Chapter 8

Statistical Sonification for Exploratory Data Analysis

Sam Ferguson, William L. Martens and Densil Cabrera

Exploring datasets can be important for discovering new ideas or hypotheses. Sonification research is reviewed, and methods are presented for performing exploratory data analysis. Statistical methods used for visualization are transformed (based on previous research) to the auditory domain to produce statistical sonifications, with example sonifications presented that are based on the iris dataset.

Reference:

Media examples: http://sonification.de/handbook/chapters/chapter8
Chapter 8

Statistical Sonification for Exploratory Data Analysis

Sam Ferguson, William Martens and Densil Cabrera

8.1 Introduction

At the time of writing, it is clear that more data is available than can be practically digested in a straightforward manner without some form of processing for the human observer. This problem is not a new one, but has been the subject of a great deal of practical investigation in many fields of inquiry. Where there is ready access to existing data, there have been a great many contributions from data analysts who have refined methods that span a wide range of applications, including the analysis of physical, biomedical, social, and economic data. A central concern has been the discovery of more or less hidden information in available data, and so statistical methods of data mining for ‘the gold in there’ have been a particular focus in these developments. A collection of tools that have been amassed in response to the need for such methods form a set that has been termed Exploratory Data Analysis [48], or EDA, which has become widely recognized as constituting a useful approach. The statistical methods employed in EDA are typically associated with graphical displays that seek to ‘tease out’ a structure in a dataset, and promote the understanding or falsification of hypothesized relationships between parameters in a dataset. Ultimately, these statistical methods culminate in the rendering of the resulting information for the human observer, to allow the substantial information processing capacity of human perceptual systems to be brought to bear on the problem, potentially adding the critical component in the successful exploration of datasets. While the most common output has been visual renderings of statistical data, a complementary (and sometimes clearly advantageous) approach has been to render the results of statistical analysis using sound. This chapter discusses the use of such sonification of statistical results, and for sake of comparison, the chapter includes analogous visual representations common in exploratory data analysis.
This chapter focuses on simple multi-dimensional datasets such as those that result from scientific experiments or measurements. Unfortunately, the scope of this chapter does not allow discussion of other types of multi-dimensional datasets, such as geographical information systems, or time or time-space organized data, each of which presents its own common problems and solutions.

8.1.1 From Visualization to Perceptualization

Treating visualization as a first choice for rendering the results of data analysis was common when the transmission of those results was primarily limited to paper and books. However, with the rise of many other communication methods and ubiquitous computing devices, it would seem better to consider the inherent suitability of each sensory modality and perceptual system for each problem, and then ‘perceptualize’ as appropriate. Indeed, devices with multiple input interface methods are becoming commonplace, and coordinated multimodal display shows promise when considering problem domains in which object recognition and scene analysis may be helpful.

Friedhoff’s ‘Visualization’ monograph was the first comprehensive overview of computer-aided visualization of scientific data, and it redefined the term: ‘Case studies suggest that visualization can be defined as the substitution of preconscious visual competencies for conscious thinking.’ [28]. Just as is implied here for visualization applications, auditory information display can take advantage of preattentive, hard-wired processing resident in the physiology of the auditory system. Since this processing occurs without the application of conscious attention (it is ‘preattentive’), the capacity of conscious thought is freed up for considering the meaning of the data, rather than cognizing its structure.

Multivariate data provides particular challenges for graphing. Chernoff notably used pictures of faces to represent a data point that varied in multiple dimensions – groups of observations with similar parameters would be seen as one type of face, while different data points would be seen as ‘outsiders’ [13]. Cleveland is commonly cited as providing the classic text on multi-dimensional data representation, as well as being involved with an important visualization software advance (Trellis graphics for S-Plus) [15].

Grinstein et al. [31] discussed the ‘perceptualization’ of scientific data, a term which may be used interchangeably with the more modern definition of ‘visualization’, although it is free of the sensory bias of the latter term. Ware [55] surveys the field of information visualization, a field distinct from scientific visualization, due to the non-physically organized nature of the information being visualized. While scientific visualization may seek to visualize, for instance, the physical shape of the tissue in and around a human organ, information visualization may wish to visualize the relationship between various causes of heart attacks.

8.1.2 Auditory Representations of Data

The auditory and visual modalities have different ecological purposes, and respond in different ways to stimuli in each domain [42]. The fundamental difference is physiological though – human eyes are designed to face forward, and although there is a broad angular range of visibility, the most sensitive part of the eye, the fovea, only focuses on the central part of the visual scene [55], while the ear is often used to monitor parts of the environment.
that the eye is not looking at currently. Eye movements and head movements are necessary
to view any visual scene, and the ears often direct the eyes to the most important stimulus,
rather than acting as a parallel information gathering system.

Auditory display methods have been applied in various fields. One area of widespread usage
is auditory alert design, where auditory design flaws can have strong effects in various critical
situations, such as air traffic control [10], warning sounds [20], and medical monitoring
equipment [45] (see Chapter 19). Much research has focused on sonification for time-series
or real-time monitoring of multiple data dimensions, such as for monitoring multiple sources
of data in an anaesthesia context [23], stock market analysis [39], or EEG signals [36]. These
types of signals are bound to time, and therefore sonification naturally is appropriate as sound
is also bound to time, and expansions and contractions in time can be easily understood.

