Proceedings of the Institute of Acoustics

APPLICATION OF SCENE ANALYSIS PRINCIPLES TO SONIFICATION

Jon Barker & Martin Cooke

University of Sheffield, Department of Computer Science, Sheffield, England.

1 INTRODUCTION

The basic premise of sonification is that the sophisticated temporal pattern processing facilities of human listeners might provide a means to extract salient cues from multi-dimensional data sets [8]. In this paper, we suggest that consideration of listeners' propensity to group and to segregate sound components is required for effective sonification. We elaborate some general principles for mapping multi-dimensional data sets on to sound, taking into account knowledge of auditory scene analysis. As a demonstration of these ideas, we have developed a software simulation of traffic flow in an arbitrarily-complex network, and have used sonification to present listeners with an aural image of this domain.

The technique of representing quantitative data using sound, termed 'sonification', is more precisely defined as follows:

> "a mapping of numerically represented relations in some domain under study to relations in an acoustic domain for the purposes of interpreting, understanding, or communicating relations in the domain under study [13]."

The data, which may be derived from physical measurements of the domain under study, or generated by a computer simulation or model, is transformed into sound, using a mapping which may involve any level of abstraction. The data may be employed directly to define a sound wave, which is sometimes termed 'auralization', or used with or without preprocessing, to drive the controls of some sound synthesis technique.

The earliest systematic study of sonification was conducted in 1954 by Pollack and Ficks [12], who compared the information transmitted by schemes employing varying dimensionality and resolution. More recent work has included the auditory analysis of seismograms [15], the use of sound for the analysis of multivariate, time-varying and logarithmic data [3], and studies comparing auditory and visual displays [7][11]. There has also been auditory display research motivated by the needs of the visually impaired, such as Lunney and Morrison's auditory interface for an infrared spectrograph [9], and Mansur's soundgraphs [10].

Sonification research blossomed at the end of the 80's, with Kramer's work on the Clarity Sonification Toolkit [8], and Smith's work on Exvis [14], a tightly coupled auditory and visual display tool. 1992 saw the first international conference on auditory display, ICAD 92 [8] bringing together 36 researchers of extremely diverse backgrounds, presenting papers largely concerned with sonification, ranging from theoretical discussions of underlying issues to applied work presenting solutions to specific problems.

The following example, taken from work conducted by Fitch and Kramer [6], should serve to illustrate the technique. Subjects were trained to play the role of anesthesiologists attempting to keep a computer simulated patient alive, monitoring eight vital signs and responding to a series of operating room emergencies. The patient's heart rate was mapped onto the rate of repetition of a pair of tones which were designed to sound similar to a heart-beat. The rate of respiration was mapped to the rate of amplitude modulation of a band-passed noise, simulating a breathing sound. Other more abstract variables were used to modulate these heart and respiratory base sounds; for example, carbon-dioxide levels controlled the brightness of the heart sound, and blood pressure controlled its pitch. Subject performance was assessed using both the standard strip charts and the sonification. Although subjects generally reported feeling initially more confident with the labelled visual displays, after some practice most showed better results using the 'auditory display.' It was concluded that in complex dynamic systems the ability of the auditory system to perceive a

number of variables simultaneously proved advantageous compared to the one-by-one perception necessitated by the visual display. From this simple example it is not hard to see the large potential of the technique, and how it may benefit a vast number of data monitoring and analysis tasks that have traditionally been conducted visually.

2 CAN SONIFICATION IGNORE SCENE ANALYSIS?

In contrast to the rapid development of sound-producing technology that has occurred over the last few decades, the auditory system, used to process the output of this technology, has evolved over many millennia. It has developed to perform a very specific task; that is, detecting naturally created air-pressure waves and analyzing them in order to extract information about the sources in the environment that created them. These sound sources adhere to physical laws and hence there are constraints placed on the nature of the auditory signals they can generate.

In the natural acoustic environment, at any one moment, the pressure waves arriving at our ears arise from the mixture of many simultaneous sources. One key task of the auditory system is to recover the individual descriptions of the separate sources so that they can be identified and located in space. Bregman terms this *auditory scene analysis* [1][19]. Using his terminology the physical entity which give rise to the acoustic events is called a *source*, and the perceptual representation of the events is called a *stream*. A great deal is now understood about the principles the auditory system applies to perform this scene analysis and several successful computer models have been designed [2][4][16].

