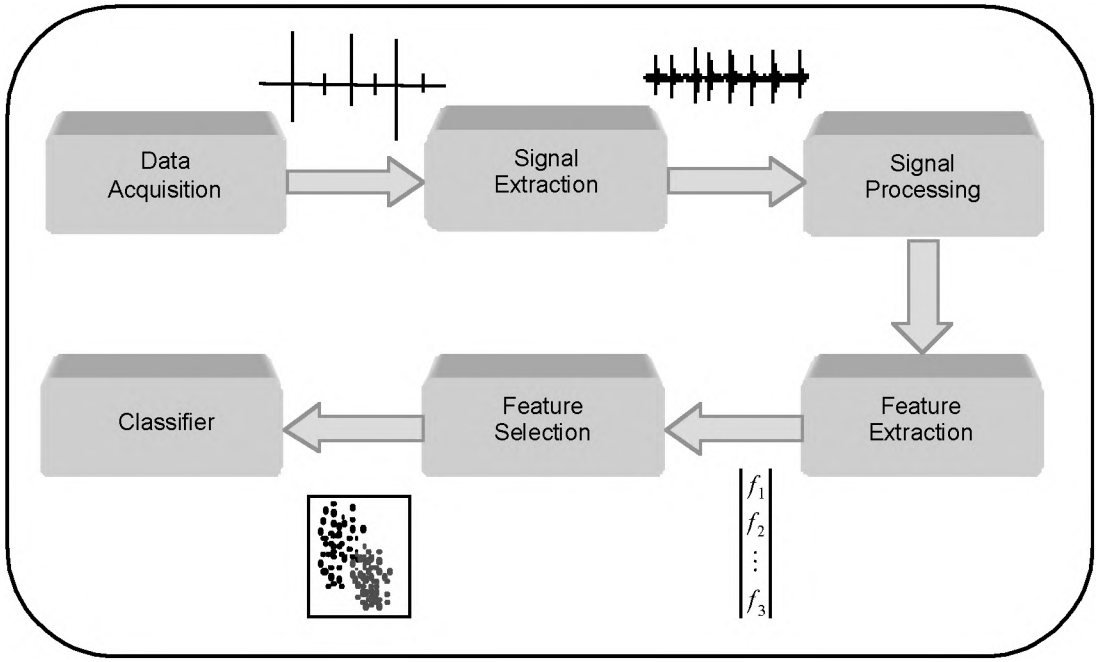


Figure 2 Heart valve classification protocol block diagram.

Classification Results

The measurement data for the classification study included the following:

- 21 SLS valves
- 29 Intact valves
- 100 signal pairs (beats with openings and closings) per valve
- 512 samples (opening)
- 1024 samples (closing)
- 50,000 signal pairs
- 7.68 million samples

The goal of classification task was to apply the fundamental classification algorithms developed for the clinical data in 1994 and 1996 to the measurements from the anechoic water tank. The goal of this project was not to perform extensive algorithm research. The fundamental algorithms developed for the clinical data are general in the sense that they can be used for classifying patterns in general signals. However, for any particular application, the algorithm parameters and the preprocessing steps applied to the data must be tuned for optimal performance. This tuning process can involve extensive trials to determine the optimal parameter settings. We performed such parameter studies and the results are presented below.

From the beginning of this project, LLNL's responsibility has been to process and classify the heart valve opening sounds. Closing sounds have been processed by other organizations. The opening sounds have the advantage that they represent the vibration of the outlet strut when it is struck by the valve disk, so they should contain information about whether or not the strut is intact or SLS. The disadvantage of opening sounds is that they have low amplitude and signal-to-noise ratio (SNR) relative to the much louder closing sounds. This low SNR makes processing and classifying difficult. The closing sounds have the advantage that they have high SNR, but the disadvantage that they include a very large contribution from the vibration of the valve ring that can mask the relatively smaller sounds from the outlet strut.

One result of this experiment is the discovery that the signal-to-noise-ratio (SNR) of the opening sounds is much lower than expected. This is manifested as excessive variability in signal energy

from beat-to-beat for a given valve. We discuss possible reasons for this result later in this report.

Experiments with algorithm parameters

The classifier was configured with the same algorithms that were used successfully during the 1996 blind test of clinical data (see Appendix III). The classifier was trained and the data were tested using the cross-validation method as discussed in Appendix III. We processed the data for classification purposes using various combinations of data types and processing parameters.

Opening Sounds

For the openings, we tested the following cases/combinations (see Appendix III):

- Width of the frequency bands used for calculating power features: 1,2,3,4,5 (frequency bins).
- Distance metric: Mahalanobis, Bhattacharyya, Kullback-Liebler, Jeffreys
- Spectral estimator: MEM, MVDR
- Number of features used: 2,3 and 5

Conclusions: The opening sounds have poor signal-to-noise ratio (SNR) and a large amount of beat-to-beat variation, which leads to very limited ability to classify the valves. We settled on using the Mahalanobis distance with 3 features.

For the opening sounds, the classifier failed to classify better than chance. The probability of correct classification (see Appendix III) was only 43%. We have not included performance plots for the opening sounds in this report, because they have little or no meaning. We present an analysis of the low SNR and large beat-to-beat variation later in this report. Given the poor classifier performance for opening sounds, we experimented with the closing sounds to determine their classification potential.

Closing sounds

For the closing sounds, we tested the following cases/combinations (see Appendix III):

- Width of the frequency bands used for calculating power features: 1,2,3,4,5 (frequency bins).
- Distance metric: Mahalanobis
- Spectral estimator: MEM
- Number of features used: 3

Conclusions: The closing sounds, as expected, have significantly greater SNR, and provide a greatly increased ability to classify the valves. We settled on using the Mahalanobis distance with 3 features.

The Probability of correct classification for closing sounds was 83%. The performance results for the closing sounds are shown in Fig. 3. The following discussion describes the performance plots and results depicted in Fig. 3.

Fig. 3 Confusion Matrix

The confusion matrix (see Appendix III) shows the performance summary of the classifier:

- Of the 29 intact valves, 25 were classified intact, giving specificity = P(INT | INT) = 86.2%.
- Of the 29 intact valves, 4 were classified SLS, giving P(false alarm) = P(SLS | INT) = 13.8%.
- Of the 21 SLS valves, 16 were classified SLS, giving P(Detection) = P(SLS | SLS) = 80%.
- Of the 21 SLS valves, 4 were classified INT, giving P(Miss) = P(INT | SLS) = 20%.
- We see that the probability of correct classification = .5[Sensitivity + Specificity] = 83%.

Fig. 3 Plots of the average SLS and INT spectra

Based upon the number of valves available, we asked the feature selector to choose the best 3 spectral features of the "sliding window" type form the set of features extracted (see Appendix

III). We see that the features chosen by the sequential forward selection algorithm are the following:

Feature 1 = Power in the frequency band of width 195.31 KHz centered at 25.39 KHz .

Feature 2 = Power in the frequency band of width 390.62 KHz centered at 96.39 KHz .

Feature 3 = Power in the frequency band of width 390.62 KHz centered at 39.55 KHz .

Fig. 3 Receiver Operating Characteristic (ROC) Curve

The ROC shows probability of detection plotted vs. probability of false alarm. The ROC curve is calculated by varying the decision threshold in the second stage of the two-stage classifier (see Appendix III). We choose the operating point on the curve to be the point for which the threshold = .5. We also display another representation of the ROC curve in which the probability of detection and probability of false alarm are plotted versus the threshold on the posterior probability P(SLS |X), where X is the feature vector (see Appendix III). The two representations of the ROC curve are equivalent.

Fig. 3 Plot of the Average Posterior Probability P(SLS| X) vs. Valve Index

The average posterior probability P(SLS | X) is plotted vs. the valve index (valve number, from 1 to 50). The fixed threshold on the posterior probability is .5 and is plotted on the figure. Valves with posterior probability greater than or equal to .5 were classified SLS and those with posterior probability less than .5 were classified intact.

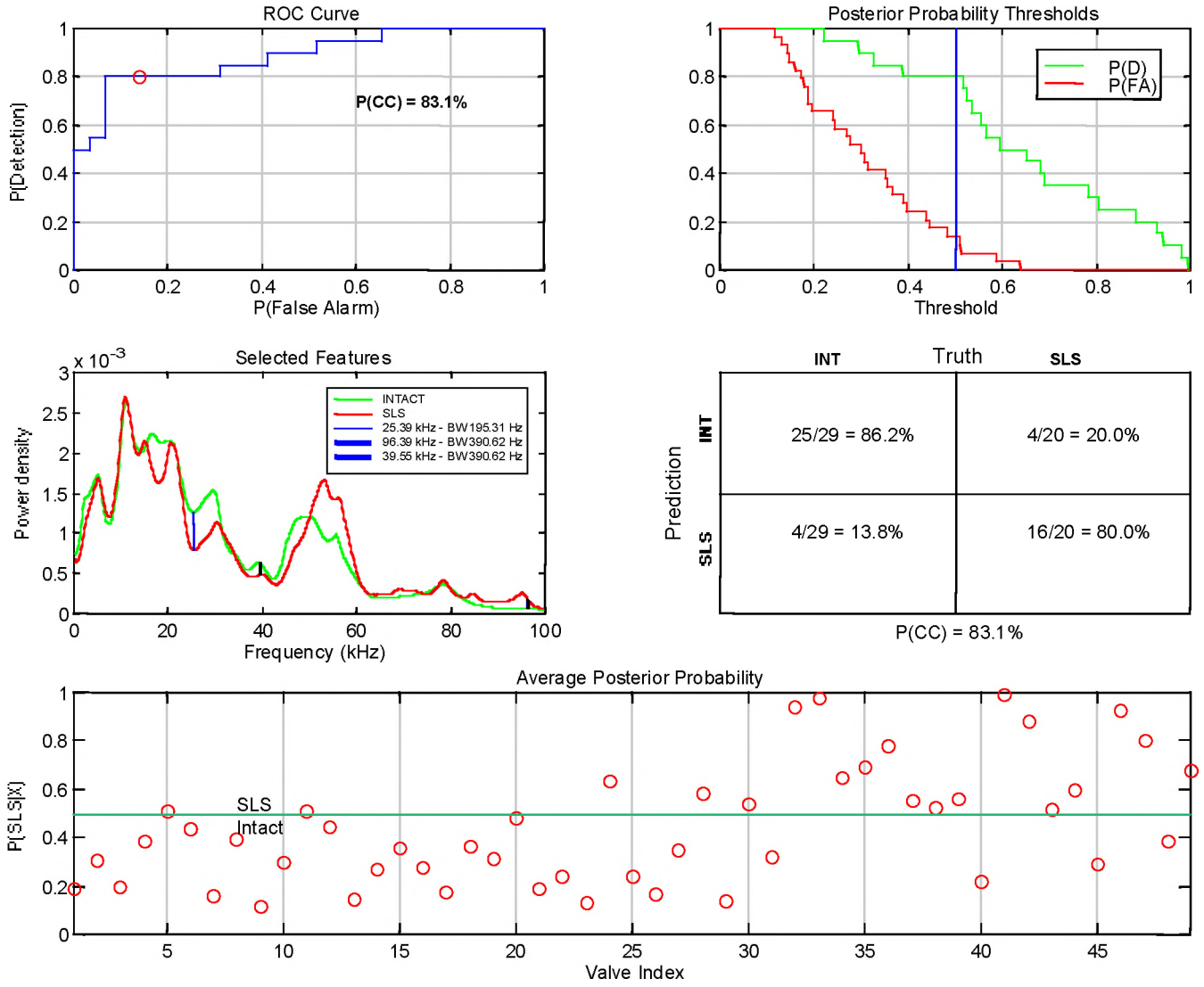


Fig. 3 The classification results are summarized for the following set of conditions: Valve #10 removed from the set, Closings sounds only, Spectral estimation algorithm used = MVDR (Minimum Variance), Class separability measure = Mahalanobis distance, Number of features used = 3.