The early development of auditory data representations was surveyed by Frysinger [30], who
highlights Pollack and Fick’s early experiments [44] which were inspired by the advances
made in information theory. They encoded information in a number of different manners and
measured the bits transmitted by each method. They found that by encoding information in
multiple dimensions simultaneously they were able to transmit more information than if the
information was encoded unidimensionally. Frysinger also mentions Bly’s 1982 work [6, 7],
in which a number of auditory data representations were developed to allow the investigation
of the Iris dataset [1]. Bly tested whether a participant could classify a single multivariate
data point as one of three iris species accurately, based on learning from many representations
of the measurements of each of the three irises (which are described in Section 8.2). Flowers
and Hauer investigated auditory representations of statistical distributions, in terms of their
shape, central tendency and variability, concluding that the information was transmitted
easily using the auditory modality [27]. Later, Flowers et al. discussed an experiment on the
visual and auditory modalities [25], again finding them to be equivalent in their efficacy for
the evaluation of bivariate data correlations. However, Peres and Lane discussed experiments
using auditory boxplots of datasets [43], and found that their respondents did not find auditory
graphs easy to use, and the error rate regarding the statistical information presented did not
improve with training as much as may be expected. They cautioned that this finding did
not necessarily generalize to the entire auditory modality and may have been influenced by
issues to do with the designs of the particular auditory graphs under investigation. Flowers
described how, after 13 years of study and development, auditory data representation methods
are still not common in modern desktop data analysis tools [24, 26].

Sonification has been defined in various ways, initially by Kramer: the process of transform-
ing data to an attribute of sound. Recently, Hermann has expanded this definition, and has
deﬁned sonification in a more systematic manner, as a sound that: reﬂects a) objective prop-
erties or relations in the input data, has a b) systematic and c) reproducible transformation to
sound, and can be d) used with different input data [35].

A common technique for sonification is parameter mapping which requires some kind of
mapping of the data to the element of sound that is to represent it (see Chapter 15). Choosing
that mapping is not a simple task [41], but Flowers [24] describes some simple strategies
that produce useful outcomes, and Walker et al. have carried out fundamental research into
strategies for mapping [52, 50], showing that certain types of data and polarities are more
naturally mapped to particular sound attributes.

The representation of probability distributions has also been discussed by various authors,
including Williamson and Murray-Smith [58, 59] who used granular synthesis as a method of displaying probability distributions that vary in time. Childs [14] discussed the use of probability distributions in Xenakis’ composition *Achorripsis*, and Hermann [33] has investigated the sonification of Monte Carlo Chain Simulations.

Walker and Nees’ research on auditory presentation of graphs has provided a description of data analysis tasks – they delineate trend analysis, pattern detection, pattern recognition, point estimation, and point comparison [53]. Harrar and Stockman described the effect of the presentation of data in discrete or continuous formats, finding that a continuous format was more effective at conveying a line graph overview as the complexity increased, but a discrete format was more effective for point estimation or comparison tasks [32]. De Campo has developed a sonification design space map to provide guidance on the appropriate sonification method (either discrete, continuous or model-based) for representing particular quantities and dimensionalities of data [11].

Hermann [34, 38] has introduced model-based sonification as a method distinct from parameter-mapping, whereby the data set is turned into a dynamic model to be explored interactively by the user, rather than sonifying the data directly. This method provides very task-specific and problem-specific tools to investigate high-dimensional data and is covered in Chapter 16. For a method that deals with large amounts of sequential univariate or time-series data, audification is a common choice, as discussed in Chapter 12. Ferguson and Cabrera [22, 21] have also extended exploratory data analysis sonification techniques to develop methods for sonifying the analysis of sound and music.

Perceptualization practice will gradually reveal when it is best to use auditory representation tools. Auditory representations can potentially extract patterns not previously discernible, and might make such patterns so obvious to the ear, that no-one will ever look for them with their eyes again. By capitalizing upon the inherently different capabilities of the human auditory system, invisible regularities can become audible, and complex temporal patterns can be “heard out” in what might appear to be noise.

8.2 Datasets and Data Analysis Methods

Tukey was possibly one of the first to prioritize visual representations for data analysis in his seminal work *Exploratory Data Analysis* [48]. He focused on the process of looking for patterns in data and finding hypotheses to test, rather than in testing the significance of presupposed hypotheses, thereby distinguishing exploratory data analysis from confirmatory data analysis. Through the use mainly of graphical methods, he showed how datasets could be summarized with either a small set of numbers or graphics that represented those numbers. In some situations (e.g. medical research), a confirmatory approach is common, where a hypothesis is asserted and statistical methods are used to test the hypothesis in a dataset drawn from an experimental procedure. In exploratory situations a hypothesis is not necessarily known in advance, and exploratory techniques may be used to find a ‘clue’ to the correct hypothesis to test based on a set of data. For a set of univariate observations there are several pieces of evidence that exploratory data analysis may find:

the midpoint of the data, described perhaps by the mean, mode or median or through some other measure of the central tendency of the data;
the **shape and spread** of the data, describing whether the data centres heavily on one particular value, or perhaps two or more values, or whether it is spread across its range;

the **range and ordering** of the data, describing the span between the highest and lowest values in the data, and the sorting of the data point values between these two extremes;

the **outliers** in the data, the data points that do not follow the data’s general pattern and may indicate aberrations or perhaps significant points of interest;

the **relationships** between variables in the data, focusing on factors that explain or determine the variability in the data.