The theory demands that this level of perceptual processing is applied to all auditory signals, whether they be natural environmental sounds or those generated artificially. The implication is that, whether we wish our sonifications to be intricately complex with multiple simultaneous voices or simply to contain all the information in a single voice, we should take scene analysis into consideration. Common experience of naturally occurring sounds may encourage the belief that there is necessarily a direct correspondence between the output of the various sound synthesis algorithms employed in a sonification and the auditory streams perceived by the listener; however since artificially generated sounds are not subject to the same constraints as naturally occurring sounds, this is not the case. For example, if coding a single time-varying variable as a sequence of pure tones, using a pitch range that is too large or a presentation rate that is too fast may cause the sequence to fragment into several perceptual streams. Whether this lack of coherence significantly effects the efficacy of the sonification is an issue that needs addressing and will no doubt depend on many factors, such as the complexity of the data and the type of task that is being performed.

3 A SONIFICATION CASE STUDY: URBAN TRAFFIC CONTROL

To provide a test-bed for sonification design automatic urban traffic control systems were considered as a suitable data domain. Traffic control systems produce high dimensional time-series data, the structure of which is not easy to ascertain visually. They also present interesting unsolved classification and prediction problems in the guise of the recognition and forecasting of traffic congestion[5].

The difficulty with traffic congestion detection is that the definition of 'congested' is highly subjective and depends largely on the context of the road link. If the parameters of an uncongested link in a town centre, such as average queue length, percentage occupancy and so on, are transferred to a link elsewhere, then the same parameters might indicate very severe congestion. Additionally, the examination of a single link in isolation of the rest of the network is not sufficient for the diagnosis or control of congestion. As it is impossible to control the flow on one link without affecting surrounding links, operators usually have to consider the state of a whole sub-area. To compound the problem further there is no single parameter, such as vehicle flow or occupancy, that can by itself act as a litmus test for congestion. Traffic controllers require much ex-

perience and they have to assimilate visual information from several sources, including CCTV screens and the output of various traffic detectors, in order to form their decisions. This is clearly an area in which sonification could be of potential benefit.

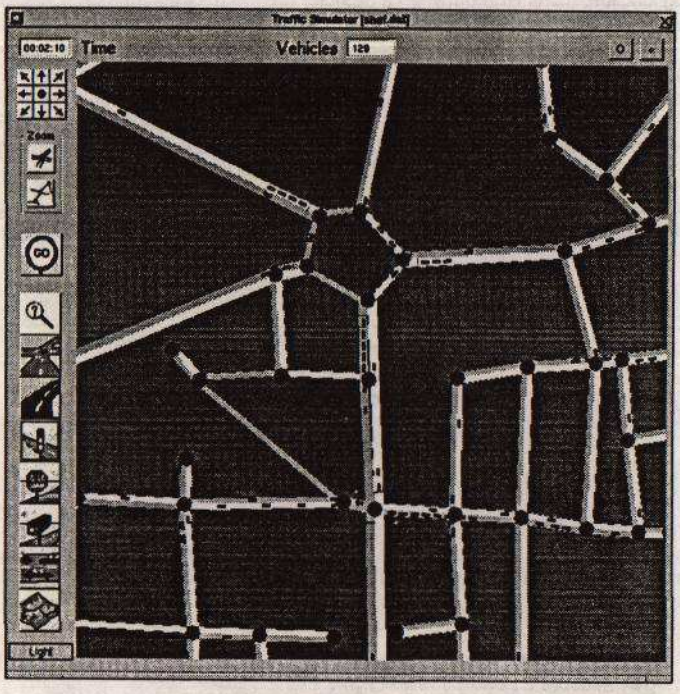

FIGURE 1. A traffic simulation of the university area of Sheffield

3.1 TRAFFIC SIMULATION

Data from the SCOOT traffic control system established at Leicester was examined initially. However, due to constraints imposed by the rather sparse network of traffic detectors, it was decided that, for increased flexibility, simulated traffic flow data would be used instead. The system that we have developed at Sheffield is a microscopic object-oriented simulation, developed in Objective-C on a NeXTStep platform, sharing many of the features of sophisticated existing systems such as NETSIM [9] and TRAFFICQ [10]. The simulator is based around a graphical user interface and incorporates a graphical network design and editing tool, allowing networks to be built, quickly and intuitively, from a palette of simple building blocks, including junctions, roads, traffic lights, giveway signs, traffic sensors and vehicles. The simulation is run as a series of discrete time steps, the computational load scaling proportionally with the number of vehicles in the network (it runs in real-time for networks containing up to about 100 vehicles).