Noise Canceling Results

We applied a noise canceling algorithm (reference) to the pre-processed opening signals in an attempt to improve the SNR before classification. The algorithm works by using a section of pre-beat noise (noise measured before the valve was excited) as a reference noise. The algorithm adjusts the magnitude and phase of the reference noise and then subtracts it from the beat signal. The result is a reduction in the noise and an increased SNR. The classification results show a probability of correct classification of 57.8%. This improvement is clearly not significant enough to recommend the classification of opening sounds for use with clinical data.

Analysis of the variability in the measurements (See Appendix IV)

During the time the experiments were taking place, visual inspection of the measurements did not reveal any obvious excessive variability in the signal power of the measurements. However, once the data were analyzed, it was found that for both opening and closing sounds, there is considerable variability from beat-to-beat for a given valve (see Appendix IV). There is also considerable variability from valve-to-valve. Appendix IV presents a discussion of the variability of the measurements. Conclusions we can draw are the following:

- Both opening and closing sounds exhibit excessive within-valve and between-valve variation in the sound signal energy.
- Opening sounds are more variable than closing sounds.

Discussion and Conclusions:

- For the opening sounds, the classifier failed to classify better than chance.
- Noise canceling applied to the opening sounds resulted in an increase in probability of correct classification to 57.8%. This improvement is clearly not significant enough to recommend the classification of opening sounds for use with clinical data.

- For the closing sounds, the probability of correct classification was 83%. We believe that the closing sounds worked better than the opening sounds for this experiment because the closing sounds have a better SNR.

- Several issues having to do with the experiments make classification difficult.

    - We observe excessive beat-to-beat variation in the sound signal energy for both opening and closing sounds. The valves are apparently not being excited consistently and with enough force, leading to excessive SNR. We believe that these effects are likely due to the distributed nature of the valve as an acoustic source. In other words, a lumped constant model of the valve may not apply.

    - We observe excessive valve-to-valve variation in the signal sound energy for both opening and closing sounds. We cannot be sure what causes this variation. Possible causes could include variability in the physical characteristics of the valves, inconsistent excitations, and environmental perturbations.

Recommendations for future work:

The opening sounds from the anechoic experiments are insufficient for classification because they have low signal-to-noise ratio (SNR), and a large amount of beat-to-beat variation. The resolution of these data issues and the improvement of classification results involves three main recommended approaches: improvements in experiment design to obtain increased SNR and reduced variability, signal processing R&D to use a nonstationary signal model rather than a stationary one and classification algorithm R&D, including work in advanced feature analysis.

Experiments

We expect that SNR can be increased and beat-to-beat variability can be reduced through improvements in the experiment design. This is the most important aspect of the recommended work, because classification results are heavily dependent on the quality of the data. Promising ideas include but are not limited to refinements in the design of the hoop that holds the valve, use of a sensor array to improve SNR, improved on-line processing to test classifier performance at the experiment site and methods for ensuring that the valve is excited exactly the same way for each beat.

Signal Processing

We recommend advanced work in signal processing, including the use of array processing and noise canceling algorithms to improve SNR. We also recommend signal processing R&D to use a nonstationary signal model rather than a stationary one. The signal is transient and therefore nonstationary, so new features using this model could result in significant benefit.

Another important aspect of this work is to ensure that all signal processing and classification algorithms are available and applied during the experiment. During this project, visual inspection of the measured signals was used to ensure data quality during the experiment. We learned later that visual inspection was insufficient to detect subtle beat-to-beat variations that were later detected during the statistical analysis. A preliminary signal processing and classification analysis should be performed on the data at the experiment site, rather than acquiring the data, ending the experiment, then processing the data later. This way, if the variability of the data is large but not detectable by visual inspection, it could be detected in time to adjust the experimental conditions to avoid the variability.

Classification

We recommend research in the area of classification algorithms. The classification results reported here show that opening sounds are not classified effectively using the algorithms and features designed for the clinical study in 1996. It is possible and quite likely that use of other features and algorithms could provide better classification results. The following steps in algorithm research are proposed as future directions.

The most important aspect of this work is feature analysis to find features which may increase classification performance. Several promising approaches have not yet been examined. These include but are not limited to the following: (1) Hierarchical multi-scale transforms, including wavelet transforms to deal with the non-stationary nature of the transient waveforms. The heart valve signals are transient and therefore nonstationary, so new features using this model could result in significant benefit. (2) A variety of specialized spectral features, (3) Specialized feature selection algorithms designed to deal with non-Gaussian-distributed data and data outliers. (4) Improvements in algorithms for dealing with small sample sizes (limited number of valves). We also recommend research in classifier algorithms to exploit recent research in the literature.

Acknowledgments

The authors gratefully acknowledge the support we have received from the Supervisory Panel for the Bowling-Pfizer Heart Valve Fund. In particular, we appreciate the technical and supervisory support we have received from Arthur W. Weyman, M.D., Donald C. Harrison, M.D., and J. Kermit Smith, M.D. The authors gratefully acknowledge the support we have received from the Shiley Heart Valve Research Center under the guidance of Dr. David Wieting, as well as the technical support from program manager, Ms. Becky Interbitzen and Shiley colleagues: Drs. Jim Chandler, Ray Chia, Stephen Schreck, and Mr. Peter Phillips. We are also indebted to Dr. Paul Stein for his medical consulting as well as co-workers Mike Buhl, Jim Candy, and Holger Jones for their many helpful suggestions ,discussions, and contributions.

Appendix I

Data Acquisition and Signal Extraction

The data for the anechoic study was collected using a National Instruments™ LabView data acquisition system. Analog signals from the hydrophone were (analog) bandpass filtered (2 Hz–100 kHz) and then time-sampled at a rate of 200 kHz providing time resolution of nearly 100 MHz, more than twice the prior in vivo studies with humans. The signals were quantized (analog-to-digital conversion) at 12 bits, providing a signal-to-error ratio of 74db [4]. We recorded data for at least 100 opening-closing cycles on each of 50 valves. Individual opening and closing sound signals were extracted off-line using our *Automated Beat Extraction Process* providing an ensemble of open and closing sounds for each valve. The report by Candy gives additional details [2].

Prior studies at LLNL [5] and elsewhere [Reynolds, 1995 #14[6]] have focused on closing sounds because they are stronger than opening sounds and easier to detect and extract. Nevertheless, we believe the opening sounds should provide more information about the condition of the valve because the occluder disc directly strikes the fractured outlet strut on an SLS when opening. As a consequence the closing sound should contain vibrational energy from the strut and our "cracked bell"_analogy obtains. Typical opening and closing sound waveforms are shown in Fig. 4. The opening sound waveform clearly exhibits a shorter duration and a lower SNR.

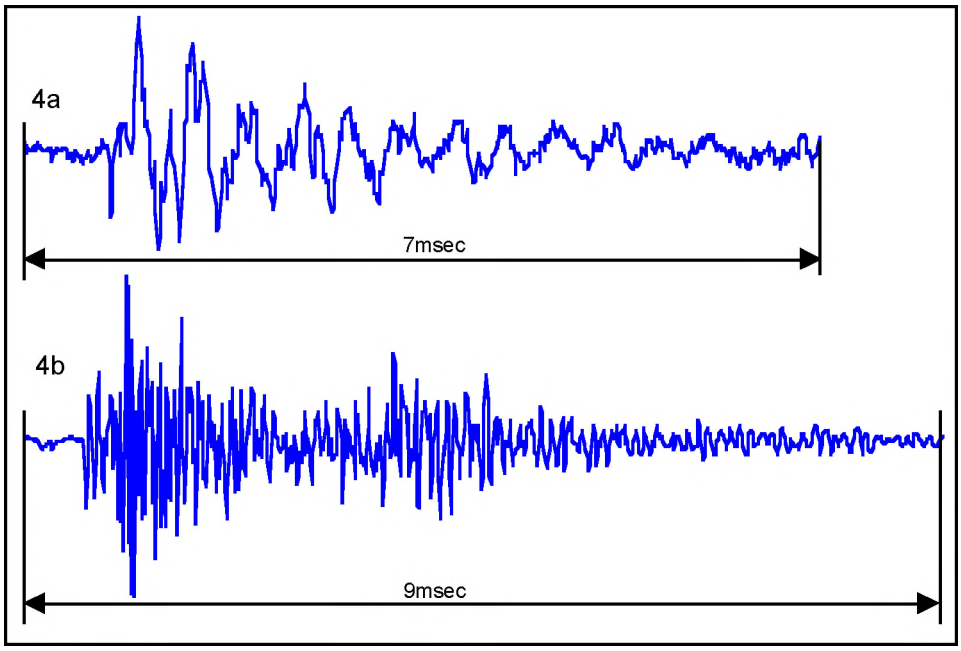


Fig. 4 (a) Extracted heart valve opening sound, (b) Extracted closing sound.

Appendix II

Signal Processing

The signal processing step converts the raw (digital) signal data into an ensemble of sound spectra used for feature extraction/selection and classification. This is accomplished by the following sequence of operations:

1. Identification and extraction of the opening and closing phases of the sound signal record using trigger pulse markers.
2. Separation of signals from leading and trailing noise
3. Trend removal and piecewise window tapering
4. Rescaling to unit variance
5. Spectrogram estimation
6. Removal of outlier spectra

We used the method known as *Minimum Variance Distortionless Response* (MVDR) [7] to calculate 100 spectra for each of the 50 valves. For comparison purposes, we also calculated spectra using Maximum Entropy Method (MEM) [8] also discussed in [9]. Because of experimental variability (discussed in more detail later in this report) the spectra exhibit scatter both within and between valves. To cope with this variability, we calculated a spectrum in the center of the scatter and retained the 50 spectra closest to this center using a distance measure based on the Median Absolute Deviation (MAD) [10]. More details on this procedure are given by Candy [2].

Once an ensemble of spectra are created and screened they are ready for the next step in the process, Feature Analysis.

Appendix III

Feature Analysis and Classification

Feature Analysis

We define *feature analysis* as an iterative procedure used to reduce the dimension of the data under consideration. Without dimensional reduction one is plagued by what Richard Bellman called "the curse of dimensionality" [11]. For example the power spectrum of a heart valve opening sound signal has dimension of 512 (the length of the spectrum vector). Since most of these dimensions contain little information useful for classification, reduction to a low dimensional subspace has the highest priority. The performance of the classifier is critically dependent on having the right set of features. Feature analysis consists of (1) *feature extraction*, or computing a large set of features based upon engineering judgment and knowledge of the physical processes that generate the data, and (2) *feature selection*, or the process of choosing an optimal subset of features from the larger set of extracted features. The process is iterative in the sense that we extract features, select features, evaluate classification performance, and loop back to the extraction step to make adjustments until optimum or acceptable performance is realized.

Feature Extraction

In general defining features is limited only by one's imagination. How we define the features depends on well we can model the underlying physical process that generated the data. In the most favorable situation, we have a full knowledge (*strong model*) of the physical processes, and we can choose features that have physical meaning. With a strong model we could numerically simulate heart valve sounds and make a meaningful interpretation of the classification. When little modeling information is available (*weak model*), we must make the best of an undesirable situation by using as much prior physical knowledge as we have, together with good engineering judgment. In the absence of prior modeling information (*no model*), we usually fail to get meaningful, interpretable results. We call this "data chasing" [12]. To mitigate against this, we perform controlled experiments where we have prior knowledge (ground truth) about the valve condition, and use intelligent search techniques to find an effective set of features.

For the heart valve classification problem, strong models are not available, moreover, we have little physical knowledge about valve responses, so our models are weak. The

challenge is to find effective features and then understand the physics of why they are effective afterward.

