Comparing User Performance on Parallel-Tone, Parallel-Speech, Serial-Tone and Serial-Speech Auditory Graphs

Prabodh Sakhardande, Anirudha Joshi, Charudatta Jadhav, Manjiri Joshi

Comparing User Performance on Parallel-Tone, Parallel-Speech, Serial-Tone and Serial-Speech Auditory Graphs

Prabodh Sakhardande[1], Anirudha Joshi[1], Charudatta Jadhav[2] and Manjiri Joshi[1]

[1] Industrial Design Center, IIT Bombay, Mumbai, India
{prabodhs, anirudha, manjirij}@iitb.ac.in
[2] Tata Consultancy Services Limited, Mumbai, Maharashtra, India
charudatta.jadhav@tcs.com

Abstract. Visualization techniques such as bar graphs and pie charts let sighted users quickly understand and explore numerical data. These techniques remain by and large inaccessible for visually impaired users. Even when these are made accessible, they remain slow and cumbersome, and not as useful as they might be to sighted users. Previous research has studied two methods of improving perception and speed of navigating auditory graphs - using non-speech audio (such as tones) instead of speech to communicate data and using two audio streams in parallel instead of in series. However, these studies were done in the early 2000s and speech synthesis techniques have improved considerably in recent times, as has the familiarity of visually impaired users with smartphones and speech systems. We systematically compare user performance on four modes that can be used for the generation of auditory graphs: parallel-tone, parallel-speech, serial-tone, and serial-speech. We conducted two within-subjects studies - one with 20 sighted users and the other with 20 visually impaired users. Each user group performed point estimation and point comparison tasks with each technique on two sizes of bar graphs. We assessed task time, errors and user preference. We found that while tone was faster than speech, speech was more accurate than tone. The parallel modality was faster than serial modality and visually impaired users were faster than their sighted counterparts. Further, users showed a strong personal preference towards the serial-speech technique. To the best of our knowledge, this is the first empirical study that systematically compares these four techniques.

Keywords: Auditory Graphs, Auditory Feedback, Sonification, Human Computer Interaction.

1 Introduction

Graphs or charts are a visual representation of some type of relational data. They are widely used across domains and disciplines. Different types of graphs like a bar graph, line graph, pie chart, histogram, etc. are chosen by the designers of the graph depending upon the nature of the information they want to convey. Due to the inherently visual

nature of graphs, it is difficult for visually impaired users to access or even understand their benefits.

Graphs provide several benefits and affordances to a sighted user. They provide multiple layers of information that allows the user to explore and discover the information in the sequence and to the level of detail that they desire [22]. Graphs allow users to quickly discover trends in the data such as “Has the value of a stock been generally increasing in the last 5 years?”, look for specific information like “What was the highest or lowest value for the stock in the last 5 years?” or discover the relationship between two points in the graph such as “Which of the two countries have similar landmass size?”. An inquisitive mind can gain multiple insights from the same graph. Graphs on web pages can be made accessible through attributes like “alt” [16]. While such mechanisms allow the author to describe the contents of the graph, it does not allow the user to explore or discover beyond what has already been “canned” by the author.

Research on making graphs accessible is not new. Two main techniques that have been explored are making graphs tangible and making graphs auditory. In this paper, we focus on auditory graphs, specifically focussing on the modalities of tones and speech, along with serial and parallel.

As we will discuss in the related work section, previous research on sonification has compared the use of synthesized speech with non-speech audio (such as tones) and found that non-speech audio decreases workload and task time [8]. However, this study was done in the early 2000s. With the recent advances in technology, speech synthesis has reached a level of maturity. Through multiple devices like smart speakers, speech input and output is now much more ubiquitous. Moreover, visually impaired users have gained a lot of experience with speech-based screen readers. Hence a comparison between speech and tone is worth investigating again.

Audio is perceived sequentially, so a potential downside of using audio is that it can slow the user down. Thus, at least some advantage of data becoming more accessible is lost because of slow speeds. Some research has explored the effect of using two audio streams in parallel instead of in series in order to speed up the process [2]. It found that parallel audio leads to faster task completion time than audio in series.

To the best of our knowledge, no study has been conducted that compares the performance of auditory graphs in both dimensions simultaneously: speech vs. tone and parallel vs. serial. In this paper, we systematically compare the effect of the four modes (parallel-tone, parallel-speech, serial-tone, and serial-speech) on user perception. We investigate the performance of both sighted and visually impaired users. We present the results of two within-subjects studies - one with 20 sighted users and the other with 20 visually impaired users.

2 Related Work

In order to make graphics like charts, graphs and images accessible to users of screen-readers, W3C [30] recommends adding a short description to identify contents of the image and a longer description that can contain detailed information such as the scales, values or trends in the data. While making the graph accessible, this method does not allow the user to explore the data creatively on his/her own and may not fulfill the needs of some users. Also, presenting all the information in speech form increases the task load when graphs are long [48].

The visual medium communicates a lot of information in parallel [35], allowing sighted users to use visual graphs in diverse ways. In contrast, by and large, audio communicates information serially. One of the challenges in designing auditory graphs is retaining this rich diversity in a serial medium. In literature, auditory graphs have been used for presenting data with several goals. SoundVis [2] uses audio to communicate detailed information of a graph. TeDUB [21, 31] describes an automatic / semi-automatic way of making diagrams accessible. Audio has also been used to quickly provide an overall gestalt effect of a graph [1, 2, 36, 37].

In this paper, we focus on enabling users to estimate (e.g. the highest point) and compare values (e.g. two similar points) of bars in a bar graph. While a sighted user may explore a bar graph for other purposes, we argue that comparison and estimation are the most common use cases of a bar graph.

Tactile [20], auditory [28, 32, 36] and combined [33, 47] modalities have been used to make graphs accessible by visually impaired users. Tactile graphs have been explored to allow exploration of line graphs [47], bar graphs [47], georeferenced data [38], etc. alongside auditory feedback. Tactile graphs themselves are limited by the amount of data that can be presented at the same time [34].

In this paper, we focus on the use of audio to make graphs accessible. In literature, several terms have been used for graphs that use audio for communication including “auditory data representation” [29], “auditory graphs” [5, 9, 13], “sonification” [2, 3, 7, 10, 11], etc. In this paper, we use the term “auditory graphs” as against “visual graphs” to include any graph that communicates numerical information through audio.

While speech is considered a natural form of communication, tones are non-speech sounds like musical instruments, earcons, everyday sounds, synthesized sounds and so on. Different parameters of sound like pitch [1, 2, 4, 13], loudness [13][17], duration [10], frequency [10], timbre [3][39], panning [2][3][4] have been used to map data to auditory graphs.