Two contrasting auditory representations of the network simulation were developed. The first focuses on a single junction and by considering flow information at three points, illustrates how the pattern of traffic flow evolves as traffic density increases. The technique exploits two competing perceptual organizations to highlight the distinction between free-flow and congestion. The second scheme renders an auditory scene,

whose structure corresponds to the graphical display, and presents simulation data redundantly to both the aural and visual modalities.

3.2 SONIFYING A SIMPLE JUNCTION

At a free flowing junction, travel time between a detector upstream of the intersection and one downstream will depend on the positioning of the detectors, the architecture of the junction and the behavioral characteristics of the driver. The delay due to interference with other vehicles will be relatively small. By delaying and combining the upstream detector outputs, a signal can be generated that is highly correlated with the downstream detector. However, the accuracy of this estimate will be reduced during periods of heavy traffic flow due to unaccounted vehicle interactions at the intersection. Therefore, a sonification that highlights the correlation between the downstream detector signal and its estimate should portray the nature of the junction traffic flow in the system.

A very basic junction was employed; two lanes, one terminated by a giveway sign, converging to form a single lane and traffic flow was monitored by three sensors, positioned as shown in the diagram below.

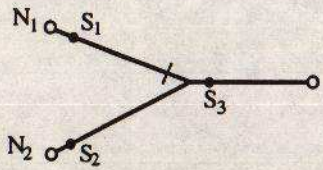

FIGURE 2. A simple junction. Traffic flow is from left to right.

The average inter-sensor delays were estimated by first generating vehicles at N_1 and measuring the mean delay between sensors S_1 and S_3, and then generating vehicles at N_2 and measuring the mean delay between sensors S_2 and S_3. Denoting these times as $t_{1,3}$ and $t_{2,3}$ respectively, the estimate of S_3 is defined as:

$$E_3(t) = T(S_1(t), t_{1,3}) + T(S_2(t), t_{2,3})$$

where the function $T(S(t), x)$ is a time delay of duration x acting on the signal $S(t)$ and $S_1(t)$ and $S_2(t)$ are the signals generated by sensors S_1 and S_2 respectively.

To highlight the correlation between the destination sensor signal, $S_3(t)$ and its estimate, $E_3(t)$ the common fate grouping principle was exploited through the application of comodulation. The sensor and estimated signals were up-sampled and passed through leaky integrators, providing the binary sensor signals with a larger and controllable dynamic range. The decay rate of the integrators was adjusted so that it was slow enough to allow the signal to build significantly in periods of heavy traffic but not so slow that detail at the level of individual sensor events was lost. After this processing, the downstream sensor and estimate signals were used to modulate the odd and even harmonics respectively of a carrier signal composed of a 200 Hz tone and its first fifteen harmonics. Various different modulations were used: amplitude modulation, frequency modulation and a combination of both. The experiments were repeated using different styles of modulation and a range of traffic densities. It was hoped that as the correlation between the two signals broke down this would be clearly detectable in the sonification as it would break into two auditory streams, one containing the odd harmonics and a second containing the even harmonics. This breakdown should be perceived as the introduction of a second 'voice' an octave above the original complex. The results have to be heard to be appreciated, but for the sake of discussion, spectrograms of several sonifications are included below. The sonification scheme compresses time 25 fold, so that the 10 second spectrograms represent

approximately two minutes of simulation time.