We decided to do our feature analysis in the spectral domain following reasons:

- Alignment of multiple opening and closing sound waveforms in time is extremely difficult [1], and we prefer to avoid it is possible.
- By using the magnitude the power spectra, we neglect the phase information and avoid the temporal alignment problem.
- Features derived from signal spectra provide more physical insight than temporal waveforms. Finite element analysis [13], laboratory studies [14], and laser vibrometer analysis [15] suggest that intact valves have a resonant frequencies missing in SLS valves. However these studies have not been conclusive for classification.
- Spectra provide a very compact representation of the information in a physical process, and compactness is desirable for reducing the dimension of the data.

Instead of using predicted or observed frequency peaks, we decided to search the entire frequency spectrum for a *set* of features that provides the best classification performance. We did this for the following reasons:

- In our opinion, the predictions from finite element modeling have not been sufficiently validated.
- Laboratory measurements show significant variability in the spectra from valve to valve.
- We have automatic techniques for feature selection that allow us to pick the best subset of features. We search an initial high-dimensional feature space for a low-dimensional subset that provides the best separation of clusters in the feature space.
- After automatically selecting features, we can still use our prior physical knowledge and engineering judgment to check that the selected features make physical sense and to learn more about the physical processes. If it turns out that the feature selector chooses the same features indicated by the models, then we will have gained a much greater confidence in the models and a much greater intuition about the physical processes. If the feature selector chooses features different than the ones indicated by

the models, then we question the validity of the models and the assumption that resonant peaks necessarily provide the best discriminants to use for heart valve classification.

Defining Features: Fixed Window Method

First we resolve the entire spectrum into N contiguous frequency bins each of width Δf. Then we form a set of *analysis windows* each of width $W\ \Delta f$, so W counts the number of frequency bins in a single analysis window, and there are a total of N/W analysis windows across the spectrum. Thus $W = 1$ gives maximum resolution, while $W = N$ gives the minimum resolution (the whole spectrum). Features are defined from the areas of the analysis windows which is the mean spectral power in the window, see Fig. 5. Our initial studies started with $W = 1$ [16] and increased later [17]. While the results were promising, we decided to abandon the fixed window technique in favor of a more sophisticated *sliding window technique* described in the next section.

Defining Features: Sliding Window Method

Since both the center frequency and width of spectral peaks are variable we have opted for a "sliding window" technique for computing features (see Fig. 5). Here, the feature we compute is the mean spectral power in a frequency band W frequency bins wide. To create the initial feature set, we let the window slide over the spectrum and compute spectral features using $W = 1, 2, \ldots, W_{\text{Max}}$. Therefore the number of features in the initial set is:

$$\sum_{i=1}^{W_{\text{Max}}} (N - i) = \frac{W_{\text{Max}}}{2}\left(2N - W_{\text{Max}} - 1\right)$$

We used $W_{Max} = 5$, so for opening sounds, we have $N = 512$ for a total of 2,5145 features in the initial set. As we show later in the paper, this very large feature set is rich in information and results in robust classifier performance.

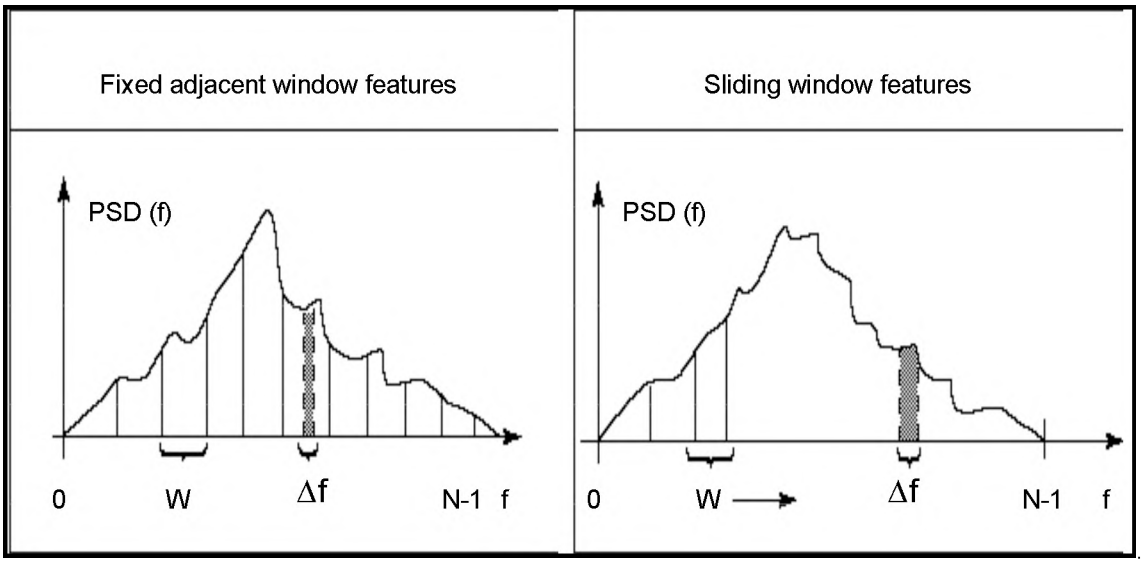


Fig. 5. For fixed adjacent window features, the number of frequency bands (= the number of features. We use the more numerous sliding window features to mitigate the possibility of missing spectral peaks.

Feature Selection

Feature selection is the process of choosing the best subset from an initial set to train the classifier. We discuss the criterion for deciding what is "best" in the next section. We do this by using a *forward selection scheme* that selects the best subset of features. Feature selection is important for several reasons.

- First, we wish to minimize the effects of the "curse of dimensionality," in the sense that the classification computational complexity increases rapidly with the dimension of the feature vector.

- We wish to use only features that add significant value to the quality of the classification results. Unimportant or redundant features add negative or zero value and should be removed [18]. Later in this section, we describe algorithms for evaluating the importance of features.

- For training sets of finite size (as we have here) the classifier performance does not monotonically increase with the dimension of the feature space, it peaks at some threshold [19]. Jain studied this peaking effect for multivariate Gaussian data [20]. While our features are not Gaussian it seems reasonable to expect the peaking effect will hold for the heart valve signal data. Clearly, our goal is to find the number of features corresponding to the knee in the curve.

- An important by-product of feature selection can sometimes be increased knowledge of the physical processes that create the data. By understanding which features are most important, we can often draw important conclusions about the physical reasons why they are important, and this can lead to productive insights that aid in the system design.

- If we use multiple sensors or multichannel measurements, we may wish to use sensor feature fusion, so feature selection helps us determine which sensors are the most important [21].

Another important consideration is a bounding relationship between the number of features used and the size of the training set. A combination of theoretical and empirical studies has led to the following *rule of thumb* [18]

$$\frac{N_S}{N_C} \geq 5\, N_F$$

Where N_S is the size of training set, N_C is the number of classes, and N_F is the number of features. For example, for 50 valves and two classes (intact, SLS) we should use fewer than 5 features.

An important implication of this rule of thumb is an upper bound on the number of features to use, given the number of independent training samples. Note that if the sample size is small, as it is in this heart valve study, it severely limits the number of features we can use. In much of our work, for example, we were limited to about 2 or 3 features, because the small sample size would not support more features. This is discussed in greater detail in section 8.

Measures of Class Separability

The best feature set is the one that will ultimately produce a classifier that will have the smallest classification error. Since it is not always feasible to select features on the basis of classification error, we need a surrogate measure. For the two group classification problem, a useful surrogate is use some measure of class separability based on an inter-group metric. Then for a given metric, the best feature set will be the one that maximizes the distance between classes. Our approach is to find the best feature set for a number of commonly used distance metrics and then use that metric which gives the smallest classification error. To demonstrate that this is a reasonable approach, we will show the relationship between classification error and inter-class distance.

We let $y \in \{0,1\}$ be the class label for intact and SLS valves respectively. The classifier is a rule $\hat{y}(\mathbf{x})$ that predicts the class membership for an observed feature vector $\mathbf{x}$. The classifier can make two kinds or errors:

1. An error of the first kind where $\hat{y}(\mathbf{x}) = 1$ when $y = 0$.
2. An error of the second kind where $\hat{y}(\mathbf{x}) = 0$ when $y = 1$

At every point in feature space the classifier will either classify correctly or make an error of the first or second kinds. The minimum error is achieved by using *Bayes' rule*

$$\hat{y}(\mathbf{x}) = 1 \text{ for } \quad f(\mathbf{x}) \geq \frac{1}{2}$$

$$\hat{y}(\mathbf{x}) = 0 \text{ for } \quad f(\mathbf{x}) < \frac{1}{2}$$

$$f(\mathbf{x}) = \Pr[y = 1 | \mathbf{x}]$$

in which case the probability of misclassification is:

$$\Pr[error] = Min\{f(\mathbf{x}), 1 - f(\mathbf{x})\}$$

The above probability is conditioned on the value of the feature vector $\mathbf{x}$. We can calculate average error for a Bayes' classifier for the whole feature space by:

$$\text{Average Error} = \int Min\{f(\mathbf{x}), 1 - f(\mathbf{x})\} p(\mathbf{x})\, d\mathbf{x}$$

where the domain of integration is the whole feature space, and $p(\mathbf{x})$ is the unconditional probability density. We can write $p(\mathbf{x})$ in terms of the prior probabilities of intact and SLS, π_0 and π_1, and the class conditional densities.

$$p(\mathbf{x}) = \pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})$$

and from the definition of $f(\mathbf{x})$ we get:

$$f(\mathbf{x}) = \frac{\pi_1 p_1(\mathbf{x})}{\pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})}$$

therefore the average error of a Bayes' classifier becomes:

$$\int Min\{\pi_0 p_0(\mathbf{x}), \pi_1 p_1(\mathbf{x})\} d\mathbf{x}$$

From the above expression, we see that if in some sense $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ are nearly identical, then minimum average classification error must be close to $\text{Min}\{\pi_0, \pi_1\}$. In other words, the classifier assigns nearly all observations to the most probable class irrespective of the observed feature. For example if $\pi_0 > \pi_1$ the classifier would classify

nearly all unknown valves as intact and the probability of misclassification is simply π_1, the prior probability of SLS. Conversely if $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ overlap very little, then the integrand in the above expression is small and the probability of error is small. Therefore we see there is a connection between misclassification and the separability of $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$. Strictly speaking this discussion applies to the Bayes' classifier which generally cannot be realized in practice because we usually don't know $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$. Nevertheless separability is a useful guideline for selecting features.

One commonly used inter-class distance metric is the *Bhattacharya distance*. This distance is directly tied to the average error of a Bayes' classifier. Bhattacharya started with the fact that the geometric mean of two positive numbers is less the minimum of the two, therefore [22]

$$\text{Min}\{\pi_0 p_0(\mathbf{x}), \pi_1 p_1(\mathbf{x})\} \le \sqrt{\pi_0 p_0(\mathbf{x})}\sqrt{\pi_1 p_1(\mathbf{x})}$$

Substituting this inequality into the expression for the average error of a Bayes' classifier gives:

$$\int Min\{\pi_0 p_0(\mathbf{x}), \pi_1 p_1(\mathbf{x})\} d\mathbf{x} \le \int \sqrt{\pi_0 p_0(\mathbf{x})}\sqrt{\pi_1 p_1(\mathbf{x})}\, d\mathbf{x}$$

which can be rewritten as follows:

$$\int Min\{\pi_0 p_0(\mathbf{x}), \pi_1 p_1(\mathbf{x})\} d\mathbf{x} \le \sqrt{\pi_0 \pi_1} \int \sqrt{p_0(\mathbf{x}) p_1(\mathbf{x})} d\mathbf{x} = \sqrt{\pi_0 \pi_1}\, e^{-J_B}$$

The integral part of the term on the right in the above equation is the basic definition of the Bhattacharya distance J_B between $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$. A feature set that maximizes this distance will minimize an upper bound of the average probability of misclassification error. When $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ are both multivariate normal, the Bhattacharyya distance becomes:

$$J_B = \frac{1}{8}(\mu_0 - \mu_1)^T \left[\frac{\Sigma_0 + \Sigma_1}{2}\right]^{-1} (\mu_0 - \mu_1) + \frac{1}{2}\log \frac{\left|\frac{1}{2}(\Sigma_0 + \Sigma_1)\right|}{\left[|\Sigma_0||\Sigma_1|\right]^{\frac{1}{2}}} \qquad (3)$$

where μ_0 and μ_1 $\vec{\mu}_i$ and $\vec{\mu}_j$ are the mean vectors computed over the feature vectors in classes 0 and 1 and Σ_0 and Σ_1 are the corresponding covariance matrices.

