January 2012

Overcoming Limitations of Serial Audio Search

Isabela Cordeiro Ribeiro Moura Hidalgo
University of South Florida, isamoura@yahoo.com

Overcoming Limitations of Serial Audio Search

by

Isabela C. R. M. Hidalgo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Dewey Rundus, Ph.D.
Rafael Perez, Ph.D.
Les Piegl, Ph.D.
Douglas Rohrer, Ph.D.
Patricia Zarate, Ph.D.

Date of Approval:
March 9, 2012

Keywords: Parallel Output, Simultaneous Sounds, Music Browser, Wii Remote, Human-Computer Interaction, Auditory User Interfaces

Dedication

This dissertation is dedicated to my family, for their support, encouragement, inspiration, and infinite patience.

Table of Contents

List of Tables

List of Figures

Abstract

The typical approach for finding audio recordings, such as music and sound effects, in a database is to enter some textual information into a search field. The results appear summarized in a list of textual descriptions of the audio files along with a function for playing back the recordings. Exploring such a list sequentially is a time-consuming and tedious way to search for sounds. This research evaluates whether searching for audio information can become more effective with a user interface capable of presenting multiple audio streams simultaneously.

A prototype audio player was developed with a user interface suitable for both search and browsing of a hierarchically organized audio collection. The audio recordings are presented either serially (serial output mode) or simultaneously (parallel output mode), spatially distributed in both vertical and horizontal planes. Users select individual recordings by simply pointing at its source location with a remote control.

Two within-subjects experiments were conducted to compare the performance of the audio player's output modes in audio search tasks. The experiments differ in the maximum number of audio recordings played simultaneously – either four or six. In both experiments, search tasks were performed about 25% faster using parallel audio output than using serial output. Over 80% of participants preferred searching parallel output.

The results indicate that using parallel output can be a valuable improvement to the current methods of audio search, which typically use only serial output.

Chapter 1

Introduction

1.1 Motivation

Computer technology has contributed to the wide availability of audio recordings in digital form, creating an increasing need for strategies that enable easier access to large audio collections. Today, audio is downloaded from websites and stored in users' computers and portable devices. The classic iPod (Apple Inc., 2011) can store up to 40,000 songs. The smaller version of this popular portable music device, the iPod Shuffle, has no visual display, and can store up to 1,000 songs. The large storage capacity of portable media players and their decreasing physical sizes introduced many challenges for user interface designers.

Music streaming and downloading websites, such as Amazon (Amazon.com Inc., 2011) that has over 11 million songs available for download, typically index their music only by metadata – textual descriptions such as the artist's name, the song's title, and genre.

Websites that offer sound effects, such as SoundBible (SoundBible.com, 2011) and FindSounds (Comparisonics Corporation, 2011), have tens of thousands of sounds in their catalogs, also indexed by metadata.

The usual approach for finding audio data online is to enter some of the metadata into a search field. The results appear summarized in a list of textual information along with a function for playing back selected audio files or all audio files in order. Keyword-based search is quite powerful for users who know the right words to use in a query, but since most music and sound effects are difficult to describe with words, the usability of this technique is limited. Some irrelevant sounds are retrieved while some relevant ones are not included in the result set.

When only a little information is provided as keywords, for instance the name of the performer of a musical piece, the result set can be very long. As an example, a search for mp3 downloads performed by artist “Neil Diamond” on Amazon returns over 900 recordings. This number does not include covers (songs written by Neil Diamond but recorded by other artists). Exploring such a list is a time consuming process especially when there are many unknown songs that need to be heard one by one before being recognized.

Even for sounds effects, the result list can be too long to be practical. A search for “crash” on the FindSounds website returns about 100 different crash sounds. The textual descriptions of these sounds are often too vague – in this example, most of them are simply “crash” – forcing the users to listen to many different sounds before finding the one they want. Listening to audio recordings sequentially is a long and monotonous way to search for sounds.

The systems for audio retrieval available today do not provide a good solution for browsing through long result sets and for users who do not know exactly what they are seeking. Research activity in audio information retrieval has focused primarily on content-based methods for search. Many approaches using Query-by-Humming (QBH) and Audio Fingerprinting have been proposed for searching audio by user-sung melodies or recorded portions of audio recordings, respectively. These methods can be useful when metadata is unavailable or unknown, but do not address the user who simply wishes to explore an audio collection.

In a recent survey of audio information retrieval (Lew, Sebe, Djeraba, & Jain, 2006), the authors suggest exploration systems as a major research challenge: “We should focus as much as possible on the user who may want to explore instead of search for media” (p.12). They also encourage research in relevance feedback methods and development of systems that interactively learn the needs of the users.

Casey et al. (Casey et al., 2008) note that the majority of the research in the field is engineering-led. They point out the need for user studies that evaluate the way music information retrieval (MIR) tools get used by untrained users and that offer a better understanding of how users navigate million-song music databases. They encourage new research in user interaction design that attempt to give users more control of the search experience.

1.2 Approach

This research study aims to find how audio data should be presented to ensure effective search and browsing. One way in which the time to present audio information can be reduced is through concurrent playback. In everyday life, people process a large amount of information simultaneously. For example, one focuses their visual attention on the road while driving without losing track of things happening in their peripheral field of vision. When browsing the web on a visual display, people can obtain a general idea of the content without having to focus on every picture or word individually.

This work aims to allow the same kind of parallel presentation for audio data, including music. Listening to multiple audio recordings at the same time also allows a real time comparison of the data. For simplicity, this dissertation will use the term "song" to refer to any musical piece, with or without lyrics.

The main research question addressed by this dissertation is the following: *Is the effectiveness of an auditory search task affected when multiple sounds are presented simultaneously?* Effectiveness is measured by the time and distance (number of steps) taken to complete the search successfully. It was anticipated that an interface that presents the audio simultaneously would allow for faster and shorter searches than an interface that only presents the audio serially.

In order to answer the research question, an audio player was developed that allows users to interact with an audio collection by listening to the recordings, presented either simultaneously (parallel output mode) or serially (serial output mode). The audio recordings are spatially distributed in both vertical and horizontal planes, and users are able to select an individual recording by simply pointing at its source location.

A within-subjects study was carried out to compare the performance of the audio player's output modes and analyze user behavior when performing search tasks.

1.3 Contributions

This dissertation introduces the concept of *searching parallel output* applied to audio search and discovery. The contributions of this work are:

- It provides evidence that parallel audio output can be used to overcome some limitations of serial audio search. It allows listeners to quickly gain insight into a large database of unfamiliar audio recordings, including music, and perform similarity-based audio searches more effectively when compared to traditional serial audio output.
- It introduces and evaluates a reduction technique for audio browsing that facilitates the elimination of uninteresting items from a set of presented recordings. This technique can be used for obtaining relevance feedback in future user studies.
- It offers a better understanding of users' browsing behavior by tracing their steps during the navigation of the audio collection.
- It produced a working prototype for an audio player that is fun and easy to use, and is the first to:

- use both vertical and horizontal planes to position audio sources in space;
- use asynchronous onset for playing simultaneous recordings;
- offer a reduction technique for browsing;
- use a remote control as the input device, which enables pointing at spatial positions to select the audio sources;
- be formally evaluated in a study with random users.

1.4 Organization of the Dissertation

The next chapters are organized as follows. Chapter 2 points out the relationship between searching and browsing, present some background on current techniques for searching multimedia collections, particularly audio, followed by a review of visualization approaches that support browsing. Finally, some concepts on the human perception of sounds are presented as the basis for the successful implementation of the approach presented in this dissertation.

Chapter 3 introduces the Parallel Audio Player, an application capable of playing six audio recordings simultaneously through spatially distributed loudspeakers, and receiving user input via a remote control adapted from the Nintendo's Wii video game.

Chapter 4 describes the two experiments performed to answer the research question. The experiments' participants used the Parallel Audio Player to browse a hierarchically organized audio collection searching for specific audio recordings.

Chapter 5 explains the results of the two experiments and Chapter 6 presents a discussion of these results, followed by a description of some limitations of the study and future work.

Chapter 2

Background and Related Work

2.1 Searching and Browsing

As defined by Spence (2001) browsing occurs "when a user scans a display to see what is there". The user is not necessarily searching for anything specific, but wants to have an idea of what is available. The activity is not limited to the passive viewing of a fixed display. Browsing can be interactive, as users usually formulate new browsing strategies as they interpret the information being displayed and find it necessary to continue exploring. Hierarchical and zoomable displays (Bederson & Shneiderman, 2003) are examples of user interfaces that support interactive browsing.

A search activity occurs when a user gives some specification of what is wanted and the retrieval system finds and brings back the information (Baeza Yates & Ribeiro, 1999). When the user is able to say exactly what is being sought, an effective query can be formulated and a good search engine will return relevant matches.

However, even well formulated queries often return a huge set of results that will require some browsing until the user recognizes the document being sought or finds a way to refine the initial query.

For example, when searching for pictures of houses on Yahoo! Image Search (Yahoo! Inc., 2011), one writes the word "house" as the query term. Over 100 million images are found and are displayed in some order of relevance determined by the search algorithm. The first five images are of the cast of the medical television show "House", quite possibly not relevant to the user. The user has the options of browsing through the results until finding a good match or formulating a new query. One way to refine the query is to add more specific terms, for instance "brick house", or to reduce the matching set by using the NOT operator, usually represented by the minus sign (-). The query "house –television" requests that images that match the word "television" are not included in the results. This reduction technique eliminates some distracters from the result set. The approach presented in this dissertation tests a reduction technique for searching and browsing audio.

2.2 Image and Video Search

Despite the great amount of work that has been done in the last decade in multimedia information retrieval, exploring a large multimedia database remains an open issue. The majority of the research focused on content-based image and video retrieval, rather than audio, which remains almost entirely based on keywords.

In a study carried out by Tjondronegoro and Spink (2008), over 100 commercial web search engines were examined and less than 1% was found to support content-based search with queries that use examples other than textual keywords as input.

Image search technologies are becoming more mature and commercially available in search engines such as Google Images (Google, 2011) and Microsoft Bing (Microsoft, 2011), where a feature is now offered to find images that are similar to any of the results from an initial query. Figure 1 shows a screen capture of some Google Images search results for the query "House –tv". When the user rests the mouse pointer on one of the results, more information about that image is displayed, along with the link "Similar", which updates the result set when clicked. A comparable system for hierarchical video browsing has been proposed in (Xingquan Zhu, Elmagarmid, Xiangyang Xue, Lide Wu, & Catlin, 2005), in which the users refine their query progressively by choosing to find similar video sequences to a selected video.

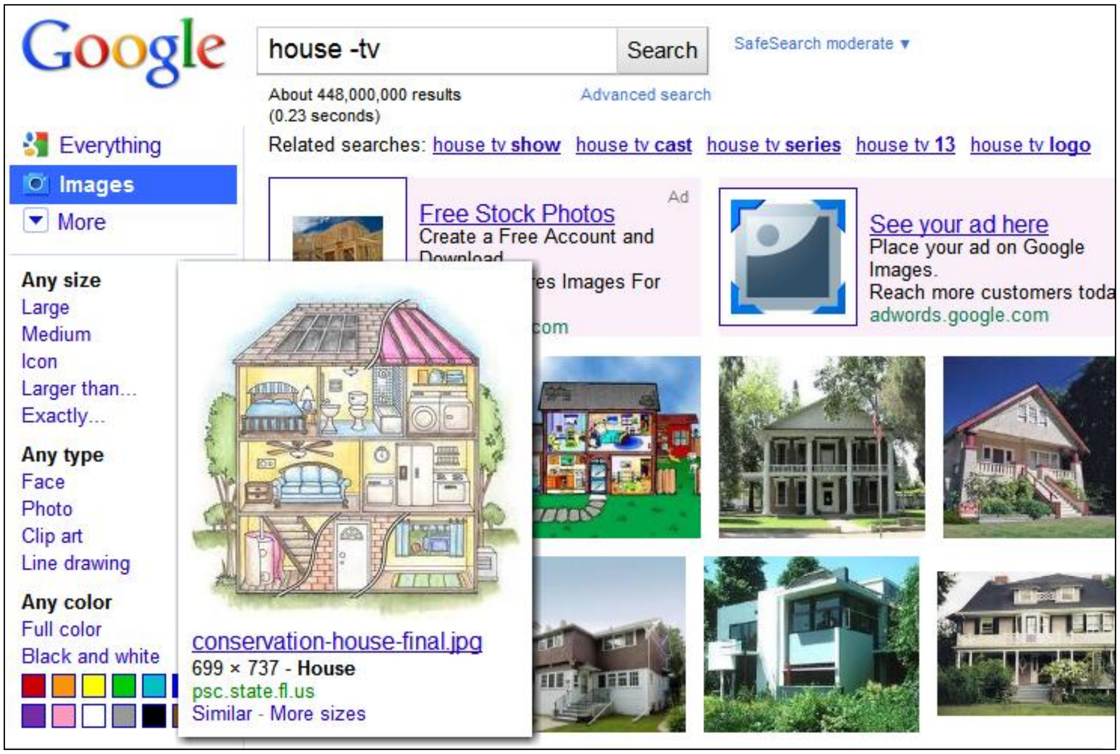

Figure 1 – Result set for the search query "House –tv"

Videos and images can be visually presented in parallel and users can effectively scan the result set to gain an overview of the contents. However, with audio content, a quick look at a list of the textual descriptions will not offer the same identification power as hearing a small piece of each audio file. There is a need for more effective ways to present audio data, and some of the latest research efforts in this area are summarized in section 2.4. This dissertation focuses on the parallel presentation of audio data in an auditory format.