Overarching each of these pieces of evidence is the desire to understand any behaviors of, or structure in, the data, to form or discard hypotheses about the data, and to generally gain some kind of insight into the phenomena being observed. As a demonstration dataset the iris measurements of Anderson [1] will be analyzed. For each iris in the set there are four measurements, the sepal’s length and width, and the petal’s length and width. Fifty irises are measured in each of three species of iris, resulting in a total of 150 measurements.

### 8.2.1 Theoretical Frameworks for Data Analysis

Bertin [4, 5], one of the pioneers of interactive data analysis techniques, described a five-stage pattern of decision making in data analysis:

1. defining the problem;
2. defining the data table;
3. adopting a processing language;
4. processing the data, and;
5. interpreting, deciding or communicating.

For data analysis, Bertin developed the permutation matrix, an interactive graphical display that used rearrangeable cards. He argued that all representations of data are reducible to a single matrix (for examples and a review see [17]). Tufte was also heavily influential through his definition of the purpose of graphics as methods for ‘reasoning about data’, and highlighting of the importance of design features in the efficiency of information graphics [46, 47]. He draws a distinction between graphics whose data is distorted, imprisoned or obfuscated – and graphics which rapidly and usefully communicate the story the data tells. ‘Above all else, show the data’, is a maxim that shows his emphasis on both the quantity and priority of data, over ‘non-data ink’ – complex scales, grids or decorative elements of graphs, although Bateman et al. have shown that some types of decoration (such as elaborate borders, cartoon elements and 3-dimensional projections) can enhance a graphic’s efficiency [3].

Another theoretical framework for statistical graphics that has already had far-reaching influence is Leland Wilkinson’s *Grammar of Graphics* [57], which has been implemented in a software framework by Wickham [56]. This is one of the few conceptual frameworks to take a completely systematic object-oriented approach to designing graphics.

Ben Fry’s *Computational Information Design* [29], presents a framework that attempts to link fields such as computer science, data mining, statistics, graphic design, and information visualization into a single integrated practice. He argues a 7-step process for collecting,
managing and understanding data: 1. acquire, 2. parse, 3. filter, 4. mine, 5. represent, 6. refine, 7. interact. Crucial to this framework is software that can simplify the implementation of each operation, so that a single practitioner can practically undertake all of these steps, allowing the possibility of design iteration to incorporate many of the stages, and facilitating user interaction through dynamic alterations of the representation.

Barrass discusses the ‘TaDa’ design template in ‘Auditory Information Design’, a Ph.D. thesis[2]. The template attempts to delineate the Task requirements, and the Data Characteristics. An important element of this design framework is that Barrass categorizes different types of data into a systematic data type and auditory relation taxonomy based on information design principles.

8.2.2 Probability

Some fundamental concepts are necessary in a discussion of statistical sonification. Statistical probability is a method of prediction based on inference from observation, rather than from induction. Data are in general understood as samples drawn from an unknown high-dimensional probability distribution. In one definition, Bulmer [9] describes a statistical probability as ‘...the limiting value of the relative frequency with which some event occurs.’ This limiting value may be approximated by repeating an experiment $n$ times and comparing that number with the number of times event $A$ occurs, giving the probability $p(\text{A})$ clearly as $\frac{n(A)}{n}$. As $n$ increases, in most situations $p(\text{A})$ moves closer to a particular value, reasonably assumed to be the limiting value described above. However, as no experiment may be repeated infinitely, the reasonableness of this assumption is strongly associated with the number of times the experiment is performed ($n$), and we can never know with absolute certainty what this limiting value is [9]. Using statistical probability, the way we infer that we have approximately a 50% chance of getting either a head or a tail when we toss a coin is by tossing that coin repeatedly and counting the two alternatives, rather than by inducing a probability through reasoning about the attributes of the coin and the throw.

8.2.3 Measures of Central Tendency

Once a set of observations has been obtained they can be quickly summarized using measures of their central tendency. The midpoint of a dataset is crucial and can be described in many ways, although a parsimonious approach is to use the median. The median is the middle point of the ranked data, or the mean of the two middle points if the dataset has an even count. The median implicitly describes the value at which the probability of a randomly drawn sample will fall either above or below it is equal. The arithmetic mean ($\bar{x}$) is another measure of central tendency that is useful for describing the midpoint of a set of data. It can be described mathematically as:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $n$ observations are termed $x_1$, $x_2$, ..., $x_n$ [9]. Other methods for calculating the mean exist, usually used when the numeric scale is not linear: including the geometric mean and the harmonic mean. The mode is the measurement that is observed the most times in a
discrete distribution, or it is the point where the probability density function has the highest value in a continuous distribution.