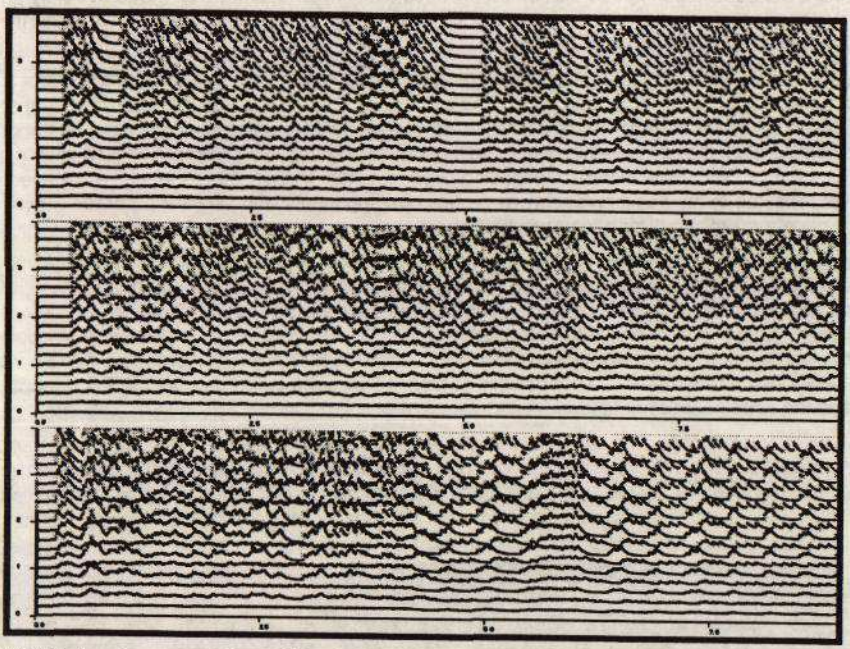

FIGURE 3. Spectrograms of the sonification of a simple junction under light (top), moderate (centre) and heavy (bottom) traffic conditions using frequency modulation.

Under light traffic conditions (figure 3, top), it can be seen that the signals are highly correlated except for short segments, for example between 1.5 and 2 seconds, which correspond to periods of transient congestion during the simulation. Poorly correlated sections of sufficient length are clearly detectable by the perception of a second voice at an octave above the perceived pitch of the background complex. The effect is especially pronounced when amplitude modulation alone is employed. In the case of moderate traffic flow (figure 3, middle) it can be seen that while there are still times when the junction is flowing freely the correlation between the two signals breaks down for much longer periods. As the traffic density is increased still further long queues form an the junction causing a steady traffic flow at the network output, averaging out the randomness found at the inputs. This can be seen as the high flat sections of the odd harmonics. When the queues back up far enough they eventually cover the network input detectors: vehicles queuing to get into the network cause a loss of randomness at the input detectors, so the estimated signal as seen in the even harmonics also becomes high and flat. Under heavy traffic conditions, a pulsing effect can be seen on the odd harmonics, caused by vehicle conflicts at the junction. When large queues are also present this can cause delayed pulses on the upstream detectors, effecting the estimated signal and hence the even harmonics. These pulses are readily perceived in the sonification as a rhythmic pattern, and the difference between either one set or both sets of harmonics being affected can be clearly heard.

4.3 SONIFYING AN ARBITRARY NETWORK

Employing a contrasting scheme to that of the previous section, an attempt was made to sonify an entire arbitrary network. It was hoped that the sonification would highlight areas of congestion, while still maintaining the discernibility of the individual vehicles.

The parameters of each vehicle, in each frame of the simulation, were taken and used to control a sound synthesis algorithm, each vehicle thus generating a component of the auditory display. The parameters relating to each vehicle change smoothly over time, so given a suitable parameter mapping, by the principle of good continuation, the components of the auditory display relating to a particular vehicle should hold together as a single perceptual stream. Congested regions, where many vehicles inhabit a confined space, should be reflected in the auditory display by groups of components with similar properties which, due to the principle of proximity, will tend to perceptually fuse.

Each display component is formed from a set of evenly spaced harmonics which are then modified by the parameters of the vehicle it is encoding. The vehicle's longitudinal, or x position is mapped, on to inter-aural time and intensity difference cues in such a way as to produce a crude left/right localization percept. This simulation parameter is also encoded redundantly as fundamental frequency, improving resolution in this dimension and helping to prevent interference of the localization cues of components generated by vehicles on opposite sides of the display. The y position is mapped onto the centre frequency of a bandpass filter controlling the component's brightness.

The vehicle speed is mapped inversely on to amplitude (thus the slowest vehicles sound loudest). Although this is physically counter-intuitive, it ensures that congested areas of the network will produce a larger signal than free-flowing areas. Additionally, the amplitude of a vehicle is slowly increased while it is moving below a threshold speed, so chronically congested areas of the network will become increasingly salient.