Another metric is *Kullback-Liebler distance* (*KL*) [Kullback, 1994 #54]. Unlike Bhattacharyya, the *KL* distance is not specifically tied to the Bayes' classifier; it is

motivated by concepts from statistics and information theory. The basic idea behind the KL is to write the unconditional log likelihood ratio of $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ as follows:

$$\log\left(\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})}\right) = \log\left(\frac{\Pr[y=0|\mathbf{x}]}{\Pr[y=1|\mathbf{x}]}\right) - \log\left(\frac{\Pr[y=0]}{\Pr[y=1]}\right)$$

The right hand side of the above equation is the log of the odds in favor of $y = 0$ given an observation of the feature $\mathbf{x}$. It is also called the *discrimination* or *weight of evidence* for $y = 0$ against $y = 1$. If we average the log likelihood ratio over all of the feature space, we get the basic defining equation for the *KL* distance:

$$KL(p_0(\mathbf{x}), p_1(\mathbf{x})) = \int p_0(\mathbf{x}) \log\left(\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

which is interpreted as the mean information for discrimination of $y = 0$ against $y = 1$ provided by the feature $\mathbf{x}$. Strictly speaking the *KL* distance is not a metric because it lacks the symmetry property $KL(p_0, p_1) \neq KL(p_1, p_0)$ and the triangle inequality does not hold. Nevertheless it has found very wide application in classification, and we use it as if it were a metric. The *KL* is also called the *directed divergence* [23]. A symmetric form of the *KL* called the *Jeffrey's distance* or the *divergence*
$J_D(p_0, p_1) = KL(p_0, p_1) + KL(p_0, p_1)$.

Again when $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ are both multivariate normal, we can get an explicit formula for the *KL* distance [23]:

$$KL(\mathrm{p}_0, p_1) = \frac{1}{2}\log\frac{|\Sigma_0|}{|\Sigma_1|} + \frac{1}{2}\mathrm{tr}\left(\Sigma_0\left(\Sigma_1^{-1} - \Sigma_0^{-1}\right)\right) + \frac{1}{2}\mathrm{tr}\left(\Sigma_2^{-1}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T\right)$$

an expression distinct from the Bhattacharyya distance. However, when the two class covariance matrices are equal, $\Sigma_0 = \Sigma_1 = \Sigma$, both the Bhattacharya distance and the *KL* distance reduce to the same metric known as the *Mahalanobis* distance [24].

$$\Delta^2 = (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)$$

Thus we can interpret the Mahalanobis distance as a bound on classification error for a Bayes' classifier, or more generally as an information metric. Usually don't know $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ nor are the features necessarily normally distributed. As a matter of practice, we use all these metrics to select feature sets and then see which set gives a classifier with the best performance.

Searching for the Best Features: The Branch and Bound Algorithm

To select a feature set, we could search all possible subsets of the largest feature set and choose the subset that maximizes a given metric. This approach can be computationally intense. For example if the large set has 50 features, and we limit the maximum number of features to 5 in accordance with our "rule of thumb", then we must search:

$$\binom{50}{1}+\binom{50}{2}+\binom{50}{3}+\binom{50}{4}+\binom{50}{5}=2,369,935$$

different subsets.

One alternative to the exhaustive search technique is the *branch and bound algorithm*, [25]. It is globally optimal in the sense that it finds the optimal feature set, but it generally does not require as much computation time as the exhaustive search method. At worst using branch and bound can be as computationally intensive as the exhaustive search, but that rarely happens. The branch and bound works by rejecting suboptimal subsets without direct evaluation of the distance metric J, and guarantees that the selected subset yields the globally best value of any criterion function J that satisfies a monotonicity condition.

$$J_1(x_1)\geq J_2(x_1,x_2)\geq\cdots\geq J_m(x_1,x_2,\cdots x_m)$$

where $J_i(x_1, x_2, \cdots x_i)$ is the criterion function evaluated using all features except $x_1, x_2, \cdots x_i$ from the feature set. The restriction of monotonicity is not severe, and is not a limitation in practice, because it simply requires that a set S of features is at least as good as any proper subset of itself for the purpose of class separation. A large number of criteria, including the Bhattacharya criterion satisfy the monotonicity condition [25]. While more efficient than exhaustive search, the branch and bound algorithm is still too computationally intensive for the heart valve problem feature set. Instead we have elected to use the *sequential forward selection* algorithm despite its suboptimal properties.

Sequential Forward Selection

Sequential forward selection (SFS) is a bottom-up process. The algorithm starts with the null (empty) set of features. Then at the first step it chooses a feature from the maximal feature set that maximizes the distance metric J. The monotonicity condition guarantees than any of the remaining features will increase J, therefore pick the one that causes the largest increase. Continue adding features until a stopping criterion is met such as the number of features.

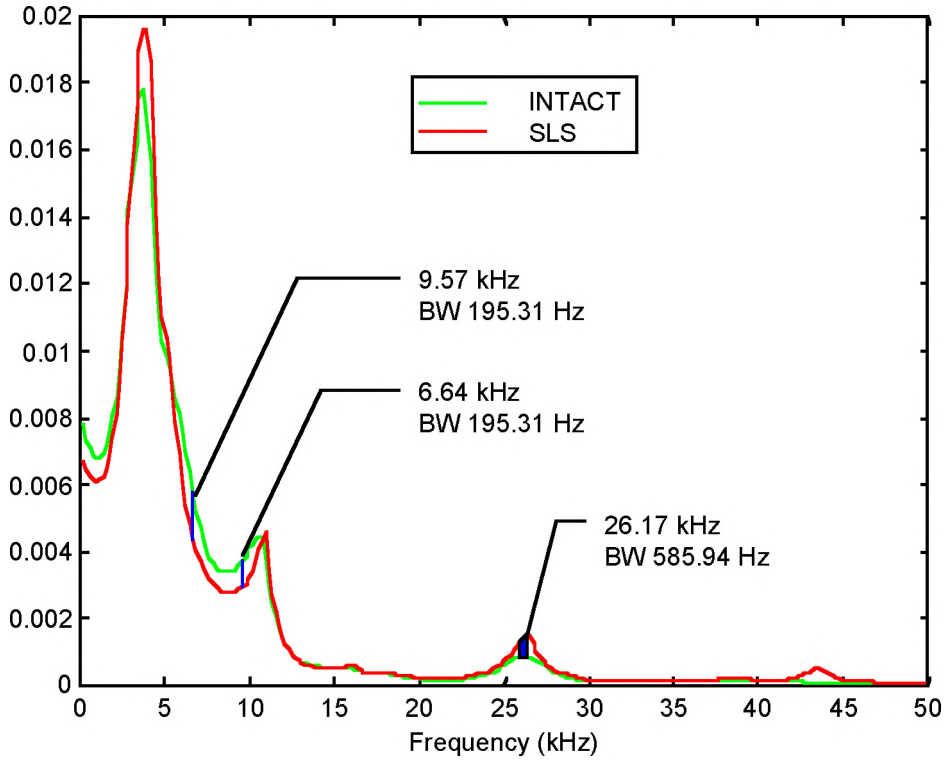


Figure 6. Features selected using the data from Nov 98. SFS, Mahalanobis, 3 features, classic covariance, two-stage classifier.

The SFS algorithm requires much less computation than branch and bound algorithm, and appears to work well in practice [17]. Nevertheless it has drawbacks. The algorithm never removes a selected feature and it could miss a combination of features that that is superior and thus SFS might not capture the optimal subset that an exhaustive search would discover. For heart valve analysis, we found the performance of the SFS algorithm to be generally satisfactory, and the optimality of the branch and bound algorithm was not worth the high computational cost.

Classification

Classification for the heart valve sounds uses the *supervised learning* approach. This approach uses a *training sample* of known cases to construct a classification rule to classify future unknown cases. Both the training sample and future unknown cases are assumed to be random samples from the same population of possible valves. The process is two-step: 1. Training and 2. Performance evaluation. With a limited amount of data (as is the case for the heart valve study) there is tension between the two steps in the sense that each step needs sufficient data to work reliability. We will discuss methods to deal

with this tension so that the available data does "double duty" for training and evaluation in a statistically valid way.

The training step makes use of a data base $T = \{\mathbf{x}_i, y_i\}$, of previously solved cases where we know the value of $y_i \in \{0,1\}$ (the group identifier) and the corresponding feature vector $\mathbf{x}_i$ for each valve *i*. Most supervised learning approaches to classification try to approximate the Bayes' classifier by "learning" the function $f(\mathbf{x}) = \Pr[y = 1|\mathbf{x}]$ from the training data. Two contrasting points of view have emerged from research on how to do this. The first, known as the *diagnostic paradigm* [26] casts the problem as one of function estimation. In other words a regression like framework. Popular methods that follow this approach are nearest neighbor methods, artificial neural nets, and logistic regression. The other point of view, the *sampling paradigm*, uses estimates of the class conditional distributions $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ with the prior probabilities π_0 and π_1 to approximate $f(\mathbf{x})$ by Bayes' formula and the *plug-in* principle. The estimated distributions (indicated by the circumflex notation) are "plugged into" the Bayes' rule as if they were the true distributions.

$$\hat{f}(\mathbf{x}) = \frac{\pi_1 \hat{p}_1(\mathbf{x})}{\pi_0 \hat{p}_0(\mathbf{x}) + \pi_1 \hat{p}_1(\mathbf{x})}$$

Examples of methods following this approach are Fisher linear discriminant analysis [19], and kernel discriminant analysis [27].

For the heart valve classifier, we have adopted the sampling paradigm with the plug-in principle, using the Parzen estimator [28] to calculate class-conditional density estimates $\hat{p}_0(\mathbf{x})$ and $\hat{p}_1(\mathbf{x})$. This approach has the desirable property that it provides the optimal Bayes classifier in the limit as the number of training samples approaches infinity. The classification rule becomes:

For $\pi_0 \hat{p}_0(\mathbf{x}) > \pi_1 \hat{p}_1(\mathbf{x})$ choose INTACT
For $\pi_0 \hat{p}_0(\mathbf{x}) \leq \pi_1 \hat{p}_1(\mathbf{x})$ choose SLS

When training the classifier, the prior probabilities π_0 and π_1 are set to their respective proportions in the training set. For clinical application π_0 and π_1 would incorporate all information about the patient and the best estimate about the incidence of SLS in the population from which the patient was selected. This step is extremely important as we expect $\pi_0 >> \pi_1$. We also see that the features must provide good separability or else the priors will dominate the decision rule.

The estimate for $\hat{p}_1(\mathbf{x})$ is given by the following expression [29]:

$$\hat{p}_1(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} m_1 s_1^d} \sum_{i=1}^{m_1} \exp\left[-\frac{(\mathbf{x}-\mathbf{X}_i)^T(\mathbf{x}-\mathbf{X}_i)}{2s_1^2} \right]$$

m_1 = number of SLS feature vectors

$\mathbf{x}_i$ = ith SLS feature vector or dimension d, $i = 1,2,\ldots m_1$

d = dimension of feature space

$\mathbf{x}$ = d dimensional point in feature space

s_1 = smoothing parameter for SLS features

The estimate for $\hat{p}_0(\mathbf{x})$ follows by an obvious change of notation. The choice of the smoothing parameters s_1 and s_2 are essential for tuning the pdf estimators. The accuracy of the decision boundaries depends upon the accuracy with which the underlying pdf's are estimated. In the limit, as $s_{1,2} \to 0$, the result is no smoothing at all, and the classifier approaches the nearest neighbor classifier [30]. As $s_{1,2} \to \infty$, the result is over smoothing and the classifier approaches a linear classifier, with the decision surface being a hyperplane.