2.3 Audio Search

This section presents a review of research related to the way audio recordings' searchable features are extracted and compared to the input of queries.

A text-based query for audio can be successful when well-defined textual descriptions are associated with the audio files. Sound effect classification, genre classification, and recommendation systems focus on producing such descriptions.

Audio fingerprinting and query-by-humming are content-based methods that aim to enable audio retrieval when text annotations do not exist or are not complete enough to provide accurate and efficient matches (Lew et al., 2006). Instead of relying only on metadata, content-based systems use information about the acoustic attributes of the recordings in their index. The idea is to allow "sounds like this" searches, by using audio examples in the query.

2.3.1 Sound Effect Classification

The textual descriptions that are associated to sounds are normally tagged by librarians. Sounds are usually placed in categories, such as animals, people, or tools. Users typically search for sounds by keyword matching or browsing category trees. The correct labeling and placement of sounds into categories is an imprecise and time-consuming task, due in part to the ambiguity of natural language and the lack of a widely accepted convention used to describe sounds.

A way to describe sounds is onomatopoeia, the formation of words to imitate sounds (buzz, crash, ring). In a letter written in 1913, Luigi Russolo categorized sounds into six groups of noises (Russolo, 2001):

- roars, thunders, bangs, booms;
- whistles, hisses, puffs;
- whispers, murmurs, mumbles, mutters, gurgles;
- screeches, creaks, rustles, crackles;
- noises obtained by beating on metal, wood, skin, stone, pottery;

- voices of animals and people: shouts, screams, shrieks, wails, hoots, howls, sobs.

However, many of these descriptions convey little information. Two people will most likely produce different sounds if asked to make the sound of a hoot, or a roar for example. Besides, onomatopoeia does not translate easily to other languages, since the words used for some sounds can be quite different in different parts of the world.

Another way to describe sounds is by semantic descriptors, which refer to the source of a sound. This approach is easier than describing the sound itself, but is less useful if users are unfamiliar with the sounds, for example, “iceberg breaking” or “toucan vocalizing”.

2.3.2 Genre Classification

Musical genre is considered a key descriptor when people define their musical preference. Similarly to what happens in sound effect classification, musical genre classification is typically performed manually. Music retailers tend to categorize artists and albums, instead of single tracks, which can distort search results when an album has one or two songs that are different from the rest, or is a compilation of different genres such as a soundtrack.

A variety of hierarchical taxonomies of musical genres is currently in use by music websites. Pachet and Cazaly (2000) analyzed three large taxonomies – Amazon.com (Amazon.com Inc, 2011), All Music Guide (Rovi Corporation, 2011), and MP3.com (CBS Radio Inc., 2011) – and found many inconsistencies in both the labels used and the semantic relations between genres and sub-genres. The lack of consistency appeared not only between taxonomies, but also within each one. In the same article, they propose a new taxonomy to classify individual tracks based on their similarity.

In an attempt to reduce the inconsistency and time constraints introduced by manual taxonomy creation, Tzanetakis and Cook (2002) proposed an automatic genre classification system. Their system organizes a music collection into hierarchical genres and has comparable performance to genre classification done by human users.

2.3.3 Recommendation Systems

A variety of tags can be used in addition to genre to describe music and measure similarity. An example of an online music recommendation system that uses annotations created by a group of music experts is Pandora (Pandora Media, 2011). Pandora’s experts tag each song from a set of 400 attributes. The consistency achieved by having a specialized group of people creating the metadata comes at a high cost. It is estimated that each expert takes about 20-30 minutes to analyze one track and write the metadata. Before entering the music catalogue, each track must be analyzed by more than one expert.

Last.fm (Last.fm Ltd., 2011) is an example of a social music website that trades quality and consistency for quantity of tags by allowing the public to contribute descriptions and ratings to their music database. The system uses this information to recommend music to other users. One problem with this approach is that user generated tags many times represent ineffective opinions, such as "awesome" or "boring". However, the major disadvantage of these approaches is what is known as the cold start problem (Levy & Sandler, 2009): only music that has been recommended for listening can be tagged, but only music that has been tagged can be recommended. This problem makes it more difficult for brand-new music to be discovered.

Slaney and White (2007) described a method that compares users' ratings of two songs to compute the similarity between the music. Their study, which used ratings from over 380,000 users, suggests that user preference data can be a more accurate measure of similarity than acoustic data when a large number of users are available and actively rating the musical pieces.

Implicit feedback, provided by user behavior such as skipping songs (Pampalk, Pohle, & Widmer, 2005), has been used to improve the quality of tags in recommendation systems. An approach that uses facial expressions and gestures as feedback has been proposed in (Vinciarelli, Suditu, & Pantic, 2009).

2.3.4 Audio Fingerprinting

Audio fingerprinting is an approach that uses a sample of the recording as the query. Acoustic features are used to compare the recordings. The identification task should return information such as the name of the recording and a description. A limitation of this type of audio identification is the difficulty in matching samples that are not identical to the recordings in a database, such as different recordings of the same sound effect, live versions of a song by the same artist or recordings of the same song by a different artist. Recent research in music fingerprinting focusing in cover song identification is summarized in (Serrà, Gómez, & Herrera, 2010).

A popular program that uses music fingerprinting methods for song identification is Shazam (Wang, 2006). Users can record a sample of the music with their cellular phones and send it to a server for identification. Much of their research has focused on creating robust recognition algorithms that can handle the distortion and background noise found in the audio samples.

2.3.5 Query-by-Humming

In early QBH research, Ghias, Logan, Chamberlin, & Smith (1995) note that "a natural way of querying an audio database is to hum the tune of a song". Several techniques have since been attempted to match audio recordings to a sample of a person humming, singing, or whistling a melody. The hummed melody is transformed into a symbolic representation, which is used to query a database of melody representations.

Most QBH techniques compare a monophonic query (one voice) to monophonic melodies in the database. Since most sound effects and music has multiple instruments and voices happening in parallel, monophonic melodies must first be extracted. Extracting these melodies directly from the audio is difficult and unreliable. MIDI (musical instrument digital interface) files are symbolically encoded music, and have been used in many QBH systems (Birmingham, Dannenberg, & Pardo, 2006).

Dannenberg et al. (2007) did a comparative study of various approaches and found that the performance of the systems is sensitive to the melody representation and the distance functions used by the matching algorithms, and very sensitive to the quality of the queries. Real world queries from the average, non-musically trained, users are in the majority, full of pitch errors and external noises, making them difficult to transcribe. Unal, Narayanan, & Chew (2004) found that individuals may not recall the tune correctly and are likely to have problems producing the correct pitch.

In an approach that attempts to eliminate the problem of comparing monophonic melodies to original recordings, Tunebot (Little, Raffensperger, & Pardo, 2007) compares a melody sang by a user with a database of melodies contributed by other users, and returns the 50 closest matches ranked by similarity (Northwestern University Interactive Audio Lab, 2010). The system also learns from the feedback provided by the user on the search results. Tunebot's main disadvantage is that it depends on user input to populate the database. In addition, the way two individuals sing the same tune may have considerable differences.

Using the same concepts and user input strategies as Tunebot, the commercial mobile application SoundHound (SoundHound Inc., 2011), and its online version Midomi

(Melodis Corporation, 2011) have a larger database of contributed samples and therefore better retrieval performance. Yet it only returns a few possible matches back to the user, often failing to match hummed samples even when they are present in the database.

In all QBH systems examined, the user has to listen to the returned matches that cannot be recognized with the presented metadata, one at a time, just to realize the searched tune is not in the list.

2.3.6 Conclusion

The approaches presented in this section have in common the fact that they all use some kind of similarity measure to classify and retrieve audio recordings. Similarity-based organization allows users to explore sounds that are similar to something they know.

However, perceived similarity between items varies between people and is often dependent on context. People may find two songs similar because they remind them of a specific person or time of their lives even if the songs are significantly different acoustically.

In addition, according to Selfridge-Field (2000), when judging the similarity of musical pieces untrained music listeners are heavily influenced by tempo which is an attribute of a performance, not the composition.

This ambiguity in human sound perception represents a challenge for audio classification systems. Novel user interfaces need to compensate for the deficiency in similarity-based classification schemes by facilitating the presentation, browsing, and management of large audio catalogues. Research on audio visualization interfaces has been encouraged by the MIR community and is summarized in the next section.

2.4 Audio Visualization

Several interfaces that rely on visualizations other than textual lists of bibliographical information have been proposed for exploring audio libraries. Many of them use self-organizing maps (SOMs), which are unsupervised neural networks, to arrange audio recordings on a map so that similar pieces are grouped together. Most of the visualization work has focused on displaying music, but the concepts could be applied to sound effects as well.

PlaySOM (Neumayer, Dittenbach, & Rauber, 2005) is a visual interface that displays music in a 2-dimensional geographical map and allows a user to move across the map and zoom into regions to select music to play. The music is clustered on the map according to similarity using content-based methods.

Risi, Mörchen, Ultsch, & Lehwark (2007) describe a similar interface where a music collection is displayed on a topographic map, but their method for determining the similarity of the music uses tags from Last.fm instead of acoustic features.

Knees, Schedl, Pohle, & Widmer (2006) describe another SOM based visual interface that includes pictures and text (metadata) on the maps to aid in the identification of the sounds. They also propose the usage of a classic video game controller instead of a mouse to interact with the interface.

A user interface for small devices was proposed by (Vignoli, van Gulik, & van de Wetering, 2004). It displays circles that represent artists, clustered by similarity. Mood, tempo, and year of release are used as similarity attributes, represented by different colors and spatial location.

MusicRainbow, presented by Pampalk & Goto (2006), also uses colors to encode different music styles. Their algorithm computes similarities between artists, based on acoustic features of their tracks. The interface displays similar artists near each other on a rainbow consisting of eight concentric circles that have different colors to represent different types of music. Word labels that summarize the artists' qualities are also displayed. The authors did not conduct a formal user evaluation of their system.

Torrens and Arcos (2004) present a hierarchical graphical interface where a music collection can be visualized as a Tree-Map, a technique described by Johnson and Shneiderman (1991). Metadata is used to classify the music. Genres are displayed as different rectangles with sizes proportional to the number of tracks of that genre. The rectangles can be divided into sub-genres, which can be split into individual artists. The color of each rectangle can denote one of a few possible attributes, chosen by the user.

The user interfaces mentioned in this section have similar objectives to the one proposed in this dissertation: to facilitate audio discovery, search, and browsing. The main differences are the lack of parallelism for audio presentation and the dependence on a visual display, which complicate their use in situations where the user's visual attention is engaged in other tasks, such as driving, reading, or walking.

2.5 Overview of Spatial Hearing

The proposed user interface for music browsing relies on the human ability to listen to multiple sounds at the same time and identify the spatial location where a specific sound originates. This section briefly describes some of the human capabilities and limitations that influenced the design of the proposed approach.

2.5.1 Sound Localization

Sound localization refers to the identification of the position (direction and distance) of a sound source (Grantham, 1995). Human listeners have the ability to localize sounds with remarkable precision (Moore, 1989). This dissertation work particularly depends on the ability of determining the direction of sound sources in both vertical and horizontal planes. Azimuth is the angular distance in the horizontal plane between the sound source and the listener's head (left or right, front or back). Elevation is the vertical angle between the sound source and the listener's head (up or down).