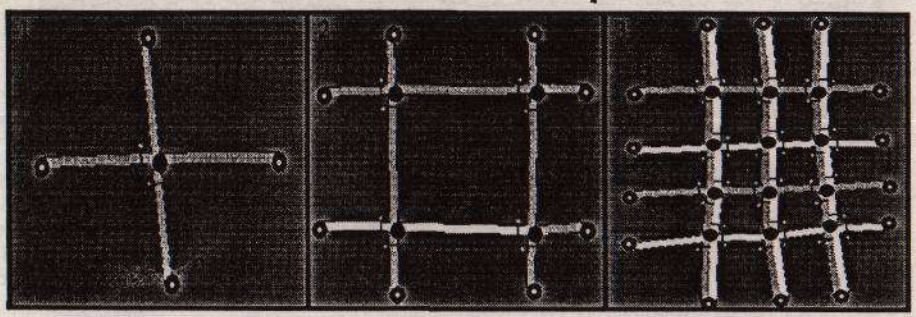

FIGURE 4. Networks of increasing complexity employed in evaluation of the sonification

The sonification was applied to a series of networks of increasing complexity (see figure 4). With network 1 under light traffic conditions it was possible to track the motion of individual vehicles. As the traffic density is increased, hearing out individual vehicles is no longer possible but the phases of the traffic-lights become very apparent. In the initial phase vehicles queue along the east going lane and the sound gradually thickens getting lower, louder and shifting towards the left of the auditory field as the queue lengthens. When the lights change, this sound dissipates being slowly replaced by centrally located sound components getting duller and louder as vehicles start to back up along the north-going lane.

In the second network the four junctions are widely spaced. The components due to queues forming at each junction are dissimilar and therefore easily distinguishable. As the situation becomes more congested and queues spread towards neighbouring junctions, the sonification starts to become confused. The present im-

plementation appears to lack sufficient segregating cues to prevent the components, which have common onset times and patterns of amplitude modulation, from fusing into a single mass of sound. In the sonification of the 3rd network this deficiency becomes very notable and segregating the components due to one area of congestion from those due to another is very difficult, if not impossible.

The confusion that occurs as the network becomes heavily congested is due to a lack of structure in the sonification. In making sense of the visual display much of the information used is incorporated in the static graphical representation of the network which is not represented in the sonification. It is hoped that by aurally encoding information of this type the sonification can be clarified. For example, comodulation and common onset cues could be used to perceptually group components in the auditory display, so as to reflect groups of vehicles that are on the same road or heading towards the same junction.

4 CONCLUSIONS

We have made the claim that auditory scene analysis considerations should be a central issue in sonification design. In the design of displays employing multiple simultaneous sources, consideration needs to be given to scene analysis principles to ensure that the display is perceived as containing the intended sources. Furthermore, even if the sonification is to contain only one apparent source, we should still consider scene analysis as there is no reason why a naively generated synthetic auditory signal should be heard as coherent source - it may instead fragment into several auditory streams.

There is, of course, another stance; it may be that the ability to interpret a given sonification is unaffected by the involvement of auditory scene analysis: That is, that the number and structure of the perceived acoustic streams is an irrelevant factor to the efficacy of the display. For example, consider coding a stream of data as a sequence of tones of variable pitch. Depending on the details of the presentation rate, the pitch range employed and the smoothness of the data, we may perceive the display as a single stream, or find the display fragmented into several streams each with a separate pitch range. Do these differences in our perception of the auditory scene, by themselves, effect our ability to interpret the data?

So this brings us to the key questions of this research:

- Does the coherence, or lack of coherence, of a simple auditory display effect the ease with which we can use it?
- If so, in what ways can we employ auditory scene analysis principles to improve the coherence of, and thus enhance, a sonification?
- When can multiple simultaneous sources usefully be employed in the analysis of complex data? Are there any tasks that require, or could benefit from, the use of complex auditory scenes?
- If so, in what ways can auditory scene analysis principles be exploited to improve the clarity of auditory displays employing multiple simultaneous sources?