### 8.2.4 Measures of Dispersion

Measures of dispersion allow us to build a more detailed summary of how the distribution of a dataset is shaped. Sorting a set of data points is an initial method for approaching a batch of observations. The top and bottom of this ranking are the *maximum* and *minimum*, or the extremes. The difference between these numbers is the *range* of the data. To provide a number that represents the dispersion of the distribution, one may take the mean and average all the absolute deviations from it, thus obtaining the mean absolute deviation. The standard deviation $\sigma$ is similar, but uses the square root of the mean of the squared deviations from the mean, which makes the resulting number more comparable to the original measurements:

$$
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}
$$

(2)

To provide more detail on the shape of the distribution we can use two more values. The central portion of the distribution is important to work out the spread of the data, and to summarize it we divide the data into 2 parts, the minimum to median, and the median to maximum. To find the 25th and 75th percentiles we then take the median of these 2 parts.
(a) A dotplot of the 150 measurements of Petal Length shows clear groupings.

(b) A histogram can show the general shape within the Petal Length data.

(c) A cumulative distribution function of the Petal Length measurements.

(d) A kernel smoothed estimation of the probability density function.

Figure 8.2: Various representations of a dimension of a dataset.

again. These numbers are also known as the 1st and 3rd quartiles, and the range between them is known as the interquartile range. A small interquartile range compared with the range denotes a distribution with high kurtosis (peakedness).

Tukey’s five number summary presents the interquartile range, the extremes and the median to summarize a distribution, allowing distributions to be easily and quickly compared [48]. Tukey invented the boxplot by taking the five-number summary and representing it using a visual method. It uses a line in the middle for the median, a box around the line for the 25th to 75th percentile range (the inter-quartile range), and whiskers extending to maximum and minimum values (or sometimes these values may be the 95th and 5th percentile). Considering the batch of data as a shape, rather than a set of (five) numbers, can show the characteristics of the data distribution more clearly, especially if the distribution is not a typical unimodal bell shape (see Figure 8.1). A graph of the distribution demonstrates the ideas of the range of the data, the midpoint of the data, but also clearly shows well-defined aspects such as skew and kurtosis, as well as aspects that are not easily described and may be peculiar to the particular dataset being investigated.

Figure 8.2 shows four representations of the Petal Length variable in the iris dataset: a dotplot, a histogram, a cumulative distribution function, and a kernel density function. A histogram is a simple way of visualizing a distribution of data by using a set of bins across the data range and representing the number of observations that fall into the bin. A cumulative
Figure 8.3: A kernel-smoothed estimation of the probability density function, grouped by the species of iris measured.

distribution function is another description of the same information. It is a graph of the probability (on the y-axis) of a random choice from the dataset being less than the value specified on the x-axis. A smoother representation of the probability density function may be obtained through the technique of kernel smoothing [54]. In this technique the distinct observations at each point are replaced by kernels, miniature symmetric unimodal probability density functions with a specified bandwidth. These kernels are then summed across the data range to produce a curve.

The distribution revealed by the kernel-smoothed probability density function (Figure 8.2(d)) does not exhibit the sort of unimodal shape that might be expected if the 150 data points were all sampled from a homogeneous population of iris flowers. Rather, there seems to be some basis for separating the data points into at least two groups, which is no surprise as we already know that multiple species of iris were measured. Although pattern classification could be based upon the ground truth that is already known about these data, it is also of particular interest whether membership of a given item in one of the three iris groups might be determined through exploration of the data. Before examining that possibility, the ground truth will be revealed here by dividing the single category into three categories based on the known iris species (Figure 8.3). The three color-coded curves in the figure show that there is one clearly separated group (graphed using a red curve), and two groups that overlap each other (the blue and green curves). What might be considered to be more typical of an exploratory data analysis process, and that which will be examined in greater depth from this point, is the attempt to assign group membership to each of the 150 items in a manner that is blind to the ground truth that is known about the iris data. That is the topic taken up in the next subsection.
8.2.5 Measures of Group Membership (aka Blind Pattern Classification)

Blind pattern classification is the process by which the items in a set of multivariate data may be sorted into different groups when there is no auxiliary information about the data that would aid in such a separation. What follows is an example based upon the four measurements that make up the iris dataset, even though the ground truth is known in this case.

The simplest approach to pattern classification would be to apply a hard clustering algorithm that merely assigns each of the 150 items into a number of groups. Of course, the number of groups may be unknown, and so the results of clustering into two, three, or more different groups may be compared to aid in deciding how many groups may be present. The most common approach to hard clustering, the so-called K-means clustering algorithm, takes the hypothesized number of groups, K, as an input parameter. If we hypothesize that three species of iris were measured, then the algorithm will iteratively seek a partitioning of the dataset into these three hypothetical groups. The process is to minimize the sum, over all groups, of the within-group sums of the distances between individual item values and the group centroids (which capture the group mean values on all four measurements as three unique points in a four-dimensional space). Of course, the K-means clustering result measures group membership only in the nominal sense, with a hard assignment of items to groups.