Mansur et al. [1] compared the performance of tactile graphs and auditory graphs for conveying characteristics of the curve such as the slope, monotonicity, convergence, symmetry and whether the line graph was exponential. The study was done with sighted

and visually impaired users. Results showed that the auditory graph was faster to use while the tactile graph was significantly more accurate than the auditory graph. The authors also concluded that the accuracy of auditory graphs could improve with some practice. In this study, music i.e. continuous tone and not discrete tones convey graph information. Ramloll et al. [8] investigated the use of speech and non-speech audio and speech-only interface for accessing numerical data (between 0-100) in 26 rows and 10 columns. They found that making data available in non-speech audio form resulted in lower subjective workload and higher task successes and significantly lower mean task completion time.

Different strategies of encoding data in sound to leverage the relationships between the sound parameters have been explored [12, 19, 28, 27]. For example, frequency and pitch have a logarithmic relationship, for logarithmic data set or for multivariate data set the dimension that varies exponentially is mapped onto frequency. Peres et al. [3] conducted an experiment to investigate whether using redundant design i.e., using two parameters of sound to represent the information redundantly is useful. They found that the benefit depends upon the properties chosen for the mappings. Using pitch and loudness ("integral parameters") to map the data to auditory graphs was found to be beneficial whereas using pitch and timing ("separable parameters") together were not found to be useful.

While some studies have suggested that panning in parallel mode may not be sufficient to make data streams differentiable [40][41], others reported that panning does help certain types of tasks when presenting multiple data series in parallel [4]. Audio is perceived sequentially, so a potential downside of the use of audio is that it can tend to slow the user down as audio needs to be perceived sequentially. Thus, at least some advantage of data becoming more accessible is lost because of slow speeds. Previous research has explored the effect of using two audio streams in parallel instead of in series to represent two or more data series [2][4]. It was found that parallel audio leads to faster task completion time than audio in series in sonified graphs with two data series, for finding intersection of the two lines. Audiograph [22] compared two ways of playing tones: note-sequence where all tones between the note for origin and that for the current value were played and note-pair, where only the notes for origin and the note for the value were played. Note-sequence had the advantage of conveying the value in pitch and in the time, while Note-pair had the advantage of being brief. Note-sequence emerged as the more accurate method. Many studies investigated the effect of providing context information in auditory graphs, the visual graph equivalent of grid-lines along X and Y axes [11][17][35].

Nees et al. [17] investigated the effect of varying the loudness of the context audio relative to the data audio on the performance and found that keeping the loudness lower or higher than that of the data audio was beneficial to the performance of point estimation tasks. Magnitude estimation tasks have been used to study preferred mappings of data to the sound parameters and whether they were the same for sighted and visually impaired users [10]. Three parameters of sound were varied: pitch, tempo, or spectral

brightness. Users were presented with a sound stimuli and asked to estimate its value. The study aimed to understand whether users associated higher pitch or spectral brightness and faster tempo with higher or lower values for entities like dollar, size, temperature, pressure and velocity and whether this was same for sighted and visually impaired users.

Different kinds of data has been represented using auditory graphs. Stockman et al. [36] investigated the use of sonification to allow users to get an overview of data in the spreadsheets. They explored auto-sonification that gives an overview of the data automatically and also manual sonification where a visually impaired user could select ranges of data to sonify in order to quickly understand trends in the selected range and identify characteristics such as outliers [36]. Flower et al. used binned data to generate auditory histograms with a musical pitch to present statistical information like shape of the distribution [49]. AUDIOGRAPH developed by Rigas et al. [22] used auditory tones to convey information like shape, size of the shape, location of the object in space, etc. to enable users to manipulate objects in space.

While the human ear can detect small pitch changes very well, studies show that there is no linear relationship between pitch perception and the actual frequency of the auditory signal and that it is user dependent [8][50][51]. In our study design, like Ramloll et al. [8], we did not include value estimation tasks. Brown et al. [3] recommend using instruments with wide pitch like the Piano and Trumpet and MIDI tones 35-100 as they are easily perceived and differentiable. In our design, we used pitch variation to show differences in values in a graph. In the next section, we describe in more detail how data in bar graphs were mapped to MIDI tones with piano timbre. While sonification gives a quick overview of the data [6, 13] and reduces cognitive load for tabular and graph datasets [8], it cannot provide an absolute perception of value at a point. Speech-based auditory graphs, on the other hand, can provide exact value at a point.

3 Design of The Auditory Graphs

In this experiment, we used a simple bar graph with a single data series, such as the one shown in Fig. 1 below. A sighted user may explore such a bar graph in different ways, depending on what task s/he is trying to do. S/he may be interested in a point estimation task (such as which city gets the lowest rainfall). Or s/he may be trying to do a point comparison task (such as which city gets similar rainfall to my city). Our goal was to provide similar flexibility of exploration to a blind user. Here we describe how the information contained in a bar graph was conveyed to the user through each of the four auditory graph techniques.

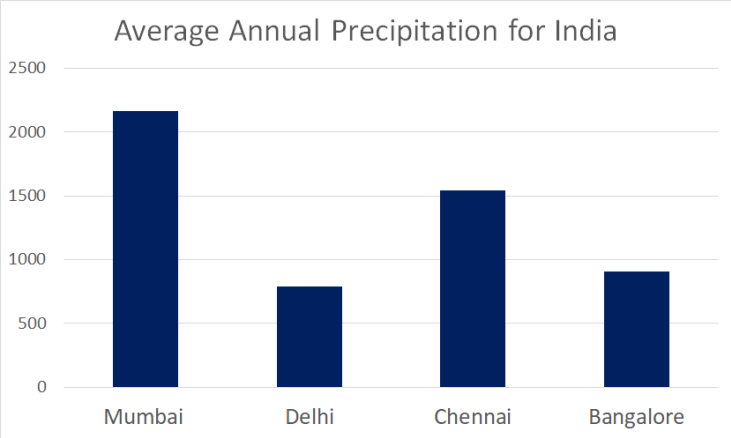

Fig. 1. Graph showing annual average rainfall in different cities in India in mm.

3.1 Serial-Parallel

By default, audio information is communicated to the user sequentially, i.e., one after the other. For example, the user would hear "Mumbai 2168, Delhi 790, Chennai 1541, Bangalore 905". We call this "serial" modality. Unfortunately, serial communication would take at least as much time as it might take to play the audio files, and more time if the user needs repetition. It is possible to speed up the audio playback, and several users (especially visually impaired users who use screen readers) are used to perceiving spoken audio at very high speeds.

Yet, another way to reduce the time is to playback multiple elements of the audio simultaneously. In the example above, we could play back the words "Rainfall in mm, (Mumbai = 2168), (Delhi = 790), (Chennai = 1541), (Bangalore = 905)". wherein the brackets imply that the words in the brackets (e.g. "Mumbai" and "2168") are played simultaneously. We call this "parallel" modality.