Misclassification can happen in two distinct ways. First we can classify an intact valve as SLS, an error of the *first kind*. Conversely we could classify an SLS valve as intact, an error of the *second kind*. Following standard statistical notation from hypothesis testing, the probability of the error of the first kind is α and the probability of a error of the second kind is $1-\beta$. Medical terminology uses the term *specificity* for the quantity $1-\alpha$, and *sensitivity* for the quantity β within the framework of hypothesis testing. If $\mathbf{X}$ is a random feature vector then the error probabilities are given by:

$$\alpha = \Pr\left[\pi_0 \hat{p}_0(\mathbf{X}) \le \pi_1 \hat{p}_1(\mathbf{X}) \middle| y = 0\right]$$
$$1-\beta = \Pr\left[\pi_0 \hat{p}_0(\mathbf{X}) > \pi_1 \hat{p}_1(\mathbf{X}) \middle| y = 1\right]$$

If $\hat{p}_0(\mathbf{x}) = p_0(\mathbf{x})$ and $\hat{p}_1(\mathbf{x}) = p_1(\mathbf{x})$ then the classifier achieves the minimum *pmc* (probability of misclassification) with

$$pmc = \pi_0\alpha + \pi_1(1-\beta)$$

and the actual values of α and β are determined by carrying out the appropriate multidimensional integrals. Following Bayes' rule will automatically determine α and β. If we want a different value of α then we must modify the decision rule which will change the value of β and the classifier will no longer achieve the minimum *pmc*. The

α–β tradeoff is known as the receiver operating curve (ROC) and the Bayes' classifier is a specific point on this curve.

The decision rule becomes:

$$\frac{\hat{p}_0(\mathbf{x})}{\hat{p}_1(\mathbf{x})} > \lambda \Rightarrow y = 0$$
$$\frac{\hat{p}_0(\mathbf{x})}{\hat{p}_1(\mathbf{x})} \leq \lambda \Rightarrow y = 1$$

The threshold λ determines the operating point on the ROC. For the Bayes' classifier

$$\lambda = \frac{\pi_1}{\pi_0}.$$

Practical Aspects of Choosing the Smoothing Parameter

We have two methods for choosing the smoothing parameter, *s*; manually and automatically. In the *manual* mode, we simply choose values of *s* and compute the resulting *probability of correct classification*(*pcc*). The curve for *pcc* generally has a "knee" or maximum, and we choose the value of *s* that maximizes *pcc*. In the *automatic* mode, we build a loop into the classifier software that allows us to automatically try a range of values for *s* and map out a curve for *pcc* versus *s*, which we can automatically search for the maximum. One possible disadvantage of this technique is that the same value of *s* is used for both intact and SLS. We plan to use different smoothing parameters for future work based on our success in other classification projects.

We have found that even small changes in the smoothing parameter affect the pcc for the heart valve data, and for this reason we opted for the automatic tuning of *s*. Others [31] report pcc as being to small changes of *s*. However Spect's comment is at least in part motivated by experimental results using 46 dimensional feature space with only 249 cases in the training set.

We used kernel-based classifier described here instead of the back propagation (artificial) neural network [30] because, it learns with only one pass through the data, does not get stuck in local minima, and can be updated easily as new training data become available. It's main drawback is that it sometimes requires more storage, but this problem can be mitigated.

The Two Stage Classifier

We use a two-stage classification scheme, because we need to *fuse* the data from multiple signals from a single valve. We fuse by using a two-stage classifier, made from two single-stage classifiers, one classifying the signals from a single valve, and the second uses the results of the first classifier as features for the second classifier which classifies valves. The two-stage classifier is diagramed in Fig. 7.

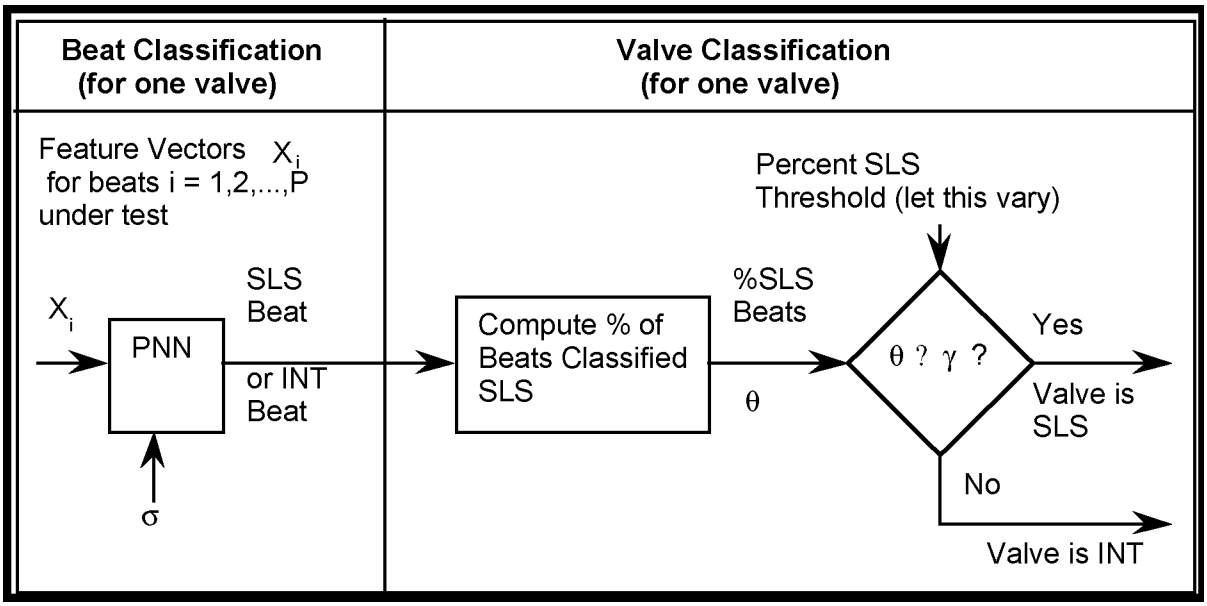


Fig. 7. Decision fusion is achieved as the two-stage classifier classifies beats for a given valve in the first stage, then classifies the valve in the second stage.

First Stage Classifier

In the first stage, the training set consists of all features from opening signal spectra from all valves. Two feature vectors are treated as independent cases even if they originate from the same valve. This produces large training set of approximately 50 x 50 = 2,500 cases. The smoothing parameter is tuned by finding the value that minimizes pmc. The output of the classifier consists of assigning each signal spectrum to a valve class, intact or SLS. These decisions are then passed off to the second stage classifier as inputs.

Second Stage Classifier

The second stage classifier works by using a one-dimensional feature based on the fraction of signal spectra classified as intact or SLS for a given valve. Thus the training set is given as:

$$T = \{\theta_i, y_i\}$$
$$\theta = \frac{\text{number classified correctly}}{\text{total in the class}}$$
$$y \in \{0,1\}$$

where the index i runs from 1 to the total number of valves from both classes. For example (θ_4, y_4) could have a value $(.8,1)$. This would mean 80% of the signal spectra were classified as class 1 (SLS) for valve 4 which is known to be SLS. The second stage classifier is equivalent to Neyman-Pearson-Wald detector for a simple (as opposed to composite) hypothesis where the null hypothesis is a valve is intact. The decision rule is:

$$\theta > \gamma \Rightarrow y = 1$$
$$\theta \leq \gamma \Rightarrow y = 0$$

By allowing the threshold γ to vary over a range of values, we can map out an ROC and select combinations of sensitivity and specificity.

Statistical Confidence Interval

The sponsors are very interested in knowing the confidence with which we can specify the performance of classifiers. In addition, they are extremely interested in knowing how many valves must be explanted from patients in order to train the classifiers to obtain acceptable performance. Clearly, these issues have great medical, social and monetary impact. In this section, we present techniques for answering these important question.

In the process of evaluating valve classification performance, we estimate conditional probabilities based upon experiments with real data and a finite number of statistical samples. We can specify the performance in terms of sensitivity and specificity. In order to fully specify the performance, however, it is desired to specify the confidence we have in the estimates of the conditional probabilities. We can do this by calculating a statistical confidence interval about the *pcc*.

If we model the classifier as a sequence of Bernoulli trials [32], the number of correct classifications has the binomial distribution with parameter *pcc* and number of trials *N*. In the absence of knowledge of prior probabilities and losses, we assume that both classes are equally probable, so $\pi_0 = \pi_1 = 0.5$. In this case we get:

$$pcc = 1 - pmc = 1 - \left[\pi_0 \alpha + \pi_1 (1 - \beta)\right]$$
$$= 1 - \frac{1}{2}(\alpha + 1 - \beta) = 1 - \alpha + \beta$$
$$= \frac{1}{2}[\text{sensitivity} + \text{specificity}]$$

The maximum likelihood estimate pcc is given by:

$$\hat{pcc} = \frac{\text{number of correct classifications}}{\text{number of test cases}}$$

We can write the 95% confidence interval about the true value of pcc as follows.

$$P\{L < pcc < U\} = .95$$

where L and U are the lower and upper bounds, respectively, of the confidence interval. The confidence interval is a random interval that covers the true probability with a frequency 95%. This does not mean a particular interval contains the true value of pcc with probability 95%. The reason for this somewhat convoluted interpretation is that pcc is an unknown constant and not a random variable, and we cannot make probability statements about constants within the frequentist framework statistics. The normal approximation for the confidence interval uses:

$$L = \hat{pcc} - 1.96\sqrt{\frac{\hat{pcc}\,\hat{pmc}}{N}} \quad , \textit{and} \quad U = \hat{pcc} + 1.96\sqrt{\frac{\hat{pcc}\,\hat{pmc}}{N}}$$

However, for our application, we are very interested in small sample sizes, because we have only 50 valves in the training set. Therefore, we are forced to use more accurate estimates of L and U which are valid for small sample sizes, and these estimates are given as follows:

$$L = \frac{N\,\hat{pcc} + 2 - 2\sqrt{N\,\hat{pcc}\,\hat{pmc} + 1}}{N + 4} \quad , \quad \text{and} \quad U = \frac{N\,\hat{pcc} + 2 + 2\sqrt{N\,\hat{pcc}\,\hat{pmc} + 1}}{N + 4}$$

We can evaluate L and U, and plot them versus pcc and $\hat{pmc}$, for various values of N, as in Fig. 9. In this case, N is the number of valves in the training set, so this plot can give us important insight into the number of training valves required to give a satisfactory confidence interval.