A more useful approach might be to determine how well each item fits into each of the groups, and such a determination is provided by a fuzzy partitioning algorithm. If again we hypothesize that three species of iris were measured, fuzzy partitioning will iteratively seek a partitioning of the dataset while calculating a group membership coefficient for each item in each of the three groups. Hence no hard clustering is enforced, but rather a partitioning in which membership is graded continuously, and quantified by three group membership coefficients taking values between 0 and 1. The result of a fuzzy partitioning of the iris measurements is shown in Figure 8.4. In the graph, group membership coefficients for all 150 items are plotted for only two of the three groups, termed here the red and the green (to be consistent with the color code used in the previous figure). Since the group membership coefficients sum to 1 across all three groups for each item, the value of the blue membership coefficient for each item is strictly determined by the values taken for the other two coefficients. To help visualize the 150 items’ continuously-graded membership in the three groups, the plotting symbols in the figure were color-coded by treating the red, green, and blue group membership coefficients as an RGB color specification. Naturally, as the group membership values approach a value of 1, the color of the plotting symbol will become more saturated. The items fitting well in the red group are thus quite saturated, while those items that have highest blue-group membership values are not so far removed from the neutral grey that results when all three coefficient values equal 0.33.

Of course, the red group items can be classified as separate from the remaining items strictly in terms of the measurement on just one column of the four-column matrix that comprises the iris dataset, that column corresponding to the Petal-Length variable. However, the distribution of Petal-Length measurement values, which was examined from several perspectives above, does not enable the separation of the green and blue group items. Other methods that are based on multivariate analysis of the whole dataset may provide better results, as is discussed in the next subsection.
8.2.6 Multivariate Data Exploration

It would be reasonable to assume that a multivariate analysis of the whole iris dataset might prove more effective in separating the 150 items into the above-hypothesized three groups, especially in comparison to an examination of the distribution of measurement values observed for a single variable. However, measurements for additional variables will only make substantial contributions in this regard if the values on those additional variables provide independent sources of information. As a quick check on this, a visual examination of two measurements at once will reveal how much independent information they might provide. Figure 8.5(a) is a scatterplot of measurement values available for each of the 150 items in the iris dataset on two of the variables, Petal-Width and Petal-Length. (Note that the plotting symbols in the figure are color coded to indicate the three known iris species that were measured.) It should be clear that there is a strong linear relationship between values on these two variables, and so there may be little independent information provided by measurements on the second variable. The fact that the two variables are highly correlated means that a good deal of the variance in the data is shared, and that shared variance might be represented by a projection of the items onto a single axis through the four-dimensional space defined by the four variables. The multivariate analytic technique that seeks out such a projection is Principal Component Analysis (aka PCA, see [19]).

PCA effectively rotates the axes in a multivariate space to find the principal axis along which the variance in the dataset is maximized, taking advantage of the covariance between all the variables. The analysis also finds a second axis, orthogonal to the first, that accounts for the greatest proportion of the remaining variance. In the case of the iris data, the scores calculated for each of the 150 items as projections onto each of these two axes, called principal component scores, are plotted in Figure 8.5(b). Scores on Principal Component 1 (PC1) separate the three groups well along the x-axis; however, scores on Principal Component 2 (PC2) do very little to further separate the three groups along the y-axis of
Figure 8.5: Principal Component Analysis rotates the axes of raw data in an attempt to find the most variance in a dataset.

the plot. This means that the blue group and green group items that inhabit the region of overlapping values would be difficult to classify, especially because in this exploration we assume that the species differences are not known. If it were a case of machine learning, in which the ground truth about the three iris species measured were known, rather than the exploratory data analysis that is under discussion here, then a linear discriminant analysis could be performed that would find a more optimal means of separating items from the three species [18]. There is some real value in evaluating the success of exploratory methods by comparing known categorization with categorization discovered blindly; however, for the introduction to statistical concepts that was deemed most relevant to this chapter, no further examination of such multivariate analytic techniques will be presented. Suffice it to say that the relative merits of visual and auditory information displays made possible employing such multivariate approaches are worth investigating in both exploratory and confirmatory analyses, though it is the former topic of exploratory sound analysis to which this chapter now turns.

8.3 Sonifications of Iris Dataset

This section presents several auditory representations of the iris dataset examined above from the visual perspective, and attempts to show the relative merit of the various univariate and multivariate representations that can be applied for exploratory data analysis. In this first subsection, the dataset is examined in a univariate manner, and this shows that groupings of items can be distinguished on the basis of their petal size measurements when sonically rendered just as when they are visually rendered (Figure 8.6).
8.3.1 Auditory Dotplot

Exploratory Data Analysis has a strong emphasis on the investigation of ‘raw’ data initially, rather than immediately calculating summaries or statistics. One visual method that can be used is the ‘strip plot’ or ‘dotplot’. This presents one dimension (in this case, the measurements of Petal-Length) of the 150 data points along a single axis that locates the measured dimension numerically. This display is not primarily used for analytic tasks, but it possesses some characteristics that are very important for data analysis – rather than showing a statistic, it shows every data point directly, meaning the numerical Petal-Length values could be recreated from the visualization. This directness helps to create a clear impression of the data before it is transformed and summarized using statistical methods, and skipping this stage can sometimes be problematic.