In our study, we compared serial modalities with parallel modalities. While it may seem natural to assume that parallel modalities will be faster than serial modalities, this may in fact not be so effective. Firstly, the amount of time taken to communicate the actual information about the elements may be not much in comparison with the other time taken. A bar graph may have other information that needs to be communicated (e.g. the title). The user may need time to perceive and process the information and to navigate the graph. Secondly, the user may find it harder to perceive the information presented in the parallel modalities in comparison with the serial modalities, leading to errors or repetitions. The goal of our study is to explore if parallel modalities actually result in time-saving, whether they induce more errors, and whether the users prefer them to serial modalities.

3.2 Speech-Tone

Our interest is to compare the effect of speech and tone in communication. “Speech” is considered a natural modality of communication. In contrast, “tones” need interpretation. As discussed in the related work, there are many ways to represent data in tones. However, for this study, we restricted ourselves to variations of pitch alone, as a single dimension of variation.

The elements of the X-axis of the bar graph may either be ordinal (e.g. ranks of athletes), related by a natural sequence or progression (e.g. months of a year), or may be unrelated by any natural sequence or progression (e.g. countries). Considering that in a general case, elements of the X-axis may be unrelated to each other in any predictable way, we decided to always use speech to communicate information about the elements on the X-axis.

In contrast, the Y-axis always contains numerical data, (e.g. time taken by athletes, rainfall in each month, or population of countries). It is possible to communicate this data through either speech or tone modalities. We chose to vary the information about the Y-axis values through the use of tone in some conditions, and speech in the others. Thus, in tone modalities, the user would hear, “Rainfall in mm, Mumbai beep 1, Delhi beep 2, Chennai beep 3, Bangalore beep 4”, where beeps 1-4 represent tones representing 2168, 790, 1541 and 905 respectively. In speech modalities, the user would hear “Rainfall in mm, Mumbai 2168, Delhi 790, Chennai 1541, Bangalore 905”.

3.3 Auditory Graphs

We used combinations of the parallel-serial and speech-tone modalities described above resulting in four modes of our study, namely parallel-tone, serial-tone, parallel-speech, and serial-speech. We selected eight datasets to create auditory bar graphs of real-world data sourced from The World Bank [24]. We selected the topics of rural population, surface area (sq km), incidence of tuberculosis (per 100,000 people), total population (thousands), population growth (%), forest area (%), CO2 emissions (kt), and GDP growth (%) of various countries. We chose these topics because we expected that participants would not be too familiar with such data, and would need the graphs to find answers to the questions asked in the tasks. We wanted the data to be realistic so that our study will have good external validity. Four datasets (rural population, surface area, tuberculosis, and total population) contained six data points on the x-axis (the “short” graphs). The other four datasets contained 15 data points on the x-axis (the “long” graphs). The number 15 was chosen to safely be beyond the short term memory limit [23]. In order to counter the variation in the time required to speak out the names of different countries (e.g. “United Kingdom” and “India”), all the short graph datasets had the same set of countries. Similarly, all the long graph datasets had the same set of countries. Thus, we created a total of (8 datasets x 4 four modes = 32) auditory graphs.

We made some specific design decisions for each modality. In the speech modalities, the numeric readout of values was limited to 3 digits (most significant bits). For example, if the value was 25.263, we rounded it off to 25.3. If it was 2.5263, we rounded it off to 2.53, and if it 25,263, we rounded it off to 25,300. As our study was done in India, we used "Animaker Voice" [26] with the Indian English female voice "Raveena" at the default speed to generate the speech.

In the tone modalities, the method of converting numerical data to auditory tones was derived from Brown et al. [2], where data values were mapped to MIDI notes. The data was mapped to MIDI notes linearly. The lowest pitch was chosen as a MIDI value of 20 and the highest was MIDI 100. We used "Sonic Pi" [25] with the "Piano Synthesizer" option as the musical instrument for generating tones as they were reported to be perceived more easily [8, 18].

In the parallel modalities, the sounds were "left aligned", i.e. both sounds started playing at the same time, although they may end at different times depending on their lengths. In both the serial and parallel modalities, the two audio streams representing the X and Y-axes were spatially separated. The audio representing the X-axis was always played to the right ear and the audio representing the Y-axis was always played to the left ear.

Once the audio clips for all modalities were generated, the audio was combined and spatially modified, mapping X-axis to the right ear and Y-axis to the left ear in Audacity [42]. All source clips were brought to the same volume gain. The means and standard deviations in seconds of the total duration of audio recordings for graphs in each mode if all information in that graph was heard without repetition or delay are shown in Table 1. An attempt was made towards making all tone audios of the same length. In most cases, the duration of tone playback (for X-Axis) was shorter than the duration of speech playback (for Y-Axis). Thus as the graphs of a set (short or long) had the same Y-Axis (countries), in Table 1 the tonal modalities show a small standard deviation.

Table 1. The means and standard deviations of the duration of audio recordings in the four modes for the two graph lengths.

	Short Graphs (s), N = 4		Long Graphs (s), N = 4	
	Mean	SD	Mean	SD
Parallel tone	3.70	0.00	9.35	0.00
Serial tone	6.43	0.10	16.45	0.12
Parallel speech	8.85	2.35	16.95	6.41
Serial speech	12.42	2.61	25.71	6.90

We created a prototype running on a laptop computer for the studies. We used Processing [46] to create the keyboard based user interface and link keystrokes to respective audio playback.

The users navigated the graphs using the up (↑), right (→) and left (←) arrow keys and the enter key (↵) of a keyboard. While we acknowledge that navigation is a crucial aspect of an auditory graph, in order to systematically study the effect of the modes on perception, we kept the navigation complexities at a minimum. When the user presses the enter key (↵), the audio corresponding to the first bar of the graph is played. When the user presses the right arrow key (→), the audio corresponding to the next bar is played after the audio corresponding to the first bar is over. When the user pressed the left arrow key (←), the audio corresponding to the previous bar is played. When the user presses the up arrow key (↑), the audio corresponding to the current bar is repeated. Audios do not interrupt or overlap each other. If the user tries to go beyond the first or the last values on the X-axis, a tone is played back indicating that the user has reached the end.

Here we illustrate each mode using data from Fig. 1. Irrespective of the mode, the user first hears a "title", such as "Rainfall in four major cities in India in millimeters". When the user is ready, s/he presses the enter key (↵). Then in the parallel-tone mode, the user hears "Mumbai" and "beep 1" simultaneously. On pressing the right arrow key (→), s/he hears "Delhi" and "beep 2" again simultaneously, and so forth. In serial-tone mode, the user hears "Mumbai", followed by "beep 1", and after (→) "Delhi", followed by "beep 2", and so forth. In parallel-speech mode, the user hears "Mumbai" and "two thousand one hundred and sixty" simultaneously, and after (→) "Delhi" and "seven hundred and ninety" simultaneously, and so forth. In serial speech mode, the user hears "Mumbai", followed by "two thousand one hundred and sixty", and after (→) "Delhi", followed by "seven hundred and ninety", and so forth. In all of these cases, the values on the X-axis ("Mumbai", "Delhi" etc.) are heard in the right ear and the corresponding values on the Y-axis (in speech or tone) would be heard in the left ear.