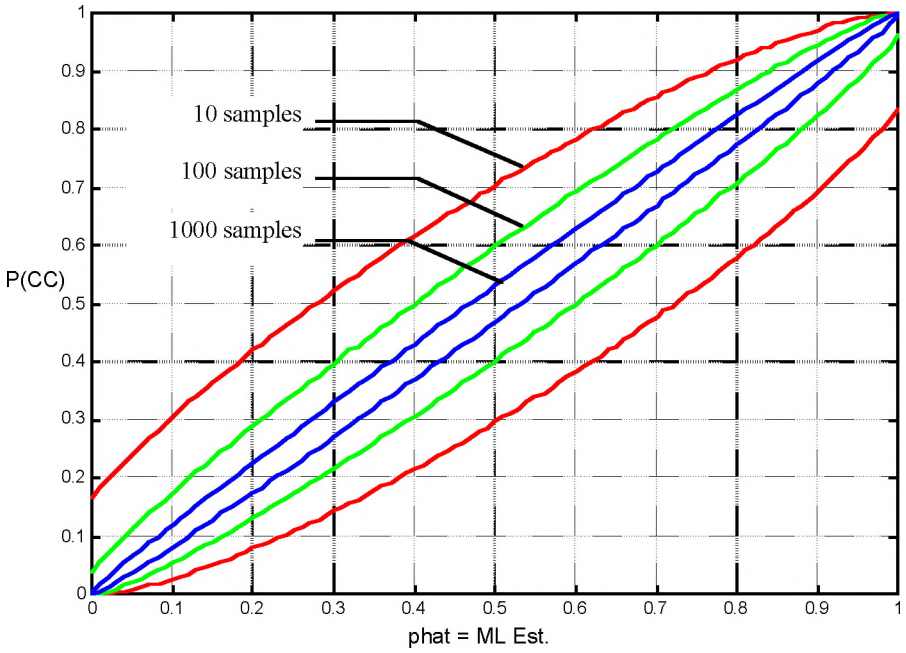


Figure 8. The lower and upper bounds (L and U respectively) for the 95% confidence interval about the true value of probability of correct classification (p) are plotted for various values of N (or n), the number of valves in the training set. The abcissa depicts the maximum likelihood estimate of p obtained from the results of training experiments with valves having known condition. The ordinate depicts the values of L and U. The approximation for large sample size is not appropriate for our application, so we use the better approximation, which is valid for small sample size.

For example, during the training process with real valves, we obtained an estimated $\hat{p}$ equal to 1. The sponsor, Shiley, is very interested in knowing the confidence with which we can specify p for a given valve, and they are very interested in knowing how many known valves they must obtain from clinical explantations in order to obtain an acceptable confidence. For the case in which $\hat{p}$=1, we can see from Fig. 9 that the upper bound, U, is always equal to one, but the value of L varies significantly as a function of the number of valves, N, as shown in Table 1. For a small number of valves, N=10, we see that L = .71, which the sponsor found to be clearly unacceptable. For N = 1000, we can achieve a very high confidence with L = .996. The sponsor, of course, finds this to be acceptable, but impractical, because it would require an enormous monetary cost and an excessive amount of time to wait for 1000 patients to become available and to actually explant 1000 valves. On the other hand, when N=100, L = .96, which is both acceptable and practical. L=.96 is high enough to be of value when decision-makers are specifying confidence. In addition, N=100 is a reasonable compromise, because we have plans to obtain information from about 100 explanted valves within the next year or so.

Table 1. This table of confidence interval bounds represents selected values of the bounds (shown graphically in Fig. 9 for the case in which the estimated probability of correct classification equals one. The importance of a large sample size (number of valves) is evident.

N, the # of training valves	Lower Bound, L	Upper Bound, U
10	.7183	1.0
17	.8095	1.0
19	.826	1.0
100	.9615	1.0
1000	.9960	1.0

For one of our earlier studies, we had only N=17 valves in the training set. In that case, we achieved P(CC)=1, and this lead to confidence interval bounds of L=.8095 and U=1.0. Our current work has N=19, which leads to L=.826 and U=1.0. These bounds are not acceptable, and we await the arrival of data from additional valve explantations that will give at least N=100, so we can obtain a more acceptable confidence interval with at least L=.96 and U=1.0.

Performance Assessment

Once a classifier is fully specified, we need to determine how well it will perform on future valves where the true condition is unknown. One commonly used performance measure is the probability of misclassification *pmc*. Usually we cannot calculate *pmc* from first principles, and we have to estimate it by running the classifier on a *test set* of known valves. The estimate $\hat{pmc}$ (the *apparent error rate*) should be near the true error rate *pmc*. We measure nearness by the *mean squared error* (MSE):

$$MSE = E\left[\left(pmc - \hat{pmc}\right)^2\right]$$

We can write the MSE in terms of the BIAS (systematic error) and VAR (random error):

$$\begin{aligned} MSE &= \left(E\left[\hat{pmc} - pmc\right]\right)^2 + E\left[\left(\hat{pmc} - pmc\right)^2\right] \\ &= BIAS^2 + Var^2 \end{aligned}$$

Using the training set as the test set (usually) produces a negative BIAS, so on average the apparent error rate is less than the true error rate, and the performance of the classifier appears better than it actually is. Using a test set independent of the training set will

guarantee an unbiased estimate of *pmc*, but at the expense of requiring a large test set. For example, if *pmc* = 5%, then for *N* valves in the test set, the standard deviation of $\hat{pmc}$ is:

$$\sigma_{\hat{pmc}} = \sqrt{\frac{pmc(1-pmc)}{M}}$$

Setting two standard deviations to be 2% and solving for *M* we get *M* = 475 for an approximately 95% confidence interval with end points of .01 and .09. Therefore an independent test set would require almost 500 valves to get a reasonably accurate estimate of pmc, nearly ten times the number of available valves. Using any valves from the training set to form the test set would also diminish the performance of the classifier. To cope with the need to use as many valves in the training set as possible for small data bases, the *hold-one-out technique* [33] is often used to estimate pmc. Here, we use all the available data samples to train the classifier, except for one which is "held out." Next, we insert the held out sample back into the training set and hold out another sample for testing. We repeat the procedure, holding out one sample and training with the remaining samples at each iteration until all of the samples have been held out once. The misclassification rate is then estimated by predicting all the held-out valves. Unfortunately, while hold-one-out provides an unbiased estimate of pmc, it sometimes has a high variance [34] because the hold-one-out training sets are too similar to the full training set. This can be a problem when the prediction rule is unstable. By accepting some bias, but less variance a smaller MSE is achieved by using *k-fold cross-validation* instead of hold-one-out. With this technique, the data base is divided into *k* equal parts (usually 5 or 10). One part is selected as the test set and the classifier is trained on the remaining k - 1 parts. This process gives *k* estimates of pmc, which are combined. Note that when $k = N$, *k*-fold cross-validation becomes hold-one-out.

Appendix IV

Analysis of Sound Data Variability

An important experimental diagnostic is the variability of the energy in an opening or closing sound signal. The electrical output (a voltage) $v(t)$ from the hydrophone at time t is proportional to the sound pressure $p(t)$ incident on the hydrophone aperture. If we integrate $p^2(t)$ over the duration T of the sound transient, the resulting quantity is proportional to the acoustic energy incident on the area of the aperture. Dividing by T gives the average acoustic power captured by the hydrophone. Therefore we can write:

$$\frac{1}{T}\int_0^T p^2(t)dt \sim \frac{1}{T}\int_0^T v^2(t)dt \approx \frac{1}{N}\sum_{i=1}^{N} v_i^2 = \overline{v^2}$$

$$\overline{v^2} = \mathrm{var}(v) + \overline{v}^2 = \mathrm{var}(v)$$

The last equation follows from the mean hydrophone voltage $\overline{v}$ being zero. We see that the (statistical) variance of the samples of sound signal amplitudes is proportional to average acoustic power. Finally taking the square root of the variance gives the RMS (root mean square) of the signal, a measure of the intensity of the sound. By studying the variation in RMS, we can identify irregular sounds.

We use the boxplot technique [10] to display the RMS data. Each box icon corresponds to data from a single valve, and it gives a graphical representation of the summary statistics: location, spread, skewness and outliers. For example, in Fig. 9 we have the RMS values of opening sounds from intact values, and we see that valve 25 has an single large outlier. This signal should be removed. We can also see that the whole of the RMS values from valve 28 are spread over a larger range than the other valves in this group. However, in Fig. 10 we have a similar plot for the closing sounds and Note both valves 25 and 28 are unremarkable. Figs. 11 and 12 are the boxplots for SLS valves. Again we can see valves with outliers such as 36 and 44. The figures give an visual impression that the RMS data for opening sounds is more variable than for closing sounds. We can compare opening and closing sounds by using the *coefficient of variation* CV, defined as the ratio of the standard deviation to the mean.

$$CV = \frac{\sqrt{Var(RMS)}}{Mean(RMS)}$$

Since there are a number of outliers, we replace the standard deviation by its robust equivilent, the *MAD* statistic [10], defined by:

$$\text{MAD}(v) = \text{Median}\{\, |v_1 - m|, \ldots |v_n - m| \,\}$$

where $m = \text{Median}(v)$

In Figs. 13 and 14 we plot the difference in coefficients of variation (opening CV minus closing CV) for both intact and SLS valves. Both groups show greater variation for opening sounds.