A sonification method that mimics the useful attributes of a dotplot may be constructed in a variety of manners. It may map the data dimension to time, and simply represent the data through a series of short sounds. This sonification (sound example S8.1) scans across a data dimension using the time axis, rendering clusters, outliers and gaps in data audible. The use of short percussive sounds allows for a large number of them to be heard and assessed simultaneously.

8.3.2 Auditory Kernel Density Plot

In Figure 8.6 the kernel density plots additionally include a dotplot at the base of each set of axes. As described in section 8.2.4 this dotplot is summed (as kernels of a specified bandwidth) to produce the curve overlaid above it. A kernel density plot is very similar to a histogram in that it attempts to show the distribution of data points, but it has a couple of differences. It employs a more complex algorithm in its creation, rather than simply counting observations within a set of thresholds, and the units it uses are therefore not easily parsed – but it does create a curve rather than a bar-graph [54]. A histogram, by comparison, simply counts the number of data points within a number of bins and presents the bin counts as a bar graph. This means the algorithm is straightforward, and the units easy to understand, but the resolution can be fairly poor depending on the choice of bin width. In the auditory modality, however, a sonification algorithm can be simpler still.

A kernel density plot in the auditory modality can be created by mapping the sorted data points to a time axis based on the value of the data (with the data range used to normalize the data point’s time-value along the time axis). This sonification achieves a similar type of summing to the kernel density plot, but in the auditory modality, through the rapid addition of multiple overlapping tones. The outcome is that groupings, and spaces between groupings, can be heard in the data sonification. This graphing technique describes the distribution of data, rather than a single summary of the data, and is easily extended to the auditory modality.

With so many overlapping data points we could give some consideration to the phase cancellation effects that might occur. Indeed, if many notes of identical frequencies were presented at the same time, some notes might increase the level more than the 3 dB expected with random phase relationships. However, it is very unlikely that two notes would be presented that have identical frequencies and spectra but are 180 degrees out of phase and
thereby cancel. A more likely scenario, with complex tonalities with temporal envelopes (created for instance using FM), is that the notes would be perceived as overlapping, but as two distinct elements – unless of course the temporal onset was simultaneous. The graphical technique that is used to avoid a similar effect is to use a specified amount of random jitter to ensure that graphical markers do not sit in exactly the same place, which is probably appropriate in this situation as well.

Sound example S8.2 is an example of the Petal-Length data presented as an auditory kernel density plot. The first grouping is heard clearly separated from the other data points, and we can also make a rough guess of the number of data points we can hear. This first group happens to be the *setosa* species of iris, while the larger, longer group is made up of the two other iris species.

### 8.3.3 Auditory Boxplot

A boxplot is another common method for comparing different distributions of data (instead of assessing a single distribution as do the kernel density and histogram). Flowers et al. have discussed the representation of distributions through the use of boxplots (or arpeggio plots for their auditory counterpart) [26]. One way to represent these is to play each of the five numbers that form the boxplot in succession forming a type of arpeggio. Another method of summarizing the data is to randomly select and sonify data points from the dataset rapidly, creating a general impression of the range, density, and center of the dataset almost simultaneously. There is no need to stop sonifying when the number of data points has been reached, and selections could continue to be made indefinitely. This means that a very high rate of sonification can be achieved, summarizing the group very quickly, resulting in a stationary sonification of arbitrary length. This is very important when multiple groups of data are compared as the speed of comparison (and therefore the memory involved) is determined by the time it takes for each group to be sonified– using multiple short sonifications of each group can be done quickly using this method.
By choosing the range of data points selected through sorting the data and specifying a range, it is possible to listen to the interquartile range of the data, or only the median, maximum or minimum. These are useful techniques for comparing different groupings. This dataset can be sorted by the species of iris, into three groups of 50 flowers each. By using this sonification we can either summarize a group with a single measure of location, such as the median, or describe the distribution. By concatenating the sonifications of each of the groups we can compare the measures against each other rapidly, allowing an estimate of whether the groups can be readily separated. Figure 8.7 shows a traditional boxplot of the Petal-Length dimension of the dataset, which can be compared to sound example S8.3 for the sonification method.

![Figure 8.7: Boxplots represent distributions of data, and when different data distribution boxplots are presented adjacently they can be easily compared.](image)

This sonification separates the three groups into different ‘blocks’ of sound, with successive narrowing of the range presented, from the full 95th to 5th percentile to 75-25 percentile and then the median only. Hence, this represents the range, interquartile range, and then the median of the data in succession. It is clear from the sonification as well as the graphic that the first group is relatively narrow in distribution, but the second and third are quite wide, and not as well separated.