4 Method

We conducted two within-subject studies to compare speed, errors, and preference of users. One study was conducted with sighted users and the other with visually impaired users. The study compared across parallel-serial and speech-tone modalities, resulting in four modes of auditory graphs - parallel-tone, serial-tone, parallel-speech, and serial speech. In each mode, we studied auditory graphs of two lengths, a 6-point graph (short) and a 15-point graph (long). On each graph, users performed three tasks - estimating the highest point (task 1), the two most similar points (task 2), and the lowest three points (task 3). The wording of the tasks was modified to suit the contents of the graph. For example, for the graph of the rural population, the task 1 was "Which country has the highest rural population?" None of the tasks required the participants to guess exact values, but only to compare values.

Thus we had 4 modes, with 2 graph lengths per mode, and 3 tasks per graph, resulting in 24 tasks per user. The order of the modes was counterbalanced across participants through a balanced Latin square. The order of graphs was kept the same across modes - the participants first used the short graph, and then the long graph. The order of the three tasks for each graph was kept the same for all graphs. Thus the complete design of the experiment across the independent variables was speech-tone (2) x parallel-serial (2) x graph length (2) x tasks (3) x user type (2 - between subjects).

All participants went through a training protocol before performing the tasks. The training protocol was the same for sighted and visually impaired participants. The experimenter first familiarised the participant with bar graphs. He presented two versions of a bar graph, namely a printed version and a transparent tactile overlay. This graph contained information about rainfall for a city by month from April to September. The experimenter asked the participant to explore this graph on his/her own and perform a talk aloud. If the participant was not familiar with graphs (which was the case for some visually impaired participants), the experimenter explained concepts of the X-axis, Y-axis points, and plotting. After this, the experimenter explained and demonstrated the four modes of auditory graphs used in the experiment. The order of modes for the training was parallel-tone, serial-tone, parallel-speech, and serial-speech. The same rainfall data was used as examples. The experimenter asked the participant one question for each mode. The participant was allowed to listen to these examples multiple times if required.

After the training, the participant performed the tasks. When the participant was ready, s/he heard the title of a graph. Then the experimenter verbally gave the participant the first task for that graph. The participant was instructed that after s/he finds the answer, s/he should immediately say it aloud. For the tasks requiring multiple answers, the participant had to be ready with all the answers before answering. Before starting with the task, the participant had the option to listen to the graph title and the task as many times as required. When s/he was ready to do the task, s/he pressed the enter key (↵) and started navigating the graph. The task time started when the participant pressed the enter key (↵). A soft limit of 5 minutes was initially kept for every question after which the task was to be abandoned, but this was not required as no one exceeded this limit. As soon as the participant started answering, the experimenter pressed a key on another keyboard, which stopped the navigation and logged the time. After the participant had told the answer(s), s/he was told if his/her answer was correct and if not, the error rank (see below). This was done to keep the participant engaged with the study.

After completing the three tasks of the short graph of the first assigned mode, the participant moved on to do the tasks of the long graph of the same mode. After completing the six tasks of the two graphs of the first mode, the participant was assigned the next mode. After completing the 24 tasks for the 8 graphs of the 4 modes, the participant was asked to fill out the system usability score (SUS) [14] to record his/her

feedback for each mode. If s/he had a problem recalling a particular mode, s/he could hear to the training graph of that mode again.

The dependent variables of the study were task time, the error rank of the given answer, and the SUS score. The task time was the time from when a participant started an exploration of the graph to when s/he started giving the answer. The error rank of a task was calculated by ranking all possible answers of that task in ascending order of correctness starting with zero. Thus the error rank of an answer was 0 if the answer given by the user was correct, 1 if it was the second best answer and so forth.

4.1 Participants

We conducted two studies. The first study was done with 20 sighted participants and the second study was done with 20 visually impaired participants. Anonymous demographic information regarding self-reported hearing ability, age, level of education, musical training and prior familiarity with graphs was recorded.

Sighted participants were university students (mean age = 24.7 years, SD = 4.65). The selection criteria were that they had to be currently enrolled and between the ages of 18 to 36. None of them were familiar with the research or with the people involved in the studies. They were recruited through word of mouth and through online groups. As these users walked to the experiment center, they were not compensated for travel. Instead, they were given a token of appreciation in the form of a gift voucher worth INR 150 (about USD 2).

Visually impaired users were recruited from across the city (mean age = 29.71 years, SD = 11.33). They were initially recruited through word of mouth and then through a snowballing process. Selection criteria was any visually impaired individual with more than 8-10 years of formal education. They were compensated for their travel to and from the experiment venue.

4.2 Experiment Apparatus

The experiment was conducted in a quiet room with minimal human movement around the seating. A Hewlett Packard laptop running Windows 10 was used for conducting the experiment. Philips DJ series headphones were used for all participants and the default volume was kept at 38. The experimenter asked the participant if s/he were comfortable with the volume and changed it if necessary. Users were given an external USB keyboard with mechanical button keys for navigating the graph. The arrangement was such that the participant could not see the screen from where s/he was seated. Only one participant participated in the experiment at a time.

5 Results

We performed two within subjects repeated measures ANOVAs with time and error as dependent variables and parallel-serial (2), speech-tone (2), graph length (2), tasks (3), and user type (2, between subject) as independent variables. We found the following differences significant for time: Tone is faster than speech ($F(1,38) = 8.093$, $p = 0.007$), parallel is faster than serial ($F(1,38) = 4.928$, $p = 0.032$), short graphs are faster than long graphs ($F(1,38) = 234.031$, $p < 0.0005$), task 1 (highest point) is faster than task 2 (two similar points) and task 3 (three lowest points) ($F(2,37) = 18.74$, $p < 0.0005$), and visually impaired participants are faster than sighted participants ($F(1,38) = 8.246$, $p = 0.007$). The significant results for errors are: speech is more accurate than tone ($F(1,38) = 15.242$, $p < 0.0005$), short graphs are more accurate than long graphs ($F(1,38) = 19.069$, $p < 0.0005$) and task 1 is more accurate than task 2 and task 3 ($F(2,37) = 15.154$, $p < 0.0005$). We elaborate on these results below.

5.1 Tone vs. Speech

Tone (mean time = 43.53 seconds, SD = 1.81, 95% CI = 39.87 to 47.20) was found to be significantly faster than speech (mean time = 49.81 seconds, SD = 1.61, 95% CI = 46.55 to 53.08) ($F(1,38) = 8.093$, $p = 0.007$, $\eta_p^2 = 0.176$). At the same time, speech (mean error rank = 0.59, SD = 0.12, 95% CI = 0.35 to 0.83) was significantly more accurate than tone (mean error rank = 1.63, SD = 0.24, 95% CI = 1.13 to 2.12) ($F(1,38) = 15.242$, $p < 0.0005$, $\eta_p^2 = 0.286$). Thus, though tone improved user speed in comparison with speech to an extent, it did so at the cost of accuracy. The implication to design is that if it is important that the users interpret the data accurately, speech works better though task time may increase.