2.5.1.1 Binaural Cues

The most important cues for localizing a sound source on the horizontal plane are binaural cues, which occur because of the position of human ears on opposite sides of the head. When a sound source is not directly in front or behind the listener, the sound will be perceived by each ear at different times (due to distance) and with different intensities (due to the head acting as an obstacle). These cues are known as interaural temporal difference (ITD) and interaural level difference (ILD) respectively (Grantham, 1995).

Binaural abilities are not only important for single sound localization. In environments where multiple sounds are presented simultaneously, the use of two ears enables the selective attention to sounds coming from one particular direction while ignoring other sounds (Moore, 1989).

2.5.1.2 Pinna

The pinna is the visible part of the outer ear - the large shell-shaped lobe located on each side of the human head (Johnston, 2009). It performs a direction dependent filtering of sounds, which is important for both vertical and horizontal localization (Kuhn, 1987).

The pinna is especially important in discrimination of sound sources located in positions where binaural cues are not sufficient, for example, sources directly in front or behind the listener, where the ITD and ILD are negligible (Grantham, 1995). The pinna is also critical in determining the elevation of a sound source (Johnston, 2009).

2.5.1.3 Localization in the Vertical Plane

Human sound localization performance in the vertical plane is less reliable than in the horizontal plane (Colburn & Kulkarni, 2005), since ITD and ILD do not contribute in detecting the elevation of sound sources. Asymmetries in the pinna and head movements can disambiguate the direction of the sounds and minimize vertical localization errors (Warren, 2008).

2.5.1.4 Loudspeakers vs. Headphones

Using headphones to present spatially distributed sounds is possible due to head related transfer functions (HRTF) (Wightman & Kistler, 1989). A HRTF describes the transformation suffered by a sound signal from the time it leaves the source until it reaches the eardrums, for a given direction and environment. It takes into consideration the shape of the head, ears, torso, shoulders, and other characteristics of the environment that can affect the perceived sound (Colburn & Kulkarni, 2005).

For an HRTF to provide localization accuracy similar to that of loudspeakers, it needs to be personalized to each listener's ears, since the size and shape of the human pinna vary considerably from person to person (Yost, 2007). There exist generalized HRTFs that are calculated to an average head and ears (Begault, 1994), however these generalized functions decrease the localization accuracy in the vertical plane and increase front-back confusions (Wenzel, Arruda, Kistler, & Wightman, 1993).

The need for individualized HRTFs for optimal vertical localization makes the use of headphones in the presented research experiments impractical due to the time it would take to prepare each participant before their trials. Since this research focuses on a solution that uses parallel output for audio search, loudspeakers will be used in the experiment. For this solution to be successfully used with headphones in future applications, its implementation needs to include two key features:

- Individualized HRTFs need to be used to simulate the spatial locations of the sounds.
- The algorithm needs to incorporate head movements, by updating the spatial location of the delivered sounds in real time. This is necessary because people naturally move their heads when trying to localize a sound source. Head movements can influence the perceived location of sounds (Colburn & Kulkarni, 2005) and should not be ignored when headphones are in use.

2.5.2 Concurrent Presentation of Sounds

Humans have the ability to focus attention on one speaker in the middle of different simultaneous conversations and background noise. This phenomenon has been the subject of extensive research and is called the cocktail party effect (Cherry, 1953).

A user interface that makes use of this ability can play multiple sounds simultaneously, instead of the typical sequential playback, and increase the amount of information that can be presented to the user in a certain amount of time. Concurrent presentation also provides an effective way to make comparisons between audio data, as it reduces the need to remember a number of sounds.

It is essential that simultaneously presented sounds be of at least similar perceivable loudness to reduce the possibility of masking (Moore, 1989). When different degrees of loudness occur, sounds presented with higher intensity may obscure the other sounds, which will not be detected by the human ear.

In order to make the use of concurrent sounds more effective, spatial distribution of the sound sources is recommended. It has been suggested that concurrent sounds coming from different places are more easily discriminated than sounds that originate from the same spatial location (Bregman, 1990).

In addition, sounds that start at slightly different times tend to be more easily discriminated (Darwin & Ciocca, 1992). McGookin and Brewster (2004) tested this concept in an auditory user interface that presents earcons (sounds used in ways similar to visual icons) concurrently, by adding a 300ms onset-to-onset delay between the presentations of each earcon. They found that staggering the onsets of earcons by at least 300ms improves the identification of those earcons. The same effects are expected in the identification of concurrently presented music.

2.5.3 Simultaneous Sounds on User Interfaces

Several researchers have designed user interfaces that take advantage of the cocktail party effect to reduce the amount of time required to present information aurally. The idea is that users are able to focus on one of a few simultaneously presented audio streams and switch attention if anything interesting is overheard in the others.

The AudioStreamer, presented by Schmandt and Mullins (1995) was an interface that simultaneously played three audio recordings of news programs, spatially separated by a 60-degree angle in the horizontal plane. An interesting feature of their interface is that head motion sensors are used to determine the user's focus of attention and increase the volume for the attended channel. One of the issues with this approach is the high probability that the louder recording will completely mask the unattended recordings, forcing the sequential listening of each channel.

Sawhney and Schmandt’s Nomadic Radio was a wearable system that informed the user of upcoming appointments, incoming email messages, and news items (Sawhney & Schmandt, 2000). The audio was played around the user, using only the horizontal plane, and followed the layout of a clock, where the spatial position denoted the time of a scheduled appointment or the time of arrival of other messages. Users interacted with the device through speech.

Fernström and McNamara (2005) described a system for browsing of a music collection that supports listening to multiple songs simultaneously. Their system, the Sonic Browser, has a graphical interface that shows the recordings in a starfield display (Shneiderman, 1998). The user selects a circular area from the display and the songs located inside the circle are played back simultaneously. The Sonic Browser does not employ any similarity measure to organize the songs on the layout, neither has a hierarchical structure for browsing. It only uses differences in loudness between the left and right channels to assist users in differentiating and localizing the sounds. The authors did not perform a quantitative evaluation of their system with random users. Instead, their prototype was evaluated by a group of ten trained musicologists in a Thinking Aloud study that indicated good recognition of previously heard tunes when using the concurrent audio interface.

An auditory interface for hierarchical menu navigation for use while driving was proposed in (Sodnik, Dicke, Tomažič, & Billinghurst, 2008). The spoken words that represented the menu options were concurrently presented and spatially distributed around the user. This interface was compared to a visual interface and another auditory interface without concurrent presentation of sounds. There are six items in each menu level. All audio sources are on the horizontal plane located around the user's head. To select an item, the sounds need to be rotated until the elected item is played on the front speaker, directly in front of the user. Louder volume was used on the front speaker. An input device attached to the steering wheel consisted of a scrolling wheel used to rotate the menu options and two buttons used to confirm or cancel a selection. The authors found that the visual interface provided faster performance but increased the perceived workload and strongly distracted the driver.

2.5.4 Conclusion

The majority of work done in user interfaces with simultaneous sounds has a particular focus on the display of sound effects or spoken words (menus, news, voicemail messages), rather than music, and use merely the horizontal plane for distributing sounds, unlike the approach presented in this dissertation. In addition, most systems that present spatially distributed concurrent audio were not formally evaluated.

The design of new user interfaces that improve the audio search experience and allow users to explore audio collections is needed and has been encouraged by the research community (Casey et al., 2008; Lew et al., 2006).

This dissertation presents the design of an interface suitable for the search and browsing of audio information (music, sound effects, and speech) which uses a simultaneous, spatially distributed presentation of the audio data in an attempt to improve upon the bottleneck of sequential search. This interface was formally evaluated through the user experiments described in Chapter 4.

Chapter 3

Parallel Audio Player

3.1 Description

In order to explore the potential of search through multiple, simultaneous audio streams, it was necessary to develop the Parallel Audio Player. This audio player is a software application capable of playing six audio files simultaneously, through different audio channels. Each channel plays one file at a time, but up to six channels can be playing at the same time. The Parallel Audio Player was designed and developed by the author to be used in the experiments described in this dissertation.

3.2 Hardware

The Parallel Audio Player uses a remote control and infrared cameras as input devices, and spatially distributed loudspeakers to output the audio. The hardware configuration used to develop and test the application is described in this section. The same hardware configuration was used in the audio search user experiments described in the next chapter.

3.2.1 Loudspeakers and Infrared Cameras

Auditory stimuli is presented on six equal stereo loudspeakers, connected to two M-Audio Delta 1010LT sound cards, which were inserted in a Dell Dimension 8400 personal computer (PC) equipped with an Intel Pentium 4, 3GHz, 1GB RAM, and running the Windows 7 Enterprise Edition, 32-bit operating system.

The loudspeakers are mounted in three 7-foot-tall poles built by the author using PVC pipes and angle brackets. There is one loudspeaker on the bottom and another on the top of each pole (see Figure 2). A seventh loudspeaker is placed behind the participant for playing the target sample (which is a sample of the audio file participants are asked to find in the audio search user experiments). The same loudspeaker plays feedback sounds from the application, such as, an announcement of the stage of search in the beginning of each stage, and an applause sound at the end of a successful search.

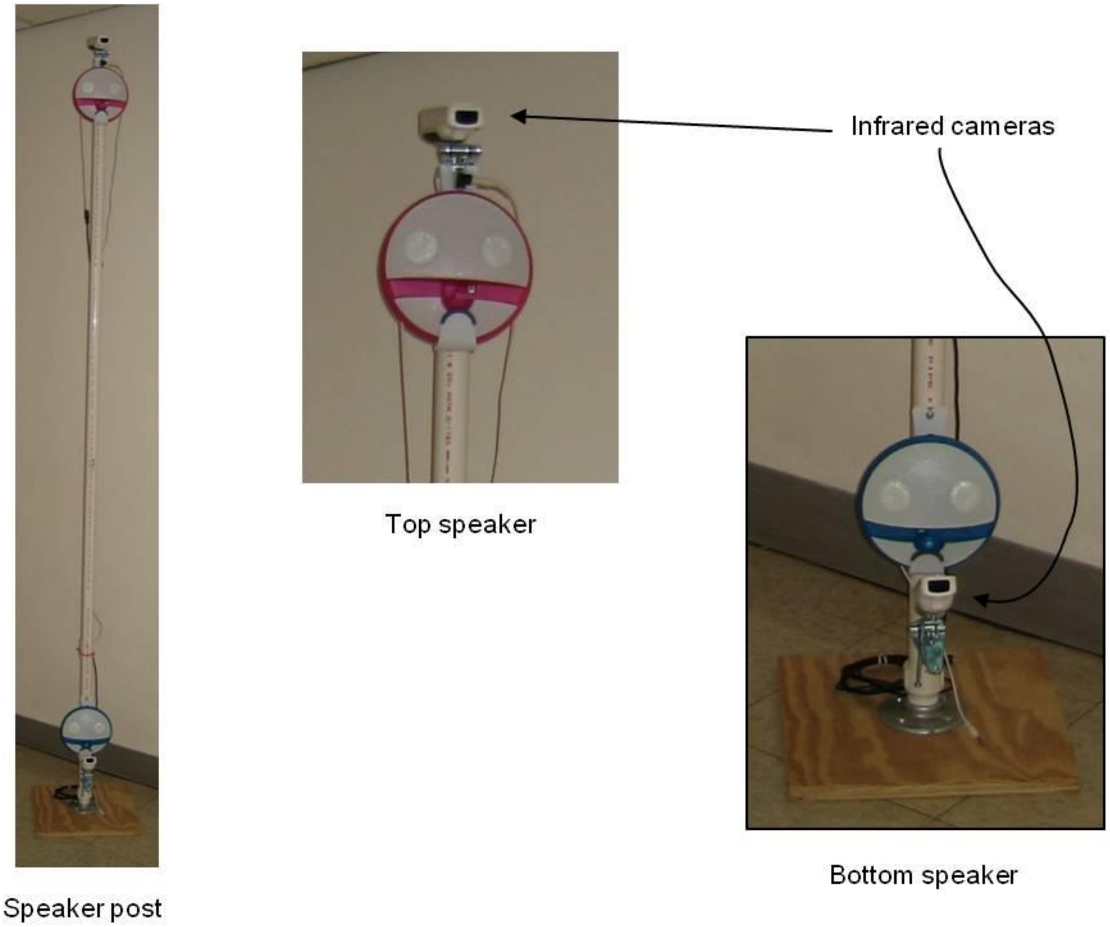

Figure 2 – Loudspeaker post with IR cameras

One infrared (IR) camera is mounted next to each loudspeaker. Each IR camera is part of a Wii remote control (wiimote), which is the controller for Nintendo’s Wii video game. The IR cameras are used to detect the loudspeaker to which the user is pointing when making selections.