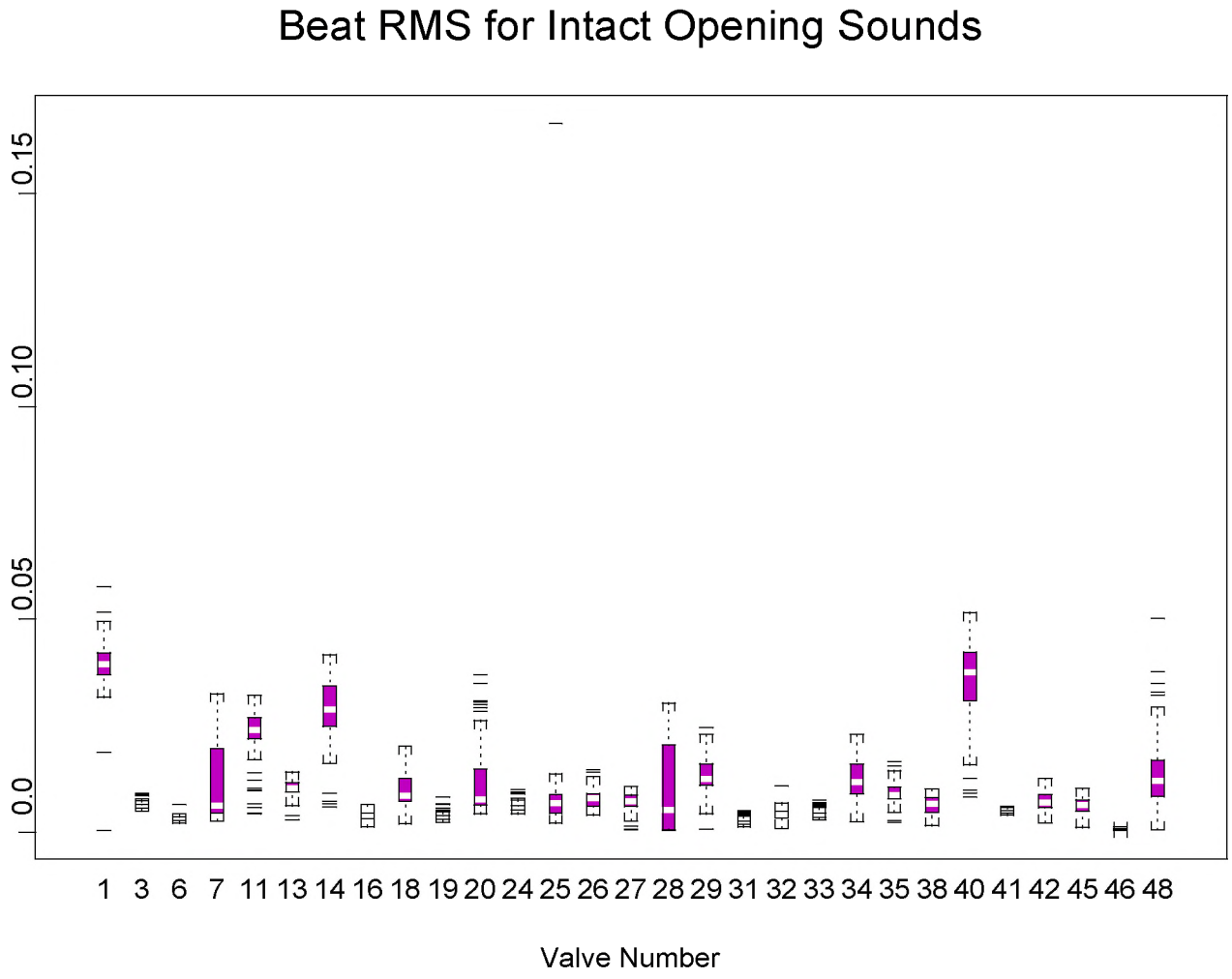


Fig. 9. Beat RMS values for intact valves, opening sounds

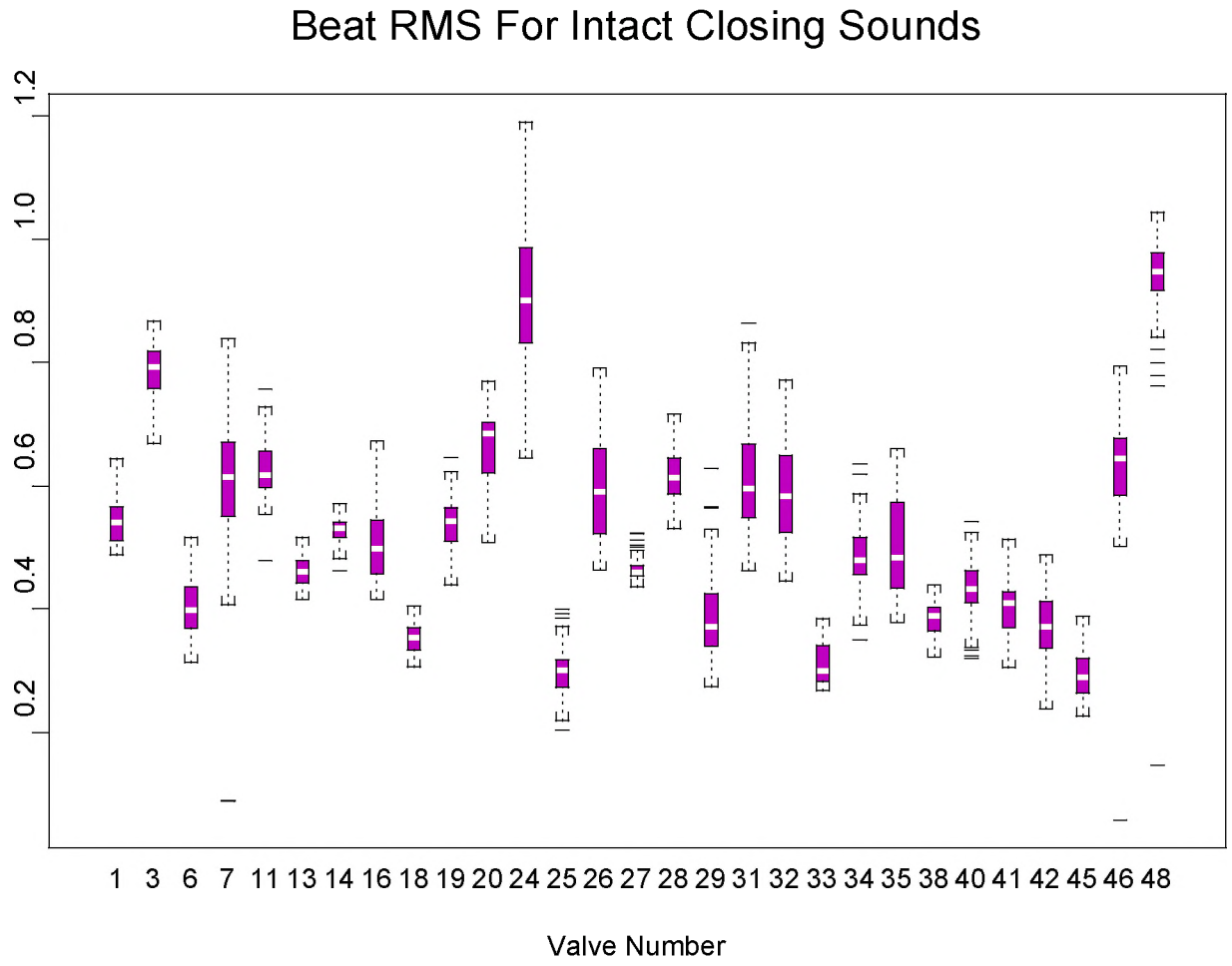


Fig. 10. Beat RMS values for intact valves, closing sounds

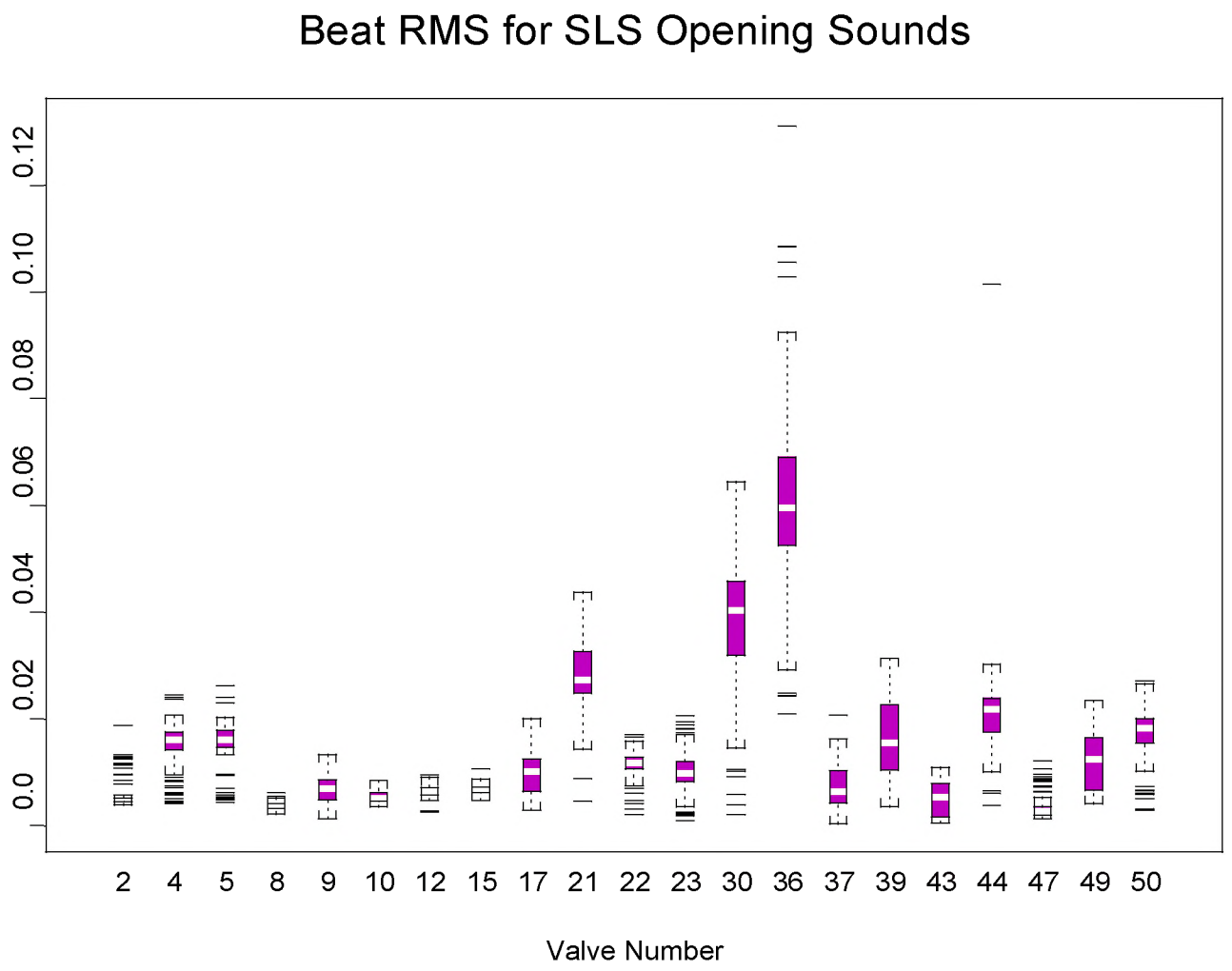


Fig. 11. Beat RMS values for SLS valves, opening sounds

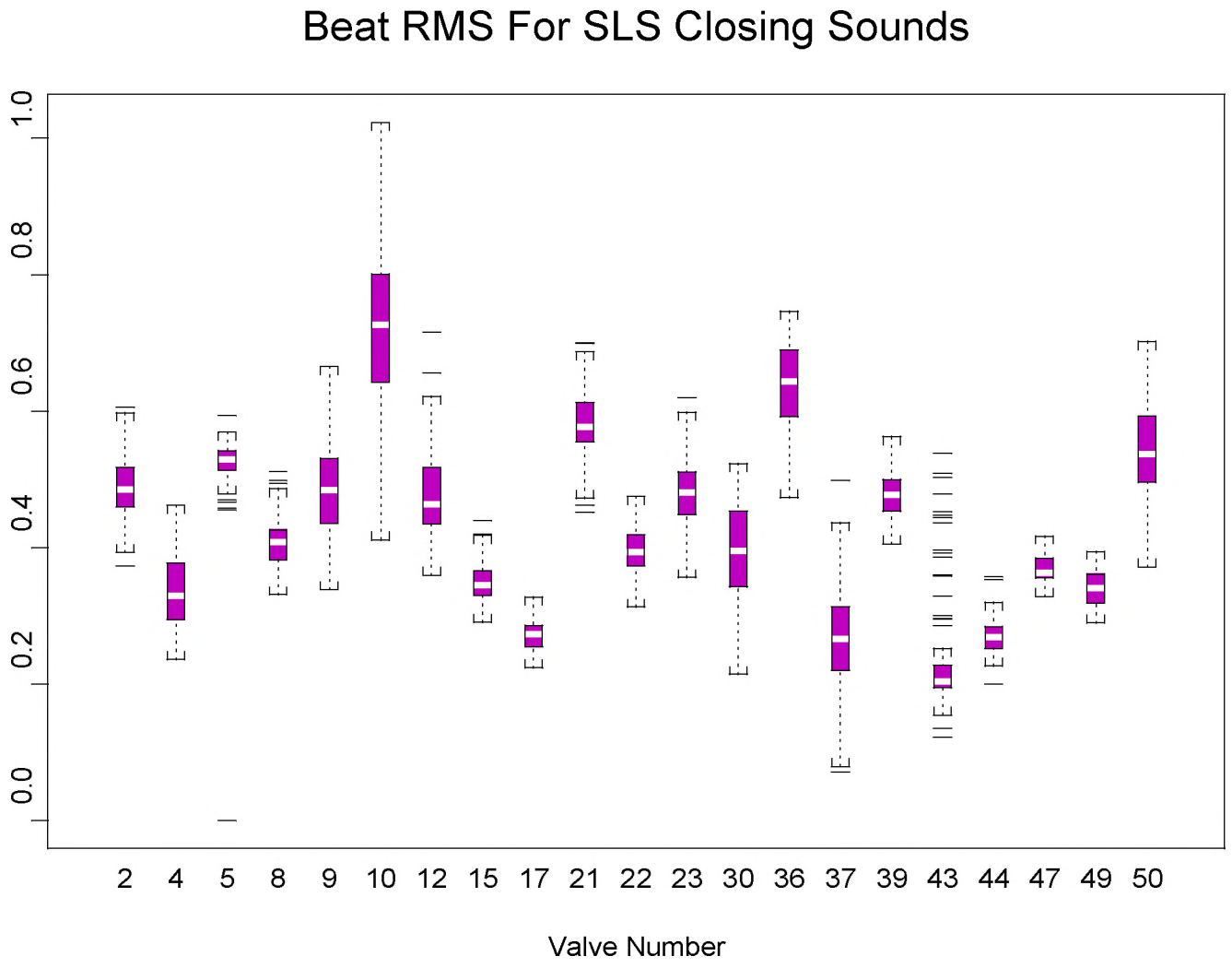


Fig. 12. Beat RMS values for SLS valves, closing sounds

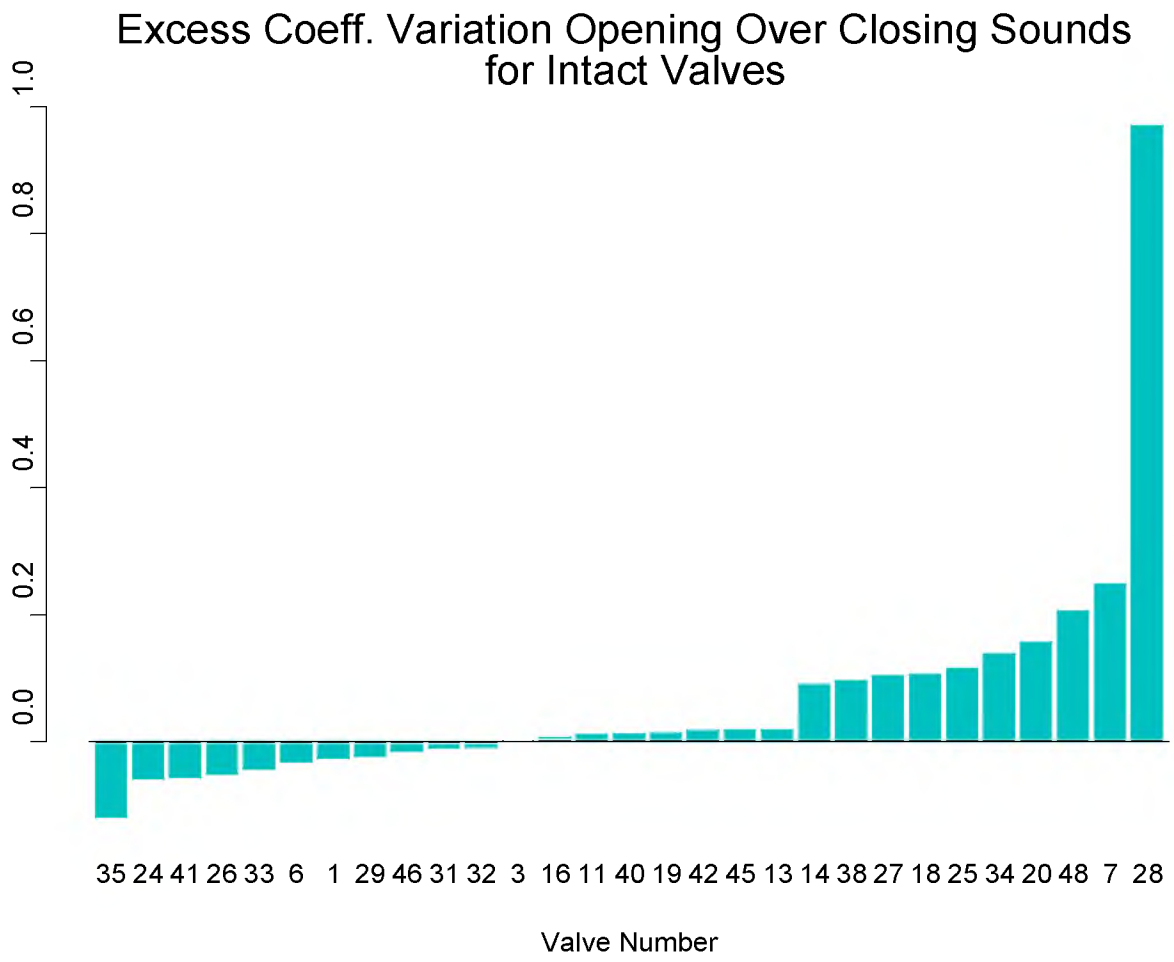


Fig. 13. Excess coefficient of variation over closing sounds for intact valves

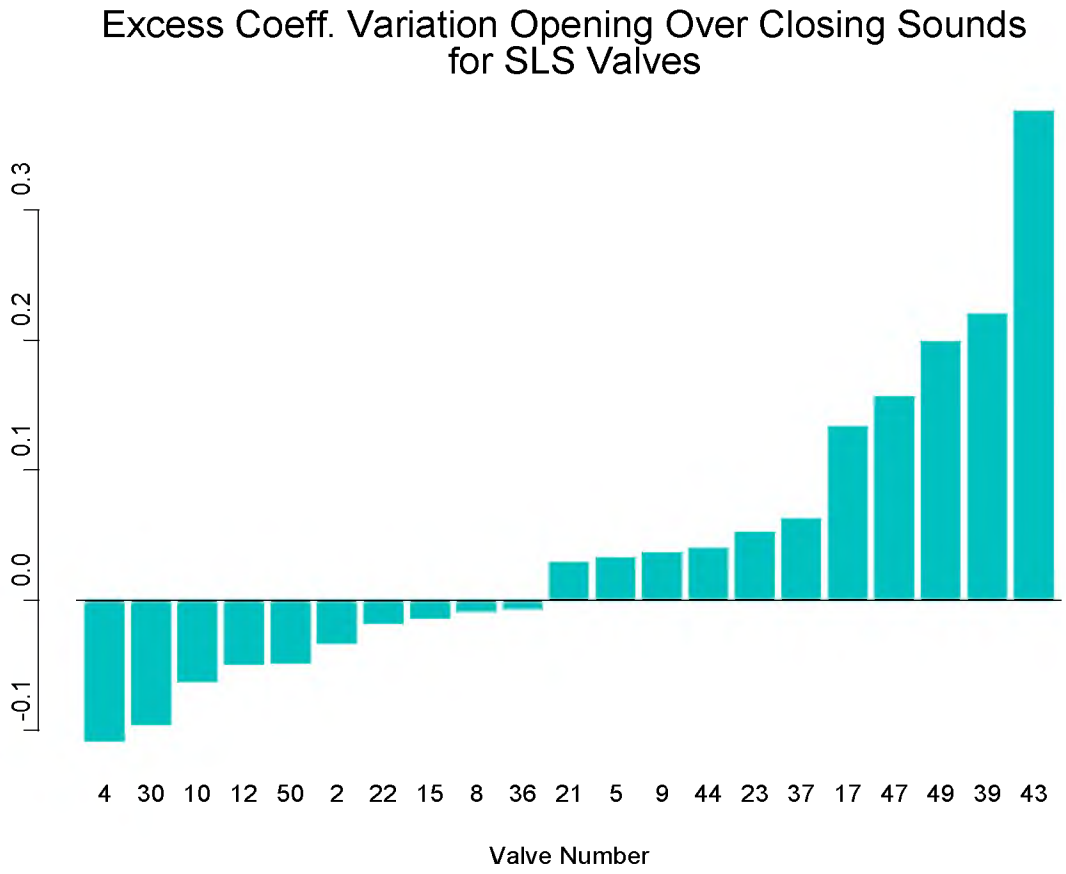


Fig. 14. Excess coefficient of variation over closing sounds for SLS valves

References

[1] J. Candy and F. Barnes, "Heat Valve Processing: A Feasibility Study," Lawrence Livermore National Laboratory, Livermore, CA UCRL-ID-107630, 1991.

[2] J. V. Candy, "**Anechoic Testing Results at the TRANSDEC Evaluation Facility**," Lawrence Livermore National Laboratory, Livermore, CA October 31, 1998 1998.

[3] C. o. E. a. C. Subcommittee on Oversight and Investigations, "The FDA and the Medical Device Industry," US Government Printing Office, Washington DC 1990 1990.

[4] K. C. Pohlmann, *Principles of Digital Audio*. New York: McGraw Hill, 1995.

[5] C. Mullenhoff, "Signal Processing of Shiley Heat Valve Data for Fracture Detection," Lawrence Livermore National Laboratory, Livermore, CA UCRL-ID113760, 1993.

[6] T. D. Plemons and M. Hovenga, "Acoustic Classification of the State of Artificial Heart Valves," *Journal of the Acoustical Society of America*, vol. 97, pp. 2326-2333, 1995.

[7] S. L. Marple, *Digital spectral analysis : with applications*. Englewood Cliffs, N.J.: Prentice-Hall, 1987.

[8] J. P. Burg, "New concepts in power spectra estimation," , 1970.

[9] S. M. Kay, *Modern spectral estimation : theory and application*. Englewood Cliffs, N.J.: Prentice Hall, 1988.

[10] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding robust and exploratory data analysis*. New York: Wiley, 1983.

[11] R. Bellman, *Adaptive Control Processes*. Princeton, NJ: Princeton University Press, 1961.

[12] J. V. Candy, "Personal Communication," .

[13] R. Chia, "Finite element analysis of vibrations of the Bjork Shiley Convexo-Concave heart valve," , 1994.

[14] A. Eberhart, M. Ward, S. Lewardowski, D. Wieting, and R. Inderbitzen, "Relationships Between Closure, Delay, Pressure, Velocity, and Outlet Strut Forces of the Bjork-Shiley Heart Valve," presented at IEEE Conference in Medicine and Biology, 1993.

[15] C. Smilor and S. Schreck, "Study of the Vibrations of the Outlet Strut of the Bjork-Shiley Convexo-Concave Heart Valve," Shiley Heart Valve Center 1994.