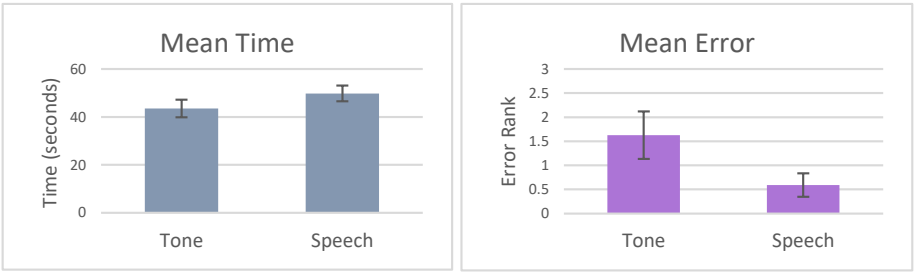

Fig. 2. Graphs showing mean time (seconds) and error rank for tone and speech modes.

5.2 Serial vs. Parallel

Parallel modality (mean time = 45.20 seconds, SD = 1.46, 95% CI = 42.24 to 48.15) was found to be significantly faster than serial modality (mean time = 48.15 seconds, SD = 1.48, 95% CI = 45.15 to 51.15) ($F(1,38) = 4.928$, $p = 0.032$, $\eta_p^2 = 0.115$). However, error was higher in the case of parallel modality (mean error rank = 1.26, SD = 0.23, 95% CI = 0.78 to 1.73) than serial modality (mean error rank = 0.96, SD = 0.12,

95% CI = 0.71 to 1.20), though the difference was not significant. Thus, the parallel modes performed significantly faster than the serial modes without affecting user accuracy by much.

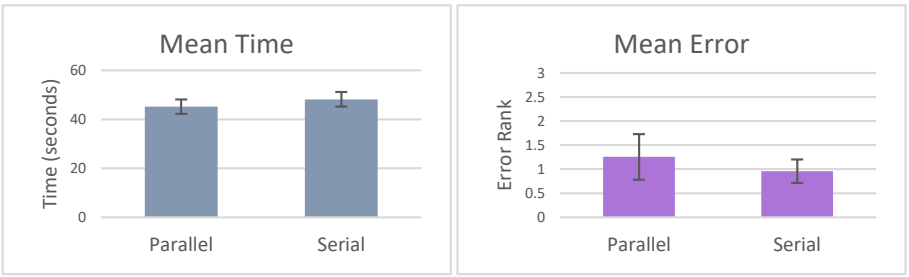

Fig. 3. Graphs showing mean time (seconds) and error rank for serial and parallel modes.

5.3 Graph Length

In case of short graphs, there were only 6 data points and users could rely solely on short term memory in order to answer after traversing the graph once. This is not the case with long graphs as there were 15 points and participants were required to go back and forth to find the correct answer. As expected, participants took significantly more time to navigate longer graphs (mean time = 58.71 seconds, SD = 1.86, 95% CI = 54.96 to 62.47) than they did shorter graphs (mean time = 34.63 seconds, SD = 1.12, 95% CI = 32.38 to 36.89) ($F(1,38) = 234.031$, $p < 0.0005$, $\eta_p^2 = 0.86$). Likewise, participants made significantly more errors with longer graphs (mean error rank = 1.62, SD = 0.25, 95% CI = 1.12 to 2.13) than shorter graphs (mean error rank = 0.59, SD = 0.07, 95% CI = 0.45 to 0.73) ($F(1,38) = 19.069$, $p < 0.0005$, $\eta_p^2 = 0.334$).

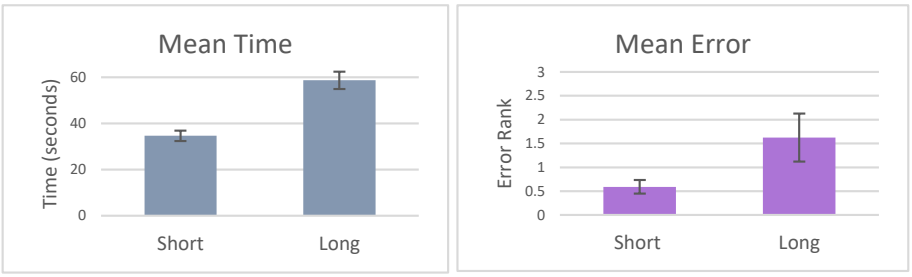

Fig. 4. Graphs showing mean time (seconds) and error rank for different graph lengths (short and long).

5.4 Task

All tasks were point estimation and point comparison tasks with varying levels of difficulty. The first task (highest point) required one answer, the second task (two most similar) required the participant to retain and compare multiple values to give two answers, and the third task (lowest three) required three answers. As expected, there were

significant differences in the ANOVA because of the tasks in time ($F(2,37) = 18.74$, $p < 0.0005$, $\eta_p^2 = 0.33$) and errors ($F(2,37) = 15.154$, $p < 0.0005$, $\eta_p^2 = 0.285$). Performing pairwise comparisons after applying Bonferroni adjustment for multiple comparisons, we found that task 1 (mean time = 39.76 seconds, SD = 1.34, 95% CI = 37.05 to 42.47) was significantly faster than task 2 (mean time = 50.97 seconds, SD = 2.00, 95% CI = 46.93 to 55.01) and task 3 (mean time = 49.29 seconds, SD = 1.82, 95% CI = 45.62 to 52.97) ($p < 0.0005$). Task 1 was also more accurate (mean error rank = 0.18, SD = 0.05, 95% CI = 0.07 to 0.28) than task 2 (mean error rank = 1.86, SD = 0.34, 95% CI = 1.18 to 2.54) and task 3 (mean error rank = 1.29, SD = 0.20, 95% CI = 0.89 to 1.69) ($p < 0.0005$). The differences between tasks 2 and 3 on task time and errors was not significant.

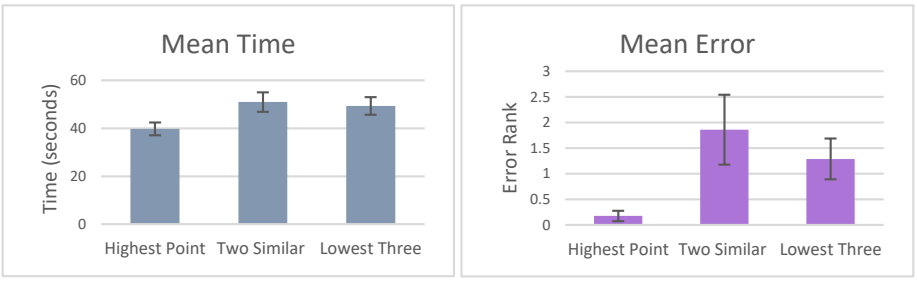

Fig. 5. Graphs showing mean time (seconds) and error rank for the three tasks.