3.2.2 Remote Control

The input device used for pointing and selecting loudspeakers is another wiimote, adapted by the author to emit infrared light through an IR led powered by one AA battery. Figure 3 shows the adapted remote control. The infrared light is detected by the infrared camera when the remote control is pointing to a specific loudspeaker and that information is transmitted to the PC wirelessly via Bluetooth. The remote control button presses are also transmitted to the PC via Bluetooth.

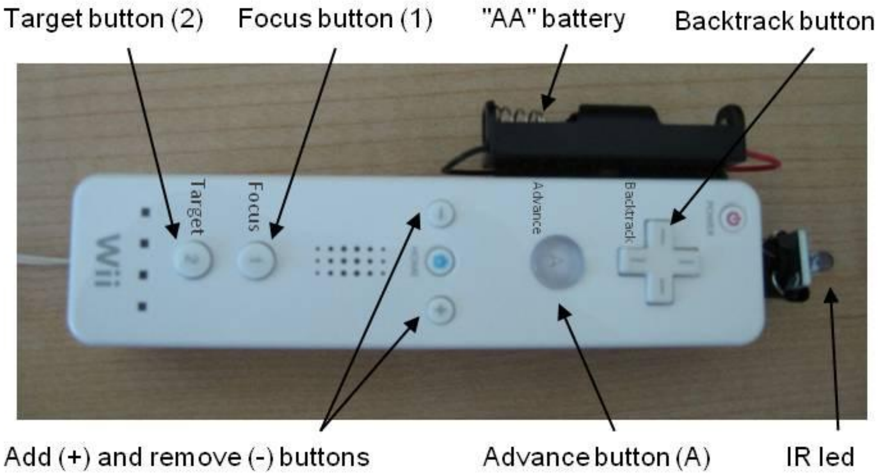

Figure 3 – User-operated remote control

The remote control was chosen as the input device since it is ideal for pointing at spatial locations and pressing buttons. The fact that users are familiar with the basic use of television remote controls promoted a positive transfer of learning (Perkins, 1994) to the wiimote, which was considered easy to use, even by users that had never played the Wii video games.

3.3 Software

The Parallel Audio Player consists of two modules: the audio module, responsible for outputting audio, and the wiimote module, responsible for receiving user commands from the input devices. The diagram in Figure 4 shows the relationship between the two modules and the hardware they support.

3.3.1 The Audio Module

The audio module is responsible for playing the audio files. It was programmed using Microsoft Visual C++ 2008 and the Microsoft DirectShow application programming interface (API), which is part of the Microsoft Windows Software Development Kit (SDK). DirectShow is an architecture for streaming media on the Microsoft Windows platform, based on the Component Object Model (COM). The output module uses COM objects provided by DirectShow to read and decode the audio files, and then pass the data to the sound cards.

This module can play a hierarchically organized collection of audio files. It is capable of simultaneously playing up to six audio files, through six different loudspeakers. There is always a short gap (300ms) between the starts of concurrently presented audio files to aid in sound discrimination (see section 2.5.2). The files start playing following a consistent sequence: top to bottom, left to right. Playback starts at the top left speaker, followed by the bottom left speaker, then the top speaker of the next speaker post to the right, and so on, until all speakers are playing simultaneously.

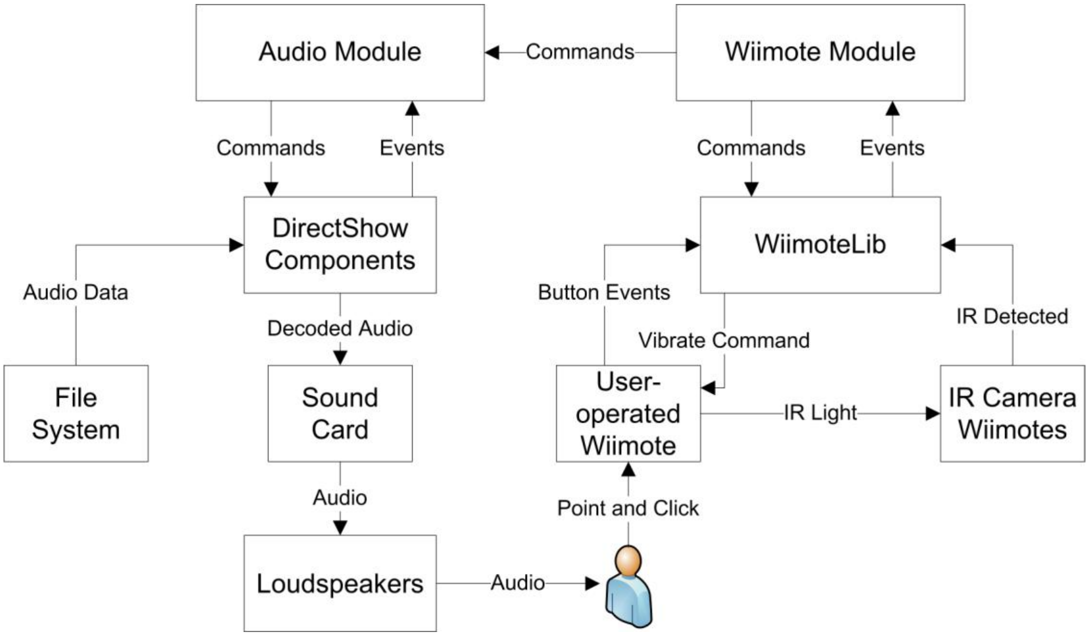

Figure 4 – Architectural components of the Parallel Audio Player

3.3.2 The Wiimote Module

The wiimote module is responsible for reading data from the wiimotes and writing data to them. This module was programmed using Microsoft Visual C# 2008 and the WiimoteLib API (Peek, 2011), a software library that facilitates the communication between the wiimotes and the application.

This module receives input commands (button presses) from the user-operated wiimote and IR information from the camera wiimotes. This information is translated into commands that are sent to the audio module. The wiimote module also sends vibration commands to the user-operated wiimote as feedback.

In order to satisfy the requirements of the audio search experiments conducted using the Parallel Audio Player, the wiimote module has a simple graphical user interface that is used by the experimenter only. In this interface, the experimenter can select the audio collection that will be browsed in the trial, the target sample, and the operating mode (see section 3.3.3). In addition, the experimenter has controls to start the trial and stop it, in the event a participant gives up.

The wiimote shown in Figure 3 is used to interact with the tool, by pointing to one of the six spatial positions and pressing buttons. The wiimote vibrates (as feedback) if the user presses a button without pointing at one of the valid spatial locations. The wiimote buttons perform the following actions on the selected spatial position:

- Advance ("A" button): selects the audio category that is playing in that spatial position and advances to the next level of the hierarchical classification – “play more sounds like this”. This button is also used to select the audio file that matches the target sample, when it is found.
- Remove ("-" button): silences the loudspeaker located in the selected spatial position.
- Add ("+" button): sounds the loudspeaker located in the selected spatial position (if previously silenced).
- Focus ("1" button - press and hold): plays only the audio file located in the selected spatial position. When released, the system returns to the previous state.

The following actions are position independent (the user does not need to point anywhere):

- Backtrack ("left" button): returns to the previous level of the hierarchy.
- Target sample ("2" button – press and hold): plays the target sample. When released, the system returns to the previous state.

The buttons on the wiimote are labeled, reducing the need for users to remember the mapping between the buttons and the functions they perform.

3.3.3 Operating Modes

The Parallel Audio Player has three operating modes:

- The *Serial* mode: allows playback of only one audio recording at a time. Audio does not play automatically in the beginning of each search stage. The room remains silent until users press and hold the Focus button to play an audio recording. That recording will play until the Focus button is released. The Add and Remove buttons are inactive in this mode.
- The *Parallel + Focus* mode: plays audio recordings simultaneously and users have the option of focusing on an individual recording temporarily. All recordings start playing automatically in the beginning of each search stage. When users press and hold the Focus button, all recordings stop playing, except for the focused one – the one playing at the loudspeaker where the user is pointing. When the Focus button is released, all recordings resume playing simultaneously. Users can focus on any recording, for as long as they want, and as many times as they want. The Add and Remove buttons are inactive in this mode.
- The *Parallel + Reduction* mode: plays audio recordings simultaneously and users have the option of stopping individual recordings while the others continue to play. All recordings start playing automatically in the beginning of each search stage. Any recording can be stopped and resumed at any time. Users press the Remove and Add buttons to stop and resume, respectively, the selected recording. The Focus button is inactive in this mode.

The state diagrams in Figures 5, 6, and 7 show the operation of all three modes for an audio collection that is organized in hierarchical levels, grouped by similarity. Each hierarchical level corresponds to a search stage. The target recording can only be found at the last search stage. A sample of the target recording (target sample) can be played at any time during the search, so users can be reminded of the audio recording they are trying to find.

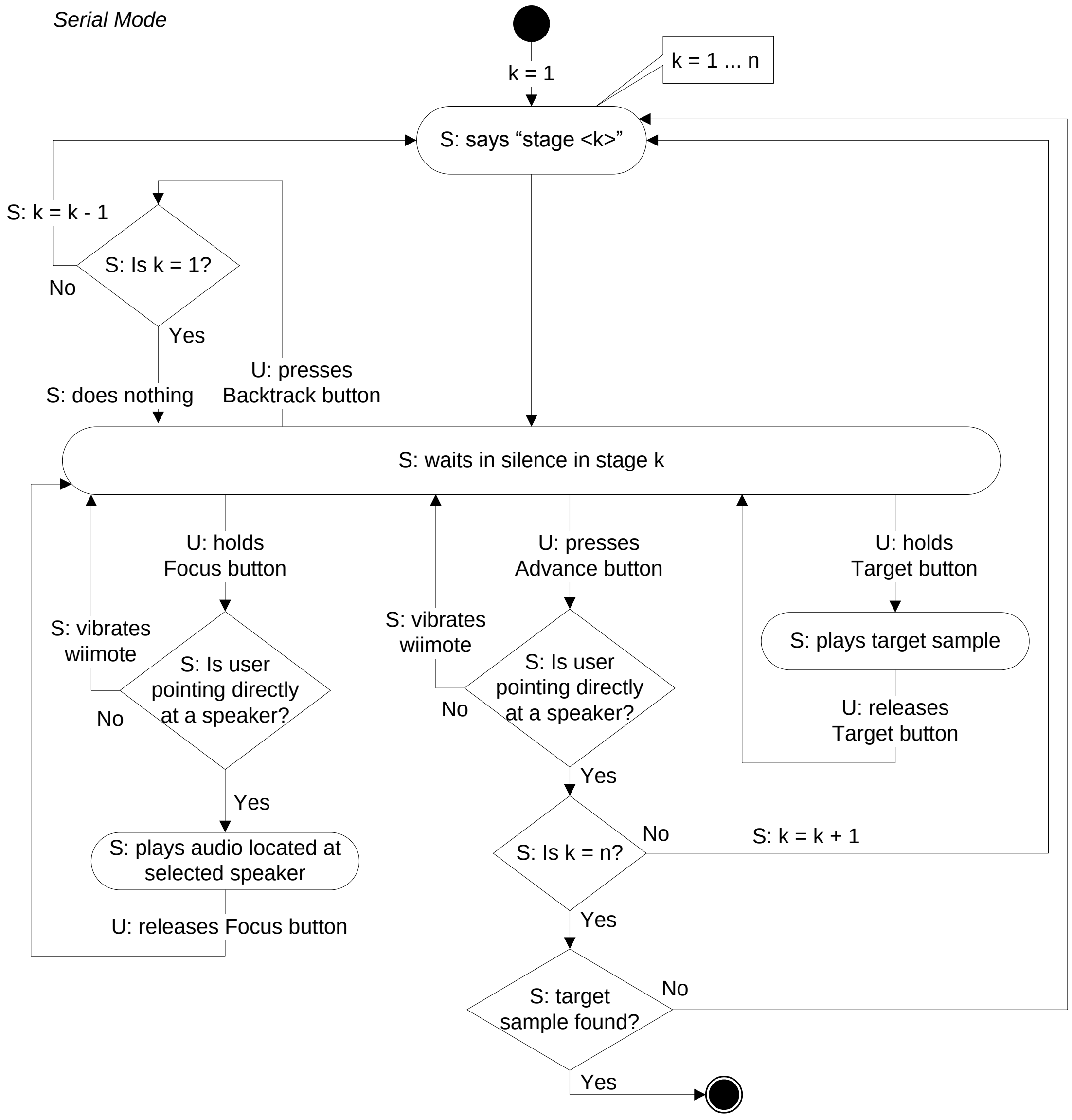

Figure 5 – State diagram showing the interaction between user (U) and system (S) in the *Serial* mode of the Parallel Audio Player