The traffic simulation sonifications outlined in the previous section, although providing interesting examples, do little towards answering these fundamental questions. To answer these it will be necessary to step back from complex and uncontrolled data domains and concentrate on simple tasks involving simple synthetic data. The extent and manner in which scene analysis interpretation affects task performance will no doubt depend to a large extent on the task itself. It is therefore necessary to make a careful study of the individual skills involved in interpreting sonifications, whether these be involved in data classification or data prediction tasks, and evaluate performance of these skills under sonification schemes employing contrasting constructions of the auditory scene. These experiments will have to be designed to provide both a means of evaluating the subjects performance of the given task, and an objective measure of the auditory scene interpretation that the subject is forming. Results should provide a set of guidelines that sonification design-

ers can employ to generate effective sonifcations making full use of the auditory scene analysis capabilities of the human auditory system.

5 REFERENCES

[1] Bregman, A.S. "Auditory scene analysis, the perceptual organization of sound." MIT Press, Cambridge, Mass., 1990.

[2] Brown, G.J. "Computational auditory scene analysis: A representational approach." Ph.D. Thesis, Department of Computer Science, University of Sheffield, U.K., 1992.

[3] Bly, S. "Sound and computer information presentation." Unpublished Ph.D. Thesis, University of California, 1982

[4] Cooke, M.P. "Modelling auditory processing and organization.", Cambridge University Press, Cambridge, MA 1993.

[5] Dougherty, M.S., Kirby, H.S. and Boyle, R.D."Models for traffic forecasting and control using neural networks" Tech. Report 313, Institute for Transport Studies, University of Leeds, 1992

[6] Fitch, W.T. and Kramer, G. "Sonifying the body electric: Superiority of an auditory over a visual display in a complex, multivariate system." In Kramer, G., editor, Auditory display: Sonification, audification and auditory interfaces, pp 307--326, Addison-Wesley, Reading, MA, 1994.

[7] Frysinger, S.P. "Applied research in auditory data representations." In *Extracting meaning from complex data: Processing, display, interaction*, volume 1259, pp 130--139. SPIE, 1990.

[8] Kramer, G. "Auditory display: Sonification audification, and auditory interfaces" Addison-Wesley, Reading, MA, 1994.

[9] Lieberman, E.B., Worral, R.D., Wicks, D. and Woo, J.L. "Netsim model" Tech. Report FHWA-RD-77-41,42,43,44,45 U.S. Federal Highway Administration, Department of Transport, 1977.

[10] Logie, D.M. "Comprehensive model for traffic management" *Traffic engineering and control* 20:516-- 518, 1979

[11] Lunney, D. and Morrison, R.C. "Auditory presentation of experimental data." In *Extracting meaning from complex data: Processing, display and interaction*, volume 1259, pp 130--139. SPIE, 1990.

[12] Mansur, D.L. "Graphs in sound: A numerical data analysis method for the blind." Unpublished Master's Thesis, University of California, Davis, 1984.

[13] Mezrich, J.J., Frysinger, S. and Slivjanovski, R. "Dynamic representation of multivariate time-series data." *J. Am. Stat. Assoc.*, 79:34--40, 1984.

[14] Pollack, I. and Ficks, L. "Information of elementary multidimensional auditory displays." *J. Acoust. Soc. Am.*, 26(2):155--158, 1954.

[15] Scaletti, C. "Sound synthesis algorithms for auditory data representations." In Kramer, G., editor, Auditory display: Sonification, audification and auditory interfaces, pp 223--252, Addison-Wesley, Reading, MA, 1994.

[16] Smith, S., Grinstein, G.G. and Bergeron, R.D. "Stereophonic and surface sound generation for exploratory data analysis." In Blattner, M.M. and Dannenberg, R.B., editors, *Multimedia interface design*. ACM Press/ Addison-Wesley, Reading, MA, 1992.

[17] Speeth, S.D. "Seismometer sounds." *J. Acoust. Soc. Am.*, 33(7):909--916, 1961.

[18] Weintraub, M. "A theory and computational model on monaural sound separation" Ph.D. Thesis, Department of Electrical Engineering, Stanford University, 1985.

[19] Williams, S. "Perceptual principles in sound grouping." In Kramer, G., editor, Auditory display: Sonification, audification and auditory interfaces, pp 307--326, Addison-Wesley, Reading, MA, 1994.