5.5 User Type

The experiment was conducted with two groups of participants, sighted and visually impaired. Fig. 6 shows the mean time and error values for both these user types. Visually impaired participants were significantly faster (mean time = 42.90 seconds, SD = 1.86, 95% CI = 39.15 to 46.66) than their sighted counterparts (mean time = 50.44 seconds, SD = 1.86, 95% CI = 46.69 to 54.20)($F(1,38) = 8.246$, $p = 0.007$, $\eta_p^2 = 0.178$). While sighted users (mean error rank = 1.01, SD = 0.20, 95% CI = 0.61 to 1.41) were more accurate in their answers than visually impaired users (mean error rank = 1.21, SD = 0.20, 95% CI = 0.81 to 1.61), this difference is not significant.

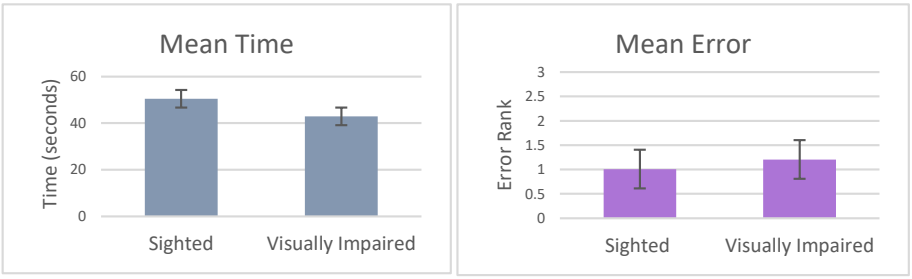

Fig. 6. Graphs showing mean time (seconds) and error rank for sighted and visually impaired participants.

5.6 Significant Interactions

In case of speed, we found four significant interactions, namely graph-length * user-type ($F(1,38) = 11.323$, $p = 0.002$, $\eta_p^2 = 0.23$), graph-length * tone-speech ($F(1,38) = 5.02$, $p = 0.031$, $\eta_p^2 = 0.117$), tone-speech * task ($F(2, 37) = 6.303$, $p = 0.003$, $\eta_p^2 = 0.142$) and tone-speech * graph-length * task ($F(2, 37) = 7.129$, $p = 0.001$, $\eta_p^2 = 0.158$). Fig. 7 represents these interactions in graphs. We note that there were no significant interactions between the two independent variables of interest in this paper - namely tone-speech and parallel-serial.

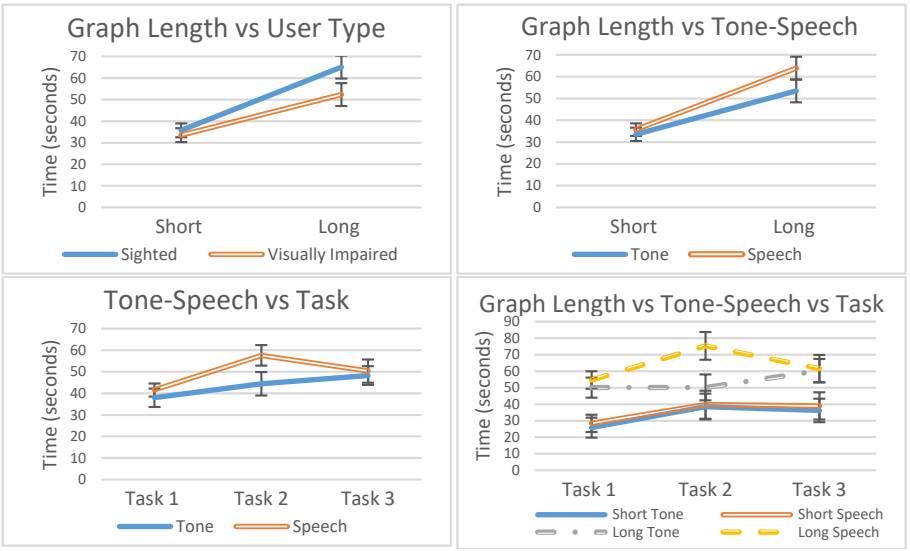

Fig. 7. Graphs showings significant interactions in speed.

In case of errors, we found two significant interactions, namely tone-speech * task ($F(2, 37) = 3.873$, $p = 0.025$, $\eta_p^2 = 0.092$) and graph-length * task ($F(2, 37) = 3.892$, $p = 0.025$, $\eta_p^2 = 0.093$). Fig. 8 represents these interactions in graphs. There were no significant interactions between the two independent variables of interest - namely tone-speech and parallel-serial.

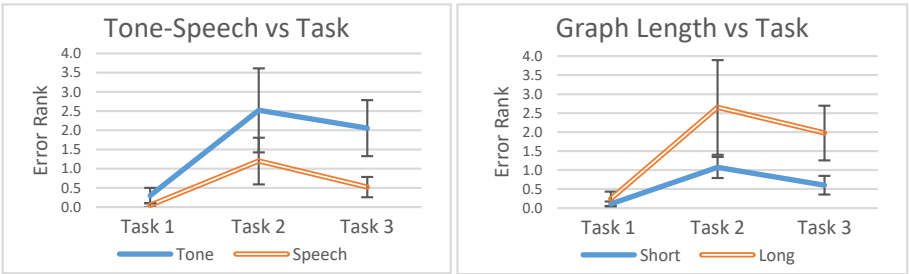

Fig. 8. Graphs showings significant interactions in error.

5.7 User Preference

User preference was calculated through a system usability score (SUS) [14] for each of the four modes, viz. parallel-tone, serial-tone, parallel-speech and serial speech (Fig. 9). In the case of sighted participants, two participants reported they were not confident enough on giving a rating as they had not paid enough attention to the technique while performing the tasks. These participants were dropped from the SUS evaluation and thus N = 18 for sighted users. All 20 visually impaired users gave SUS ratings confidently.

We performed a within subjects repeated measures ANOVA with SUS scores as dependent variables and parallel-serial (2), speech-tone (2) and user type (2, between subject) as independent variables. We found that users preferred speech (mean SUS = 83.63, SD = 1.97, 95% CI = 79.63 to 87.63) over tone (mean SUS = 61.25, SD = 2.80, 95% CI = 55.56 to 66.93) ($F(1,36) = 46.283$, $p < 0.0005$, $\eta_p^2 = 0.562$). Users also preferred serial modes (mean SUS = 77.51, SD = 2.05, 95% CI = 73.35 to 81.66) over parallel modes (mean SUS = 67.37, SD = 2.03, 95% CI = 63.25 to 71.48) ($F(1,36) = 26.157$, $p < 0.0005$, $\eta_p^2 = 0.421$). Visually impaired users gave significantly higher SUS ratings (mean SUS = 76.31, SD = 2.45, 95% CI = 71.34 to 81.29) compared to sighted users (mean SUS = 68.56, SD = 2.59, 95% CI = 63.32 to 73.80) ($F(1,36) = 4.732$, $p = 0.036$, $\eta_p^2 = 0.116$). This was probably because visually impaired users are a lot more familiar with audio-interfaces in general than sighted users.