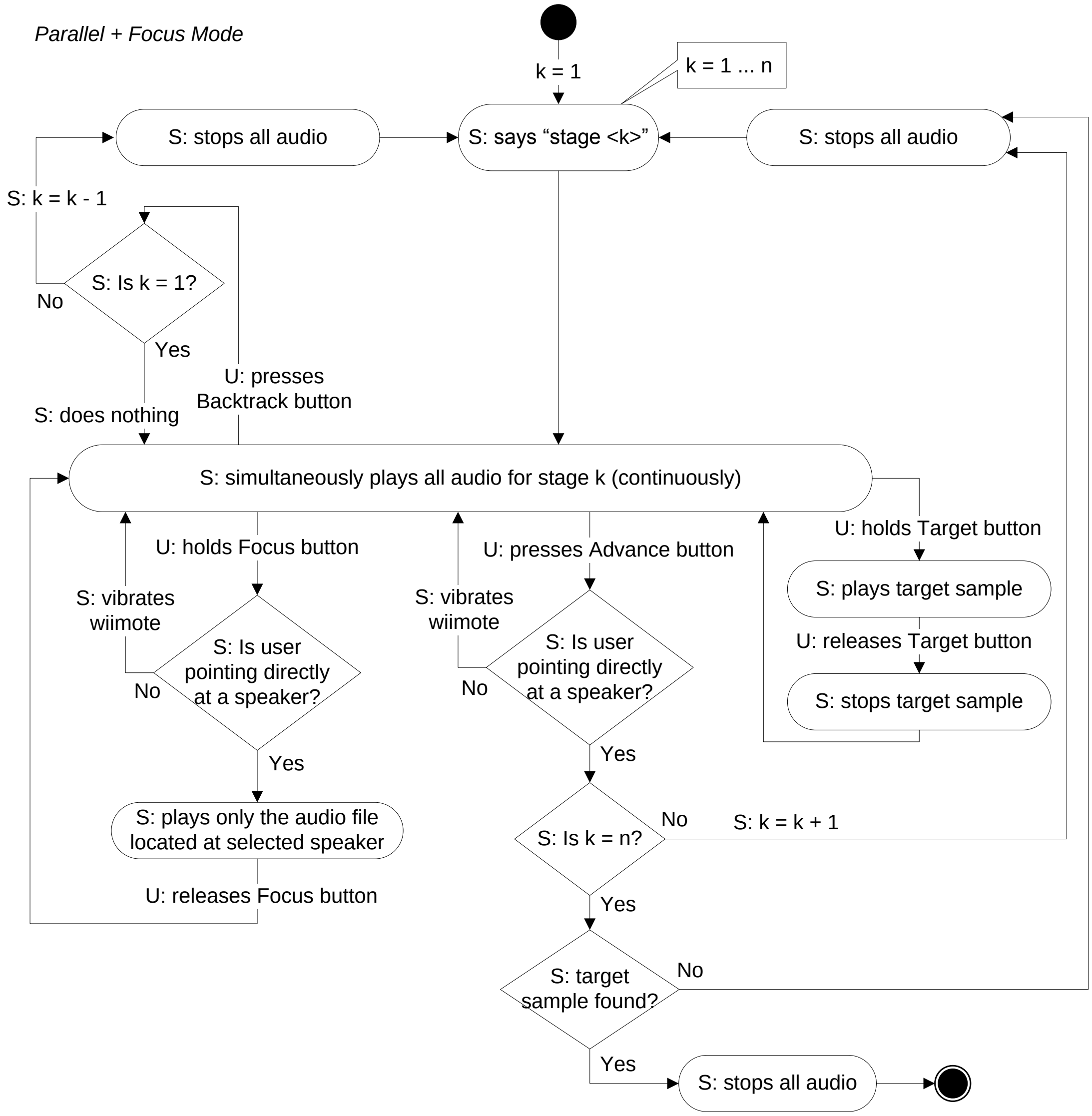

Figure 6 – State diagram showing the interaction between user (U) and system (S) in the *Parallel + Focus* mode of the Parallel Audio Player

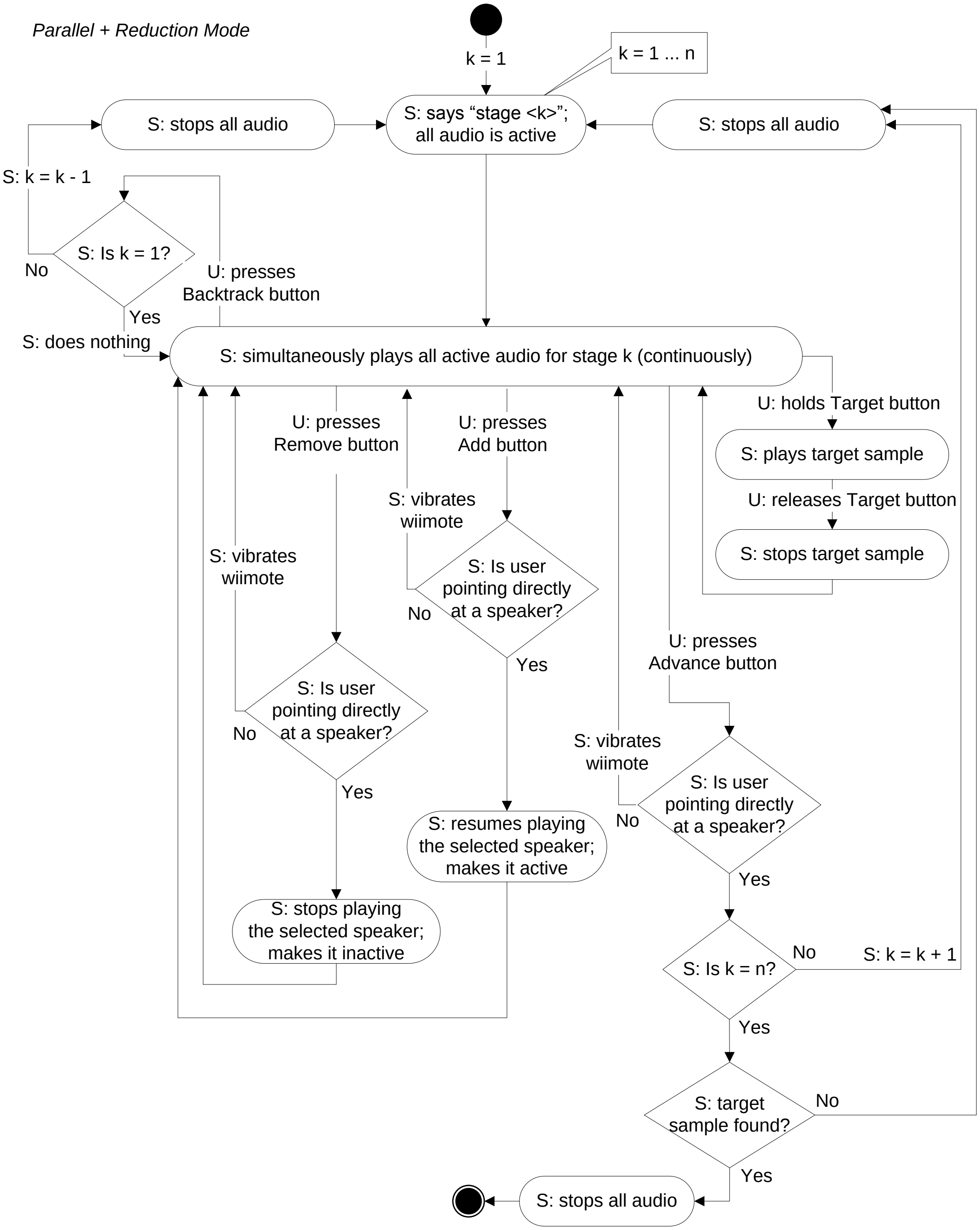
Parallel + Reduction Mode
k = 1
k = 1 ... n
S: stops all audio
S: says "stage <k>"; all audio is active
S: stops all audio
S: k = k - 1
S: Is k = 1?
No
U: presses Backtrack button
Yes
S: does nothing
S: simultaneously plays all active audio for stage k (continuously)
U: holds Target button
S: plays target sample
U: releases Target button
S: stops target sample
U: presses Remove button
U: presses Add button
S: vibrates wiimote
S: Is user pointing directly at a speaker?
No
Yes
S: vibrates wiimote
S: Is user pointing directly at a speaker?
No
Yes
U: presses Advance button
S: vibrates wiimote
S: Is user pointing directly at a speaker?
No
Yes
S: resumes playing the selected speaker; makes it active
S: stops playing the selected speaker; makes it inactive
S: Is k = n?
No
S: k = k + 1
Yes
S: target sample found?
No
Yes
S: stops all audio

Figure 7 – State diagram showing the interaction between user (U) and system (S) in the *Parallel + Reduction* mode of the Parallel Audio Player

Chapter 4

Experimental Methods and Design

4.1 Goal

This audio search study was designed to evaluate how the effectiveness of an auditory search task is affected when multiple complex sounds are presented simultaneously.

In each trial, participants listened to an audio sample – the target sample – for 30 seconds, and then browsed an audio collection, searching for that target sample.

Effectiveness was measured as speed and distance. Total task speed is defined as the time (minutes, seconds) necessary to complete the recognition task. Distance is measured as the number of steps taken in the path to the target sample. When a participant did not finish the task within the time limit, the task was recorded as incomplete.

4.2 Questions

The following questions were answered by this study:

- When browsing a hierarchically organized audio collection, if a number of audio recordings is played simultaneously rather than sequentially, then:
  - Does a search for a specific recording take less time?
  - Does a search for a specific recording require fewer steps?
  - Is the number of incomplete searches the same?

  - Is the number of mistakes matching an audio to its category (to move down the hierarchy) the same?
  - Is the *Parallel* + *Focus* mode of the Parallel Audio Player (the ability to select audio recordings to be briefly presented in isolation) more beneficial than the *Parallel* + *Reduction* mode (the ability to select audio recordings to be eliminated from the set of currently presented audio)?
- Do users prefer to browse an audio collection through a user interface that presents the audio simultaneously or sequentially?

In order to answer the above questions, two experiments were conducted. The experiments differ in the maximum number of audio recordings played simultaneously – either four or six. All the above questions were answered for each experiment.

The effect sizes from the first experiment were compared to those from the second experiment to detect any significant differences in performance that could help determine the best speaker configuration for a parallel audio player.

The following section describes the first experiment, which uses four simultaneous audio files in the parallel presentation conditions.

4.3 Experiment 1: Method

A repeated-measures design was used in this experiment. There were three experimental conditions: serial output (SO), parallel output + focus (POF), and parallel output + reduction (POR). The conditions differ in the way the audio is presented to the participants and in the browsing strategies available to them. As a counterbalancing technique, participants were randomly assigned to one of the six possible sequences of conditions presented in Table 1. The same number of participants was assigned to each sequence, which means the total number of participants had to be a multiple of six.

In all conditions, the participants used the Parallel Audio Player (described in Chapter 3) to search for a specific audio recording: the target sample. Each participant did three search tasks in each condition. Participants had their search time assessed as well as the number of incomplete tasks. Other observations were made as described in section 4.3.5.

Table 1 – Possible sequences of conditions

		Stage		
		1	2	3
Sequence	1	SO	POF	POR
	2	POF	POR	SO
	3	POR	SO	POF
	4	SO	POR	POF
	5	POR	POF	SO
	6	POF	SO	POR

4.3.1 Participants

Participants for this study were adults (18 to 65 years old) of any gender, race, ethnicity, or occupation, recruited from the University of South Florida through classroom announcements, e-mail announcements, and fliers posted on campus. Participation was voluntary and no identifiable information was recorded. Participants were identified only by a random identification number assigned at the beginning of their study session.