### 8.3.4 Auditory Bivariate Scatterplot

Two-dimensional approaches can provide more differentiation than those using one dimension. In Figure 8.5(a) we see a bivariate scatterplot that compares two parameters. Using parameter mapping, we sonify the data again (sound example S8.4), this time using FM-synthesis (as introduced in chapter 9) to encode two data dimensions in one note. The petal length parameter is mapped to the pitch of the tone, while the petal width parameter is mapped to the modulation index for the tone. This method of representation allows the user to listen to the data and build an auditory representation quickly, internalizing it like
an environmental sound for later recall. Single sounds can then be compared against the internalized sound for comparison and classification.

8.3.5 Multivariate Data Sonification

As suggested in the above discussion on visualization of the iris dataset, it has been assumed that a multivariate exploration of the whole iris dataset might prove more effective in separating the 150 items into three groups than would a univariate examination of measurement values. The same question should be asked here in comparing univariate, bivariate, and multivariate sonifications of the iris data.

The initial investigation of this idea is a demonstration of the difference between a Petal-Length sonification and some bivariate sonifications. The first bivariate sonification is strictly analogous (sound example S8.4) to the visual graph shown in Figure 8.5(a), the scatterplot of measurement values available for each of the 150 items in the iris dataset on Petal-Width and Petal-Length. Although there is some linear dependence between values on these two variables, there does seem to be a substantial amount of independent information provided by measurements on the second variable. The second bivariate sonification (sound example S8.5) is analogous to the visual graph shown in Figure 8.5(b), which plotted scores for each item on the first two principal components that were found when the iris dataset was submitted to PCA. As the PCA rotated the axes to maximize variance along the principal axis, and then to maximize the remaining variance along a second orthogonal axis, two auditory attributes of graded perceptual salience were applied in a PC-based sonification (as suggested by Hermann [34]). Comparing this PC-based sonification to the direct two-parameter sonification also presented here does not provide a convincing case for an advantage given the PC-based approach. To most listeners, the PC-based mapping does not produce a sonification that makes the distinction between groups any more audible than did the more straightforward two-parameter sonification. Might it be that correlation between item values on the original variates, when mapped to distinct sonification parameters, could be relatively more effective in the case of presenting the iris data, despite the inherent redundancy this would display?

Therefore, a four-dimensional sonification was created using measurement values on all four of the original variables (sound example S8.6). This sonification is related to the visualization using Chernoff’s [13] faces that is illustrated in Figure 8.8. The sonification is not strictly analogous to the visualization, however, since the sonification allows individual items to be displayed in rapid succession, and also allows many repeat presentations to aid the observer in forming a concept of how sounds vary within each group of 50 items. A comparable succession of faces could be presented, but the more typical application of Chernoff’s [13] faces is to present many faces simultaneously at plotting positions that are spread out in space. It may be that the opportunity to visually scan back and forth presents some distinct advantage over the strict temporal ordering of item sonifications followed in the sound example, but is is difficult to argue for a more general conclusion outside the context of the investigation of a particular dataset. Indeed, the authors are not aware of any empirical evaluation of the relative effectiveness of spatial vs. temporal distribution of faces for human discrimination of patterns in visualized data, let alone sonified data.

The four-dimensional sonification created for this discussion is not the first such sonification
Statistical Sonification for Exploratory Data Analysis 191

Figure 8.8: A Chernoff-face visualisation of the mean values for each of three groups of iris measured on four variables. Underlying each of the three Chernoff faces can be seen a color-coded patch illustrating the convex hull of principal component scores for the 50 items in each of the three groups contributing to the mean values visualized. Although the faces are constructed to show the four mean parameters as a four-dimensional visualization, each is positioned over the centroid of these patches in the bivariant plotting space to show the spread of the data in each of the three groups.

created to represent the iris dataset. In fact, about 30 years ago Sara Bly [7] presented a closely related sonification that was developed during her doctoral research. Her sonification used the following mapping from iris measurement variables to sound synthesis parameters: Variation in petal length was mapped to duration, petal width was mapped to waveshape, sepal length to pitch, and sepal width to volume. After a training session in which listeners heard example sonifications representing each of the three iris species, they were presented with 10 test items that could be more or less well classified. She reported that most casual observers could place her sonifications into the appropriate groups with few errors. No formal empirical evaluation of the success of the current sonification was undertaken, but informal listening tests suggest that similar good performance would be expected using the mapping chosen here, which is summarized as follows: The first two parameters, Petal-Length and Petal-Width, were mapped to the most elementary auditory attributes, pitch and duration, while the third and fourth parameters were mapped to timbral attributes, perhaps providing more subtle indication of parameter variation than the mappings for the first two parameters. These two timbral attributes could be heard as variations in tone coloration similar to that of vowel coloration changes characteristic of human speech. More specifically, the dataset values were mapped so as to move the synthesized tones through the vowel space defined by the first two formants of the human vocal tract, as follows: The measured Sepal-Length values modulated the resonant frequency of a lower-frequency formant filter, while Sepal-Width values were mapped to control the resonant frequency of a higher-frequency formant.
filter. Applying this co-ordinated pair of filters to the input signals that varied in pitch and duration resulted in tones that could be heard as perceptually rich and yet not overly complex, perhaps due to their speech-like character.