We found only one significant interaction, namely parallel-serial * user-type ($F(1,36) = 5.775$, $p = 0.022$, $\eta_p^2 = 0.138$). Fig. 9 represents these interactions in a graph. We note that there were no significant interactions between the two independent variables of interest in this paper - namely tone-speech and parallel-serial.

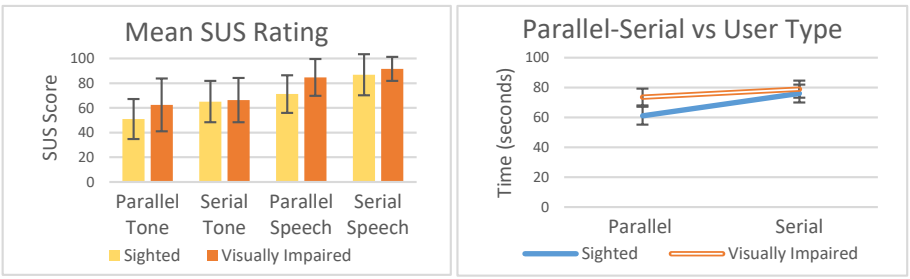

Fig. 9. Mean SUS rating (out of 100) and significant interactions in SUS ratings by participants.

5.8 Qualitative Findings

Most participants found tone modalities harder, especially for values that were numerically close to each other. They could not identify or remember the tones for the comparison based tasks or recall the scale. One participant said that though tone modalities took less time to play out, it took longer to analyze the data as comprehension still took

more time. Participants stated that tone modalities were easy for comparing adjacent points, but difficult for longer graphs.

Likewise, users found it harder to concentrate in parallel modalities as they had to pay attention to both ears. Users noted that this might be hard when they are listening casually, while multitasking, or in places with noise. Interestingly, users found parallel speech more distracting than parallel tone. In spite of this, the response for parallel-speech was generally positive as participants found that it was fast and conveyed all the information necessary. Users found parallel tone acceptable only for adjacent values. For longer graphs, this was much harder.

Participants made several design suggestions. For example, using different instruments in the tonal modalities to indicate data values. A participant also suggested using different voices to indicate the magnitude of a data point as well as changing the pitch of voices. Previous research has explored using multiple instruments to denote multiple data series [3]. It would be worth exploring a technique where different instruments are used to denote the magnitude of the data.

We noticed that users rarely used the up arrow to repeat a bar, preferring instead to go back and forth with left and right arrow keys, leading to more repetitions. When asked why, they mentioned it helped get a better relational reference.

6 Discussion

Tone modalities were faster than speech modalities. Through this study, we found that while the tone modalities have substantially less recorded time than speech modalities, this results in only a small amount of saving in the actual task time. Further, this saving comes at a cost of significantly higher error rates and lower user preference. Tones are a poor substitute for speech to communicate numeric data, but they provide a way to reach the required data quickly. While prior work has found that tone modalities may provide context and help in augmenting the information in graphs [15], our users felt that tones do not convey the complete information and demand higher attention. User preference was higher for speech modalities than tonal modalities.

Parallel modalities showed more promise, though they too have a price. As expected, the parallel modalities were faster than the serial modalities in recorded time. Our studies show that task times have some gains in parallel modalities compared to serial modalities, and although accuracy is degraded, the difference is not significant. These results are consistent with earlier studies [2] which showed the effectiveness of using parallel audio streams. While users did not prefer parallel modalities over serial modalities because of increased attention demands, participants were nonetheless positive about parallel modalities, as they conveyed all the information in a short time.

We acknowledge some limitations in our studies. Our participants were familiar with speech technologies, including conversational agents such as Google Assistant [43] and Siri [45], and in case of visually impaired users with screen readers such as TalkBack [44]. None of the participants had prior familiarity with tones which depict numerical values. Our studies did not give the users an opportunity to practice the tone and parallel modalities. The advantages of using these modalities may become more evident after practice.

Our studies have identified opportunities for future work. Firstly, we focussed on bar graphs only. Several other types of graphs could be explored. The tasks performed in our studies were point estimation and point comparison tasks. Choosing a different task could lead to different results. We did not explicitly evaluate the effect of practice of the first task on subsequent tasks. This is something that could be systematically controlled for, and explored. It would be interesting to evaluate interfaces which switch between modalities, either automatically or manually. There was a significant increase in task time and errors from short graphs to long graphs. It will be interesting to explore techniques required for very long graphs containing hundreds of points. Future studies could study more interactive graph navigation techniques, data sorting and filtering techniques, and/or voice input based interactions.

7 Conclusion

In this paper we presented a systematic evaluation of tone and speech and parallel and serial modalities for auditory bar graphs in four modes - parallel tone, serial tone, parallel speech, and serial speech. We found that although tone was consistently faster than speech, it was less accurate. Parallel modalities were significantly faster than serial modalities without affecting accuracy significantly. We also found that users preferred serial modalities over parallel modalities and speech modalities over tone modalities. Serial speech mode was the most preferred. To the best of our knowledge, this is the first paper that presents a systematic comparison of these modes. We hope this work will be beneficial to creators of auditory graphs and pave the way towards new methods of effective auditory graph generation.

Acknowledgements. This research was supported by Tata Consultancy Services. We would like to thank all the participants who took part in this study for their valuable feedback. We would also like to thank the anonymous reviewers of this paper for their reviews which helped shape the final version of this paper.

References

1. Mansur, D. L.: Graphs in sound: A numerical data analysis method for the blind (No. UCRL-53548). Lawrence Livermore National Lab., CA (USA) (1984).
2. Brown, L., Brewster, S., Ramloll, R., Yu, W., & Riedel, B.: Browsing modes for exploring sonified line graphs. In IN PROCEEDINGS OF BCS-HCI (2002).