Participants responded to announcements by e-mail. An initial e-mail contact determined the participant's eligibility and an appointment was set up for conducting the study. To ensure that they were capable of performing the study related tasks, potential participants were asked if they had normal hearing. Additionally, participation was limited to those with normal use of at least one arm and hand, since participants were required to use a remote control (point and press buttons) with average speed and accuracy.

The first experiment had 36 participants. The absence of preliminary data precluded a reliable estimate of effect size for performing a power analysis. Therefore, a moderate effect size of 0.50 was hypothesized, following Cohen’s (1988) recommendation.

Assuming a moderate effect size, a moderate correlation of 0.60 among the repeated measures, and a medium dispersion of group means, a sample size analysis (Bausell & Li, 2002) indicated that at least 36 participants would be needed in order to produce an 80% or higher chance of obtaining statistical significance ($p < .05$).

4.3.2 Instruments

4.3.2.1 Parallel Audio Player

The hardware and software described in Chapter 3 were used in the experiment. The maximum number of audio files presented simultaneously was set to four; consequently, only five loudspeakers were used: four to play the audio collection and one to play the target sample and feedback sounds.

4.3.2.2 Survey

After participants finished all search tasks, they were asked to complete a user experience survey consisting of five questions. In the first three questions, participants rated the difficulty in using each of the three modes of the Parallel Audio Browser in a five-point scale (with values of 1-very difficult, 2-difficult, 3-neutral, 4-easy, 5-very easy). The forth question asked which mode they preferred to use for searching audio. The fifth question measured their perception of search speed, by asking which mode allowed them to find the target sample faster.

4.3.3 Materials

4.3.3.1 Musical Genre Taxonomy

An audio taxonomy is a hierarchical tree consisting of audio categories. For this experiment, musical pieces were used because they are highly complex sounds, made of many other sounds (instruments and voices). If users can search more effectively for music when the recordings are being presented simultaneously, it is highly likely that a search for other types of audio, such as sound effects or speech, will also benefit from parallel presentation.

The participants of this study navigated through a collection of music, categorized by genre. The organization of the music in this study is adapted from the taxonomy designed by Pachet and Cazali (2000), and the classifications used by online music retailers such as Amazon (Amazon.com Inc., 2011) and emusic (eMusic.com Inc., 2011). Pachet and Cazali analyzed several genre classifications used by the music industry and noticed many inconsistencies between them. In addition, the classifications were mostly used to describe albums or artists, as opposed to individual music titles. They designed a new taxonomy to classify individual titles and to include similarity relations between genres. However, no existing taxonomy was balanced enough to fit this experiment's needs. An unbalanced taxonomy could add an unnecessary increase in variance between the conditions.

The music trees used in this experiment are subsets of a larger taxonomy. There are three levels, with four genres in each level, totalizing 4 primary genres, 16 sub-genres, and 64 leaf songs per music tree. In order to reduce practice effects, three different music trees were created, so that participants performed only three searches per tree. The music classification for one of the music trees can be seen in Figure 8.

Note that any hierarchical classification could have been used, for example, a chronological taxonomy, or an organization by artist then album. The categorization by genre was chosen because it does not require specific musical knowledge. This was especially important since the music used in the experiment was unfamiliar to the participants. Participants did not need to know the genre of a specific song, but they could recognize that two songs sounded similar enough to each other to belong to the same genre.

4.3.3.2 Audio Files

The audio files used in this experiment are musical pieces, retrieved from the author's personal music collection, and websites that offer free music downloads from new artists, such as emusic (eMusic.com Inc., 2011). All the music files are in the mp3 (formally MPEG-1 Layer 3) format. Only 30-second snippets of the music files were used. An effort was made to select samples that are good representatives of each music style in the taxonomy, but that are not very well known to the public. This created a more balanced experiment as we expected the participants not to recognize most of the songs. Participants had to identify the music style by listening to it instead of knowing that a certain artist belongs to the same style as another artist. An unfamiliar target sample was played to a participant that tried to recognize it in the music collection by selecting similar sounding music to move down the taxonomy.

In order to level the volume of the samples, an open source software called MP3Gain (MP3Gain Development Team, 2009) was used. This software analyses MP3 files and determines how loud they sound to the human ear. This information is used to adjust the files so that they have similar perceivable volume.

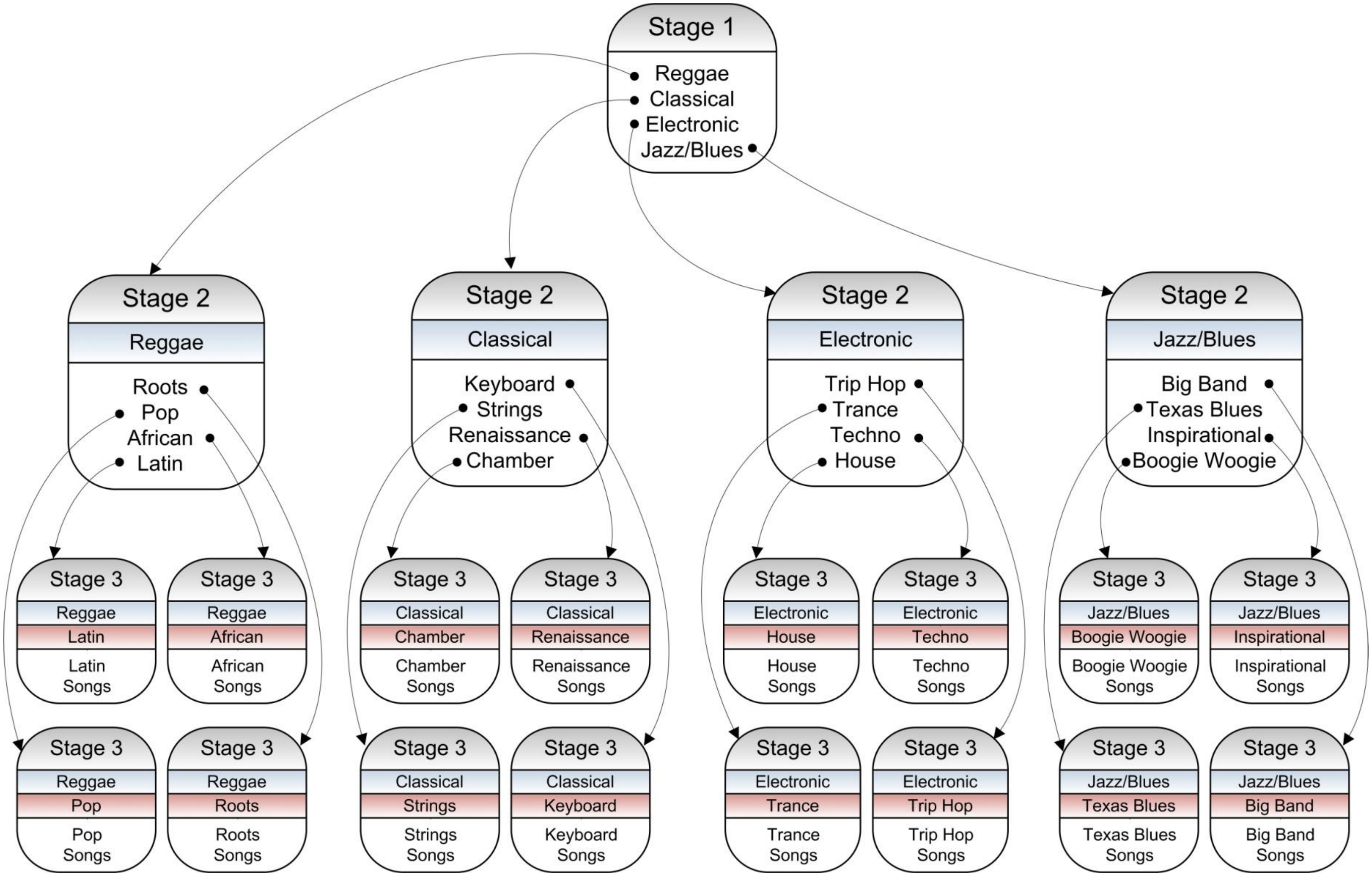

Figure 8 – Musical genre taxonomy used in experiment 1

4.3.4 Procedure

Pilot trials were conducted to test the experiment's procedure, the equipment, software, instructions, and to determine how much time would be needed from each participant. It was calculated that in one hour each participant would be able to complete nine search tasks, after receiving instructions and practice trials, and answer the survey. A time limit per search task was set to four minutes.

4.3.4.1 Potential Risks and Benefits to Participants

The research presented no risk of harm greater than that encountered in the participants' daily lives. There were no direct benefits to the participants, but this study contributes to a better understanding of audio search strategies and aid in the design of better human-computer interfaces for audio search and browsing.

4.3.4.2 Pre-Manipulation

The participants had the opportunity to talk to the principal investigator on the phone or via e-mail prior to setting an appointment for participation. At that time, they were given a brief overview of the study procedure and the opportunity to ask questions. All qualified participants were scheduled for an individual appointment, held in room ENB 313 at the Engineering Building at the University of South Florida. Upon arrival, each participant was required to read a Human Research Informed Consent Form, receive instructions for the experiment, and verbally agree to participate.

Participants received a gift certificate for a free lunch at a local pizza restaurant, as compensation for their time. They were made aware that they were allowed to stop the experiment and leave at any time without any penalty.

Practice trials were conducted to enable participants to become familiar with the experimental apparatus and procedures.

4.3.4.3 Manipulation

The loudspeakers were positioned 5 feet away from the participant's chair. For all search tasks, the participant was sitting on the chair, listening to the music and making selections by pointing the remote control at a loudspeaker and pressing the remote control's buttons (one at a time).

The experiments were conducted by the author (the experimenter), who was sitting behind the participant to avoid providing any clues as to the goodness of choices. The room layout is shown in Figure 9.

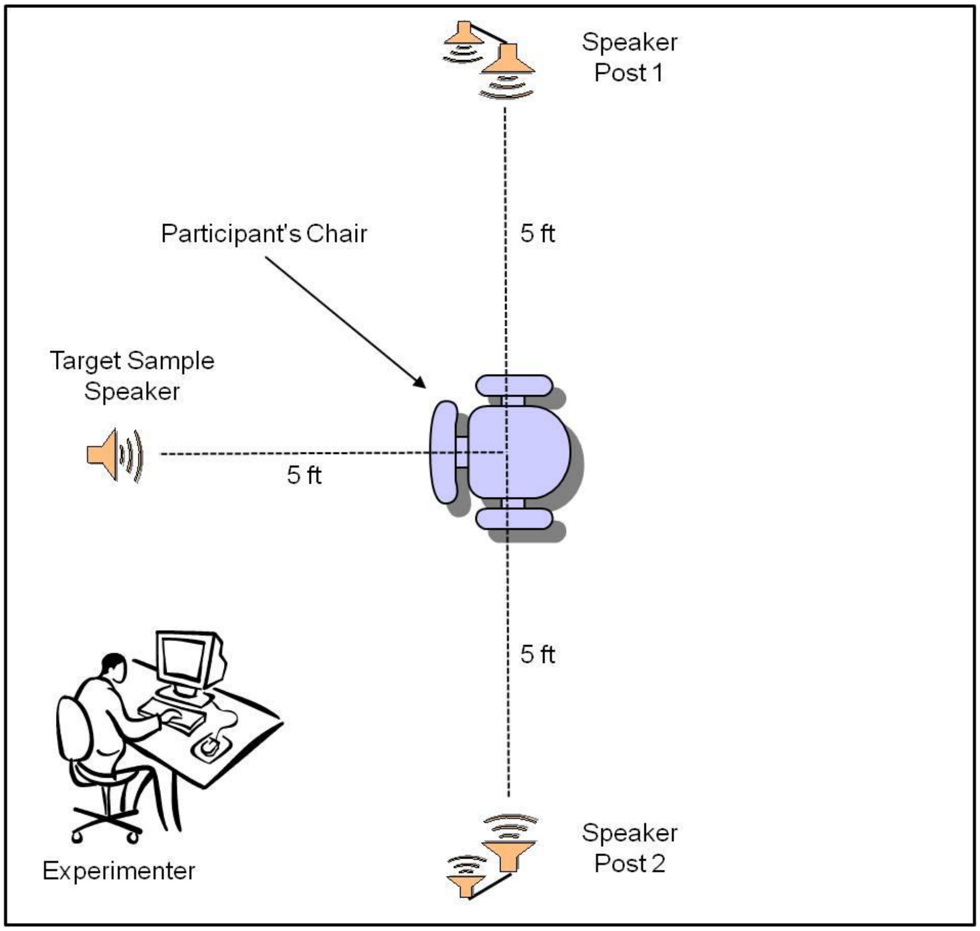

Figure 9 – Layout of the experiment room during the first experiment

Each participant was exposed to all three experimental conditions, in one of six possible orders of trials, as shown in Table 1. There were three search tasks per condition, totalizing nine search tasks per participant.