[16] S. Crawford and G. Thomas, "In-Vivo Classification of Bjork-Shiley Convexo-Concave Heart Valve from Acoustic Signitures," Lawrence Livermore National Laboratory, Livermore, CA UCRL -ID-114819, 1993.

[17] M. Buhl, G. Clark, J. Candy, and G. Thomas, "Detection of Single-Leg-Separated Heart Valves Using Statistical Pattern Recognition with the Nearest Neighbor Classifier," Lawrence Livermore National Laboratory, Livermore, CA UCRL-ID-114802, 1993.

[18] T. Y. Young and K. S. Fu, *Handbook of pattern recognition and image processing*. Orlando: Academic Press, 1986.

[19] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*. New York: Wiley, 1992.

[20] A. Jain and W. Waller, "On the Optimal Number of Features in the Classification of Multivariate Gaussian Data," *Pattern Recognition*, vol. 10, pp. 365-374, 1978.

[21] G. Clark, ., S. Sengupta, R. Sherwood, J. Hernandez, M. Buhl, P. Schaich, R. Kane, M. Barth, and N. Delgrande, "Sensor Feature Fusion for Detecting Buried Objects," presented at SPIE 1993 International Symposium and Exhibition on Optical Engineering and Photonics in Aerospace and Remote Sensing, Conference on Underground and Obscured Imaging and Detection, Orlando, FL, 1993.

[22] A. Bhattacharyya, "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions," *Bull. Calcutta Math Soc.*, vol. 35, pp. 99-110, 1943.

[23] S. Kullback, *Information Theory and Statistics*. New York.: Dover Publications, 1997.

[24] P. Mahalanobis, "On the generalized Distance in Statistics," *Proc. Nat. Inst. Sciences India*, vol. 2, pp. 49-55, 1936.

[25] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. C-26, pp. 917-22, 1977.

[26] Dawid, "Properties of Diagnostic Data Distributions," *Biometrics*, vol. 32, pp. 647-658, 1976.

[27] D. J. Hand, *Discrimination and classification*. Chichester Eng. ; New York: Wiley, 1981.

[28] E. Parzen, "On the Estimation of a Probability Density Function and Mode," *Ann. Math. Stat.*, vol. 33, pp. 1065-1076, 1962.

[29] B. Silverman, *Density Estimation*. New York: Chapman and Hall, 1986.

[30] R. Lipmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, pp. 4-22, 1987.

[31] D. F. Specht, "Probabilistic Neural Networks," *Neural Networks*, vol. 3, pp. 109-118, 1990.

[32] W. Feller, *Introduction to Probability Theory and Its Applications*. New York: John Wiley, 1968.

[33] W. Highleyman, "The Design and Analysis of Pattern Recognition Experiments," *Bell System Tech. Journ.*, vol. 41, pp. 723-744, 1962.

[34] L. Breiman and P. Spector, "Submodel Selection and Evaluation in Regression: the x-Random Case," *International Statistical Review*, vol. 60, pp. 291-319, 1992.