3. Brown, L. M., & Brewster, S. A.: Drawing by ear: Interpreting sonified line graphs. Georgia Institute of Technology (2003).
4. Brown, L. M., Brewster, S. A., Ramloll, S. A., Burton, R., & Riedel, B.: Design guidelines for audio presentation of graphs and tables. International Conference on Auditory Display (2003).
5. Flowers, J. H.: Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions. Georgia Institute of Technology (2005).
6. Kildal, J., & Brewster, S. A.: Providing a size-independent overview of non-visual tables. In 12th International Conference on Auditory Display (ICAD2006) (pp. 8-15) (2006, June).
7. Peres, S. C., & Lane, D. M.: Sonification of statistical graphs. Georgia Institute of Technology (2003).
8. Ramloll, R., Brewster, S., Yu, W., & Riedel, B.: Using non-speech sounds to improve access to 2D tabular numerical information for visually impaired users. In People and Computers XV—Interaction without Frontiers (pp. 515-529). Springer, London (2001).
9. Walker, B. N., & Mauney, L. M.: Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. ACM Transactions on Accessible Computing (TACCESS), 2(3), 12 (2010).
10. Walker, B. N., & Lane, D. M.: Psychophysical scaling of sonification mappings: A comparison of visually impaired and sighted listeners. Georgia Institute of Technology (2001).
11. Bonebright, T. L., Nees, M. A., Connerley, T. T., & McCain, G. R.: Testing the effectiveness of sonified graphs for education: A programmatic research project. Georgia Institute of Technology (2001).
12. Nees, M. A., & Walker, B. N.: Encoding and representation of information in auditory graphs: Descriptive reports of listener strategies for understanding data. International Community for Auditory Display (2008).
13. Peres, S. C., & Lane, D. M.: Auditory graphs: The effects of redundant dimensions and divided attention. Georgia Institute of Technology (2005).
14. Brooke, J.: SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7 (1996).
15. Smith, D. R., & Walker, B. N.: Effects of training and auditory context on performance of a point estimation sonification task. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 48, No. 16, pp. 1828-1831). Sage CA: Los Angeles, CA: SAGE Publications (2004, September).
16. Resources on Alternative Text for Images, https://www.w3.org/WAI/alt/, last accessed 2019/01/20
17. Nees, M. A., & Walker, B. N.: Relative intensity of auditory context for auditory graph design. Georgia Institute of Technology (2006).
18. Brewster, S. A., Wright, P. C., & Edwards, A. D.: Experimentally derived guidelines for the creation of earcons. In Adjunct Proceedings of HCI (Vol. 95, pp. 155-159) (1995, August).
19. Nees, M. A., & Walker, B. N.: Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension. Proceedings of ICAD 2007 (2007).
20. Wall, S. A., & Brewster, S.: Sensory substitution using tactile pin arrays: Human factors, technology and applications. Signal Processing, 86(12), 3674-3695 (2006).
21. Horstmann*, M., Lorenz, M., Watkowski, A., Ioannidis, G., Herzog, O., King, A., ... & King, N.: Automated interpretation and accessible presentation of technical diagrams for blind people. New Review of Hypermedia and Multimedia, 10(2), 141-163 (2004).
22. Jankun-Kelly, T. J., Ma, K. L., & Gertz, M.: A model and framework for visualization exploration. IEEE Transactions on Visualization and Computer Graphics, 13(2), 357-369 (2007).

23. Cowan, N.: The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences, 24(1), 87-114 (2001).
24. The World Bank, https://data.worldbank.org/, last accessed 2019/05/01.
25. Sonic Pi - The Live Coding Music Synth for Everyone, https://sonic-pi.net/, last accessed 2019/05/01.
26. Animaker Voice, Create free human-like voice overs for your videos, https://www.animaker.com/voice, last accessed 2019/05/01.
27. Pollack, I., & Ficks, L.: Information of elementary multidimensional auditory displays. The Journal of the Acoustical Society of America, 26(2), 155-158 (1954).
28. Bly, S.: Presenting information in sound. In Proceedings of the 1982 conference on Human factors in computing systems (pp. 371-375). ACM (1982, March).
29. Yeung, E. S.: Pattern recognition by audio representation of multivariate analytical data. Analytical Chemistry, 52(7), 1120-1123 (1980).
30. Web Accessibility Tutorials, https://www.w3.org/WAI/tutorials/images/complex/, last accessed 2019/05/01.
31. The TeDUB-Project (Technical Drawings Understanding for the Blind), https://www.alasdairking.me.uk/tedub/index.htm, last accessed 2019/05/01.
32. Brown, A., Pettifer, S., & Stevens, R.: Evaluation of a non-visual molecule browser. In ACM SIGACCESS Accessibility and Computing (No. 77-78, pp. 40-47). ACM (2004, October).
33. Kennel, A. R.: Audiograf: A diagram-reader for the blind. In Proceedings of the second annual ACM conference on Assistive technologies (pp. 51-56). ACM (1996, April).
34. Challis, B. P., & Edwards, A. D.: Design principles for tactile interaction. In International Workshop on Haptic Human-Computer Interaction (pp. 17-24). Springer, Berlin, Heidelberg (2000, August).
35. Carpenter, P. A., & Shah, P.: A model of the perceptual and conceptual processes in graph comprehension. Journal of experimental psychology: applied, 4(2), 75 (1998).
36. Stockman, T., Nickerson, L. V., & Hind, G.: Auditory graphs: A summary of current experience and towards a research agenda. Georgia Institute of Technology (2005).
37. Gardner, J. A., Lundquist, R., & Sahyun, S.: TRIANGLE: a tri-modal access program for reading, writing and doing math. In Proceedings of the CSUN International Conference on Technology and Persons with Disabilities, Los Angles. Converging Technologies for Improving Human Performance (pre-publication on-line version) (Vol. 123) (1998, September).
38. Zhao, H., Plaisant, C., Shneiderman, B., & Lazar, J.: Data sonification for users with visual impairment: a case study with georeferenced data. ACM Transactions on Computer-Human Interaction (TOCHI), 15(1), 4 (2008).
39. Flowers, J. H., Whitwer, L. E., Grafel, D. C., & Kotan, C. A.: Sonification of daily weather records: Issues of perception, attention and memory in design choices. Faculty Publications, Department of Psychology, 432 (2001).
40. Deutsch, D.: Grouping mechanisms in music. In The psychology of music (pp. 299-348). Academic Press (1999).
41. Bregman, A. S.: Auditory scene analysis: The perceptual organization of sound. Cambridge, MA, US (1990).
42. Audacity - Free, open source, cross-platform audio software, https://www.audacityteam.org/, last accessed 2019/05/01.
43. Google Assistant, https://assistant.google.com, last accessed 2019/05/01.
44. Talkback - Google Accessibility, https://www.google.com/accessibility/, last accessed 2019/05/01.
45. Apple Siri, https://www.apple.com/siri/, last accessed, 2019/05/01.

46. Processing Programming Language, https://processing.org/, last accessed 2019/05/01.
47. McGookin, D., Robertson, E., & Brewster, S.: Clutching at straws: using tangible interaction to provide non-visual access to graphs. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1715-1724). ACM (2010, April).
48. Gomez, C. C., Shebilske, W., & Regian, J. W.: The effects of training on cognitive capacity demands for synthetic speech. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 38, No. 18, pp. 1229-1233). Sage CA: Los Angeles, CA: SAGE Publications (1994, October).
49. Flowers, J. H., & Hauer, T. A.: "Sound" alternatives to visual graphics for exploratory data analysis. Behavior Research Methods, Instruments, & Computers, 25(2), 242-249 (1993).
50. De Cheveigne, A.: Pitch perception models. In Pitch (pp. 169-233). Springer, New York, NY (2005).
51. Newman, E. B., Stevens, S. S., & Davis, H.: Factors in the production of aural harmonics and combination tones. The Journal of the Acoustical Society of America, 9(2), 107-118(1937).