In the SO condition trials, participants used the Parallel Audio Player in the *Serial* mode. In the POF trials, the audio player was used in the *Parallel* + *Focus* mode, and in the POR trials, the *Parallel* + *Reduction* mode was used. The three modes were explained in section 3.3.3.

The procedure for all conditions is as follows:

The experimenter starts the trial. The target sample plays for 30 seconds and then stops. No information about the target sample (title, artist, genre, etc.) is given. Participants are asked to find that sample in the audio collection, which is organized hierarchically, as presented in section 4.3.3.1. The search is divided in stages that correspond to the taxonomy levels. In each level, each category is identified by an audio sample representative of that category. Participants move through taxonomy levels (search stages) by listening to the samples that represent the categories in each stage, and selecting the one that sounds the most similar to the target sample. The process repeats until one of the following happens: (a) the participant recognizes the target sample; (b) the participant decides to give up searching for that target sample; or (c) the maximum task time is reached and the experimenter stops the trial. The target sample can only be found in the last taxonomy level and, as in a perfect n-ary tree, there is only one path that leads to it (see Figure 10).

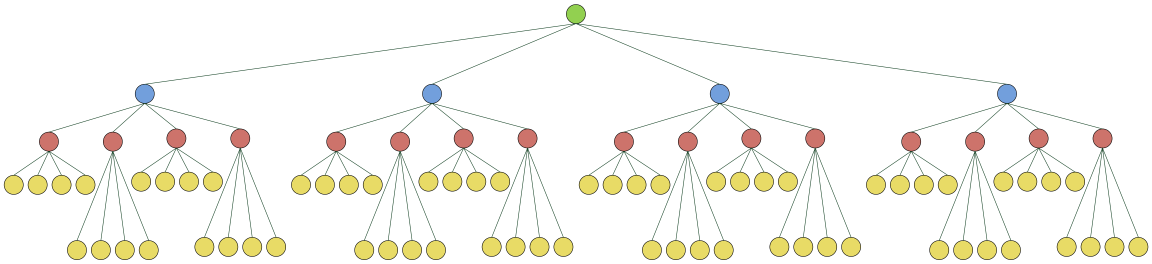

Figure 10 – Quaternary (4-ary) tree structure. Any of the leaves (yellow nodes) can be the target sample.

The participant is asked to select the target sample once it has been identified by pressing the "A" button on the remote. In the unlikely event that the sample is incorrectly identified, the search task continues.

When the target sample is correctly found, the trial stops, the software records the total task time along with the other performance measures listed on section 4.3.5, and the next search task begins (until the last trial for that participant). If the participant gives up or the time limit is reached, that task is recorded as "incomplete".

While browsing, participants are allowed to backtrack – go up a level in the taxonomy – as many times as needed, which is useful if they select the incorrect audio category. They are also allowed to replay the target sample at any time.

4.3.4.4 Post-Manipulation

Upon completion of the search tasks, the participant completes a questionnaire (described in section 4.3.2.2) to assess their user experience with the Parallel Audio Player.

4.3.5 Performance Measures

The following data was recorded by the computer software, for each search task:

- Task completion time: number of minutes and seconds from the beginning of the search until the target sample is correctly identified. The total task time was recorded, as well as the time spent in each search stage. When the search task could not be completed within the time limit, it was recorded as an incomplete task.
- Number of total steps: The number of steps for each completed task was recorded, as well as the number of steps in each search stage.
    - SO condition: A step is recorded every time the participant holds the Focus button to play an audio file. If the same file is played multiple times, multiple steps are recorded.
    - POF condition: A step is recorded every time the participant holds the Focus button to focus on an audio file. If the same file is being focused multiple times, multiple steps are recorded.
    - POR condition: A step is recorded every time the participant presses the Remove button to mute an audio file. If a removed file is later added, and then removed again, another step is recorded.
- Time spent listening to the target sample, per search stage: The number of seconds participants spent holding the Target button to listen to the target sample during each search stage.
- Time before the first step, per search stage: The time was recorded from the beginning of a search stage until the first step taken in that stage. The first step was taken when participants played (SO), focused (POF), or removed (POR) an

audio recording, backtracked, or selected a recording to advance to the next stage.

- Number of backtracks: Number of times a participant uses the Backtrack button to return to a previous search stage.

The above measures were recorded per search stage so that the differences between stages could be analyzed. For each search task, participants were considered to be in one of the search stages only when they were in the correct path to the target sample. For instance, in the music trees used in this experiment, there were three search stages in the correct path towards the target sample. If participants made incorrect selections that led to incorrect sub-trees (outside of the correct path), the time and steps spent in the incorrect sub-tree are recorded as "lost time" and "lost steps".

Figure 11 illustrates an example where a participant made one incorrect selection before finding the correct path to a target sample. The correct path was Electronic → Trance → Song “Aurora Borealis”:

1. In stage 1, the participant browsed for 19.8 seconds before correctly selecting the electronic song.
2. In stage 2, she incorrectly chose the sub-genre *House* after browsing for 23.5 seconds. She went to level 3 of the tree, but in the incorrect sub-tree, so she is not considered to be in stage 3 but “out-of-path”.
3. Since she was outside of the correct path to the target sample, the time spent browsing the songs belonging to the *House* sub-genre (18.7 seconds) was counted as "lost time". After realizing the target sample was not in that subset, the participant backtracked to stage 2.
4. In stage 2, she browsed for another 11 seconds, and correctly selected *Trance*.
5. In stage 3, she spent 15.3 seconds before selecting the target sample.

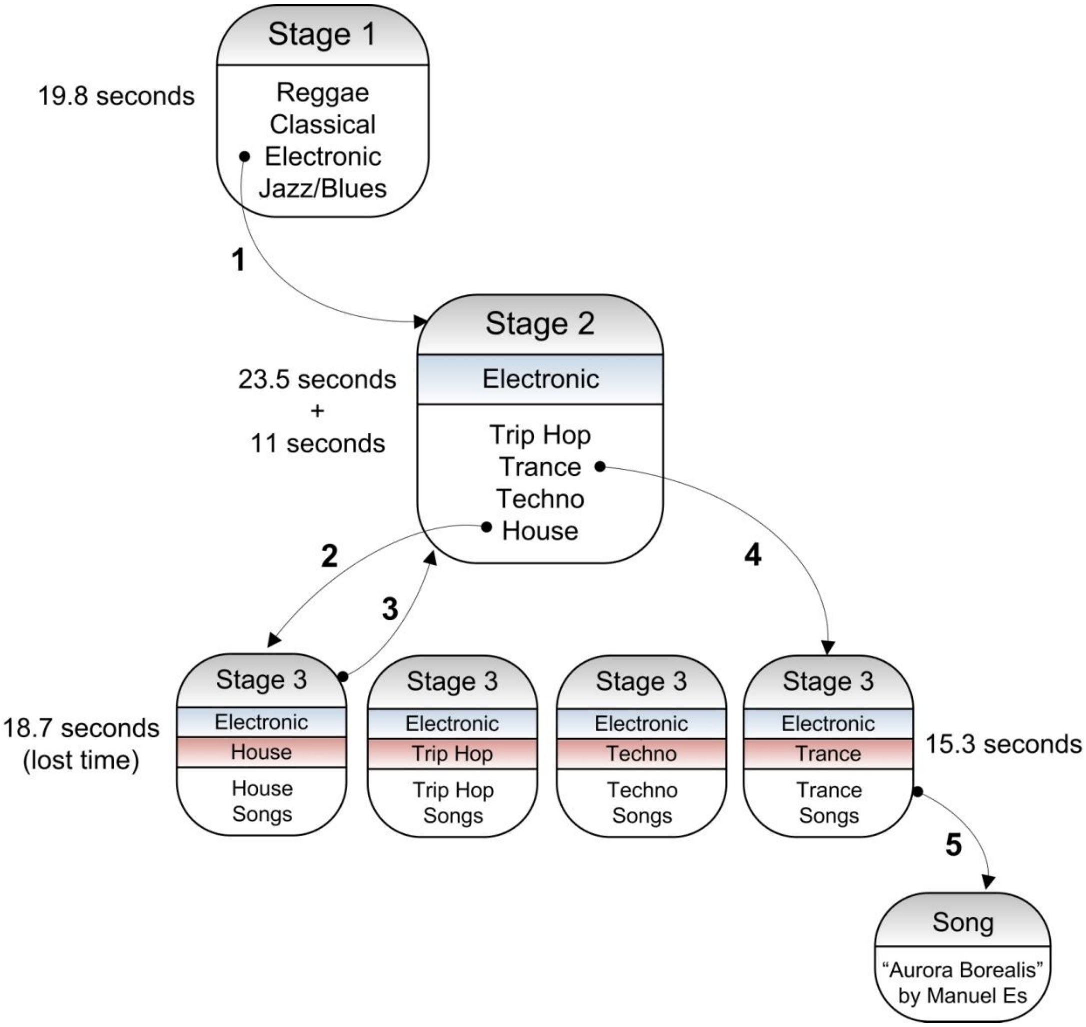

Figure 11 – Example of how the duration of the search stages was calculated

The times per stage in this example were computed as follows: stage 1: 19.8 seconds; stage 2: 23.5 + 11 = 34.5 seconds; stage 3: 15.3 seconds; lost time: 18.7 seconds; total search time: 88.3 seconds.

4.4 Experiment 2: Method

Since participants in the first experiment were able to use the Parallel Audio Player successfully, with four files being played simultaneously in each search stage, a second experiment was conducted, with six audio files being presented at the same time. The second experiment had a new group of participants, and followed the same procedures for data collection and analysis as the first experiment.

A repeated-measures design was used, with the same experimental conditions as experiment 1: serial output (SO), parallel output + focus (POF), and parallel output + reduction (POR). Participants were randomly assigned to one of the six possible sequences of conditions presented in Table 1.

4.4.1 Participants

The second experiment had 18 participants, recruited from the University of South Florida the same way as the participants from the first experiment.

4.4.2 Instruments

4.4.2.1 Parallel Audio Player

The hardware and software described in Chapter 3 were used in the experiment. The maximum number of audio files presented simultaneously was set to six; consequently, seven loudspeakers were used: six to play the audio collection and one to play the target sample and feedback sounds (Figure 12).

4.4.2.2 Survey

After participants finished all search tasks, they were asked to complete a user experience survey consisting of the same five questions as the survey for experiment 1 (section 4.3.2.2).