8.4 Discussion

Different priorities exist in the auditory modality, and an important difficulty for making sonification design decisions is the purpose of the representation. Sonification can be used for enabling blind and visually impaired people to use data representations, for ‘ears-only’ representations in high-workload monitoring environments, for multi-modal displays, and for representations where the auditory modality is more efficient than the visual.

Sonification is an important part of the process of developing next-generation data representations. The data representations described in this chapter could be used in isolation, but would also be appropriate for inclusion in a multi-modal interactive interface, to redundantly encode the information for better description. As data analysis slowly moves off the page and onto the computer screen, touch-screen or mobile phone, interfaces of this nature will become more important.

Interactivity is key to a high-quality representation, and sonification of many types benefits immensely from a strong interactive interface. Many of the representations described are designed as a single-pass presentation of an overview of the data, although many information visualizations employ an interactive zoom-and-filter technique. This technique could be appropriate for the interactive control of sonifications. The speed at which the data is presented is one of its strengths, allowing the sonification to be started and repeated rapidly, and therefore capable of responding to interaction in a time-appropriate manner while remaining audible. Natural user interfaces employing multi-touch technology have now appeared in many human-computer interaction situations, and sonification research has already started to address priorities associated with this form of data interface [37, 8, 49]. Interfaces such as these may be predecessors to widespread use of sonification exploratory data analysis.

8.4.1 Research Challenges or Continuing Difficulties?

Data sonification is possibly not as common as it could be for a few reasons. One that has been mentioned by many authors is logistical in nature – many practitioners have little access to, or cannot easily use, sonification software. While much of the available software (such as the sonification sandbox [51], or SonEnvir [12]) is free and easily available, it does not necessarily fit within a typical data analysis workflow, and is non-existent within most major statistical packages or spreadsheet programs [26].

Another logistical problem is that sonification is often time-bound, while data representations of all types usually need to be scanned in a non-linear fashion, as the user seeks to build a conception of the relationships between different points by making comparisons between data points, axes and labels. The eye is able to quickly move between multiple elements in the graph obtaining various pieces of information. Where the time-axis is employed as a mapping axis in a sonification, the sonification must be replayed and the particular elements
to be compared must be listened for. This is analogous to taking a visual graph, wrapping it onto a cylinder and reading it by rotating the cylinder. The eye has a great advantage over the ear in this regard, as it is capable of scanning non-linearly, in a way that the ear cannot. Of course, it is not necessarily problematic that sonifications are time-bound, since the sensitivity of the human auditory system to variations in temporal patterns can be quite acute under many circumstances. For example, in highly interactive sonification interfaces, a variety of alternatives to the linear presentation format are available, and these user-driven interactive explorations can reveal the benefits of the auditory modality, with significant improvements resulting from the possibility of listening and re-listening, comparing and re-assessing the presented data (see Chapter 11 for more discussion of interactive sonification).

Open questions remain - there are not many methods that exist for sonification of datasets that simultaneously present more than five or six dimensions. This difficulty exists in the visual domain, and is often solved through multiple views of the same data presented and interacted with simultaneously (see ggobi for instance [16]) – an auditory analogue faces obvious difficulties as the ear cannot ‘shift its gaze’ as easily as the eye. With careful interaction design however, similar or complementary possibilities may prove possible for sonification. For very large datasets the mapping of distinct auditory elements to each data dimension is not practical, and future research may investigate possible methods for developing methods that scale well for large numbers of data dimensions. Common methods used in the study of genetic data, for instance, use large permutation matrices in the form of a heat map, with sophisticated methods of interaction to highlight differences or outliers in the data (see Fry [29], Chapter 4 for a discussion). No auditory analogue yet exists that does not use some form of data reduction, but the auditory sense’s capability for processing large amounts of information seems well suited for this type of data.

Also, visualization can provide numerical indicators of values (e.g., numbers on axes), while it is difficult for sonification to be so specific. Auditory tick-marks (including defining the start of the sonification, time 0), and exploiting either physical or psychoacoustical scales can help, but in general, statistical sonification will often provide a general sense of the data distribution without providing the user with access to specific numeric values.

### 8.5 Conclusion and Caveat

This chapter has described previous and current methods for representing multivariate data through statistical sonification for the purposes of exploratory data analysis. It must be said that the current state of the art must be considered to be quite immature as yet, with many challenges for sonification research to tackle in the future. In fact, it might be proposed that the best approach to take in designing and developing statistical sonifications in particular would be one that includes critical evaluation of the results at each attempt. Indeed, in the early development of scientific visualization methods, such a summary of practical case studies did appear in the published collection entitled ‘Visual cues: Practical data visualization’ [40]. Perhaps the sonification case studies that are presented in the handbook in which this chapter appears provide a useful beginning for such an endeavor.

In the absence of a well-established paradigm representing the consensus of practitioners in this field, a strategy might be taken in which the effectiveness of sonifications, in contrast to visualizations, would be put directly under test, so that ineffective sonifications could most...
easily be rejected. Through such rigor practitioners may become confident that their attempts have real value; yet without such rigorous evaluation, less useful sonification approaches may be accepted as worthy examples to be followed, before they have been adequately examined.

Bibliography


Statistical Sonification for Exploratory Data Analysis 195