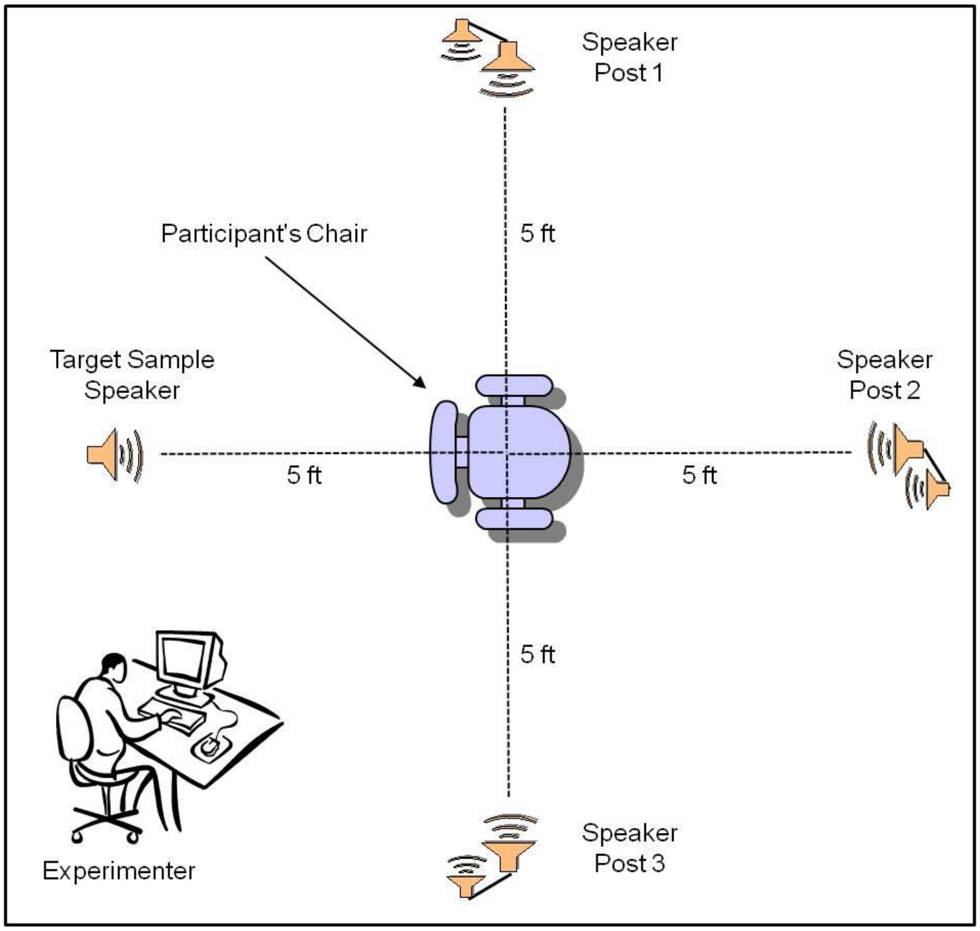

Figure 12 – Layout of the experiment room during the second experiment

4.4.3 Materials

4.4.3.1 Audio Files

The audio files used in this experiment are 30-second snippets of musical pieces in mp3 format. As in the previous experiment, the selected samples were expected to be unfamiliar to the participants.

4.4.3.2 Musical Genre Taxonomy

The music trees used in this experiment were derived from the same taxonomy as the trees from the first experiment. However, this experiment had six songs per search stage. Consequently, each music tree is a perfect 6-ary tree, with three levels and six genres in each level, totalizing 6 primary genres, 36 sub-genres, and 216 leaf songs. In order to reduce practice effects, three different music trees were used, so that participants performed only three searches per tree.

4.4.4 Procedure

A time limit per search task was set to five minutes, one additional minute than in the first experiment, since there were two extra audio recordings to browse in each search stage. Nevertheless, it was calculated that each participant would be able to complete nine search tasks, after receiving instructions and practice trials, and answer the survey, in less than one hour.

The experimental procedure was the same as in experiment 1 (section 4.3.4), except for the room layout (Figure 12), which was slightly different, with an extra loudspeaker post, placed directly in front of the participant's chair.

4.4.5 Performance Measures

Experiment 2 has the same performance measures as experiment 1 (section 4.3.5).

Chapter 5

Results

5.1 Background Information

Both the first and second experiments compare the effort required to find a target audio sample in three different audio output conditions. In the *Serial Output* condition participants listened to one audio recording at a time while in the two parallel conditions, they listened to multiple recordings at the same time, with the option to either listen momentarily to any individual recording in isolation (*Parallel Output* + *Focus* condition) or stop one or more recordings from playing (*Parallel Output* + *Reduction* condition). The maximum number of audio recordings played simultaneously was four in the first experiment and six in the second. The effort to find the target sample is represented by the two dependent variables: *search time* and *number of steps*.

Nine trials were administered to each participant, three for each audio output condition. At the end of this data collection procedure, the mean scores of each condition were obtained, so that each participant contributed one score per condition for each performance measure. For participants that had an incomplete search in any condition, the mean of the other two scores for that condition was used. Participants did not have more than one incomplete search per type of audio output.

For each experiment, a repeated-measures MANOVA was used to detect any effect of the type of audio output (serial vs. parallel) on the audio search effort. Significant MANOVAs were followed by separate repeated-measures ANOVAs on each of the dependent variables.

Additional effects, such as the duration of the search stages, the time before taking the first step, and the time spent listening to the target sample in each type of audio output, were also analyzed using repeated-measures ANOVAs.

All statistical tests were conducted using α = .05.

5.1.1 Effect Sizes

Effect sizes for the significant repeated-measures ANOVAs were calculated using the following equation for omega squared (Field, 2009):

$$\omega^2 = \frac{\left[\frac{k-1}{nk}(MS_M - MS_R)\right]}{MS_R + \frac{MS_B - MS_R}{k} + \left[\frac{k-1}{nk}(MS_M - MS_R)\right]}, \tag{5.1}$$

where *k* is the number of conditions in the experiment, *n* is the number of participants, MS_M is the mean square for the model, MS_R is the residual mean square, and MS_B is the between-participant mean square. For interpreting the resulting ω^2, values of .01, .06, and .14 represent small, medium, and large effects respectively.

When the results of the main repeated-measures ANOVAs were significant, the Pearson's correlation coefficient was used as a measure of effect magnitude for the planned contrasts and was derived from each contrast's F-ratio and residual degrees of freedom (df_R) according to the following equation:

$$r = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}} \tag{5.2}$$

When paired samples t-tests were used for post-hoc pairwise comparisons, the equation used to calculate the effect magnitude was (Rosenthal, 1991):

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{5.3}$$

For interpreting the magnitude of *r*, the values of .10, .30, and .50 correspond to small, medium, and large effects respectively, following Cohen's (1988) guidelines.

5.1.2 Assumption of Normality

Even though the Shapiro-Wilk test suggested that the data violated the assumption of normality, the results of parametric tests are reported in this chapter with the belief that the *F* statistic is robust to deviations from normality (Lindman, 1974). However, to verify the reliability of the results, the data was normalized using a log transformation, and another analysis was performed for the main effects of this study, with similar results obtained. This analysis is described on Appendix A.

5.1.3 Order and Sequence Effects

The experiment was designed to minimize practice effects by allowing participants to get familiar with the browser's user interface in a few practice trials. Order and sequence effects were controlled by counterbalancing the order of conditions (see section 4.3). These effects were expected to balance out across the different orders. An analysis was conducted to detect if the performance changed as the experiment progressed (indicating the presence of practice and/or fatigue effects) and to detect if there was any sequence of conditions that was significantly different from others (indicating the presence of sequence effects). No significant practice, fatigue, or sequence effects were found in either experiment. This analysis is described in Appendix B.

5.2 Experiment 1: Main Effects

Data from all 36 participants were used in this analysis. The results of the overall MANOVA revealed a significant effect of audio output on the effort required to find the target sample, $V = 0.58$, $F(4, 140) = 14.20$, $p < .05$, using Pillai's Trace. The results of univariate ANOVAs were then examined and are presented in the following sections.

5.2.1 Effect of Audio Output on Search Time

On average, participants were able to find the target sample faster when the audio was available simultaneously, regardless of the type of parallel output (*Focus* or *Reduction*), as shown by the chart in Figure 13.

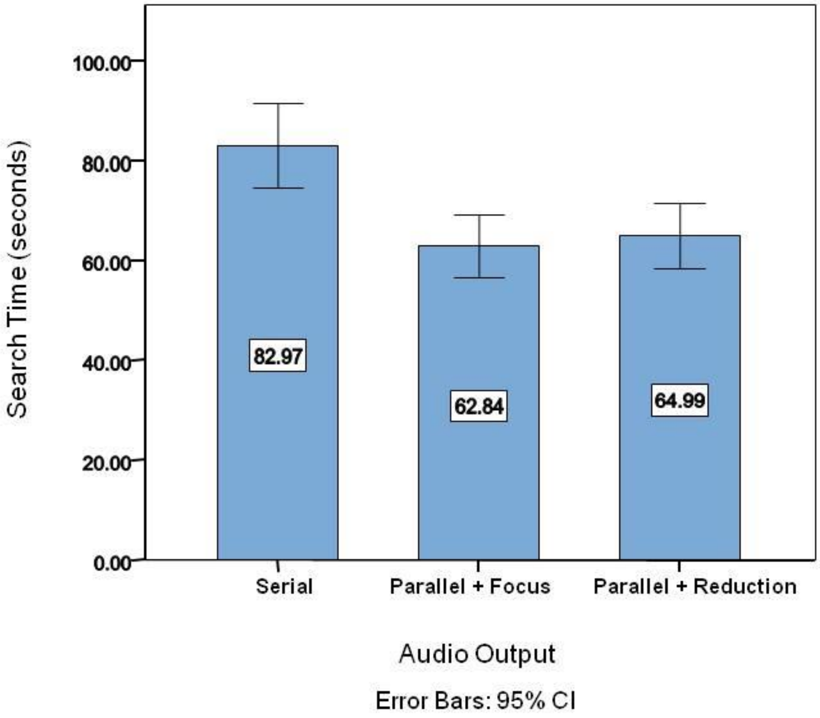

Figure 13 – Error bar chart of the mean duration of audio searches (lower is better) in three different forms of audio output (experiment 1)

A one-way repeated-measures ANOVA was carried out to confirm this outcome. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 5.56$, $p > .05$, therefore no correction was needed.

The time to find the target sample was significantly affected by the type of audio output during the search, $F(2, 70) = 6.59$, $p < .05$, $\omega^2 = .06$. Planned contrasts were used to follow-up this finding and showed that the search time was significantly higher when the audio output was serial as opposed to parallel, *F*(1, 35) = 9.43, *p* < .05, *r* = .46. However, there was no significant difference between the two parallel conditions, *F*(1, 35) = 0.20, *p* > .05.

5.2.2 Effect of Audio Output on Number of Steps

Figure 14 shows that the average number of steps required to find the target sample with the serial audio output was much larger than with either parallel forms of audio output.

A one-way repeated-measures ANOVA was conducted on these data. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 8.13$, *p* < .05, therefore the degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity (ε = .86).

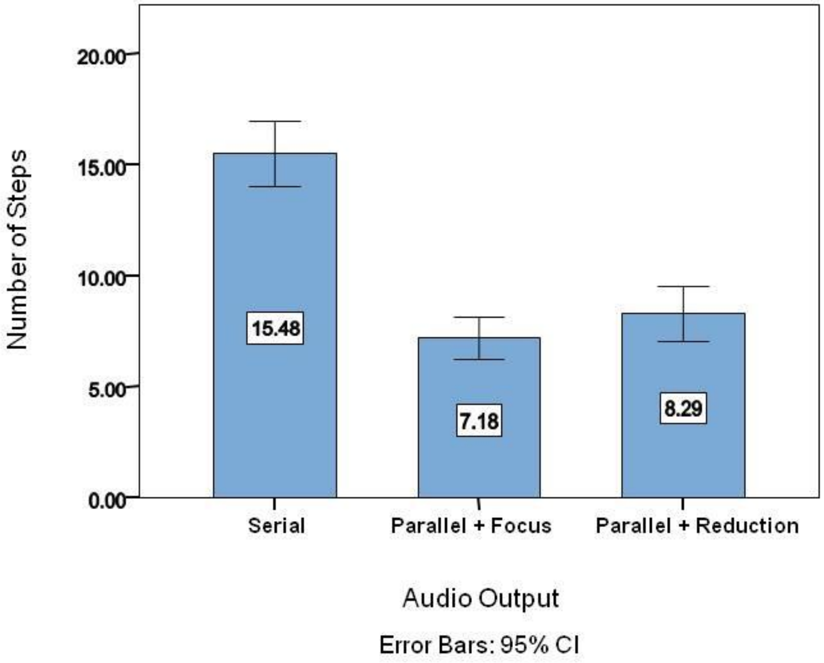

Figure 14 – Error bar chart of the mean number of steps taken during audio searches (lower is better) for three audio output conditions (experiment 1)

The analysis revealed a significant effect of the form of audio output on the number of steps taken to find the target sample, $F(1.72, 60.21) = 35.82$, $p < .05$, $\omega^2 = .35$. Contrasts confirmed that the number of steps was significantly higher in the serial search condition than in the parallel conditions, $F(1, 35) = 50.44$, $p < .05$, $r = .77$, but the two types of search in parallel output were not significantly different from each other, $F(1,35) = 1.81$, $p > .05$.

5.3 Experiment 1: Additional Effects

5.3.1 Background Information

Each search task performed in the experiment was composed of three stages. Each stage corresponds to listening to the songs that belong to a level of the music tree described in section 4.3.3.1 and selecting one of these songs in order to move to the next stage.

The following analysis compares the three search stages in the correct path towards the target sample. If participants made incorrect selections that led to incorrect sub-trees (outside of the correct path), the time and steps spent in the incorrect sub-tree are recorded as "lost time" and "lost steps" (section 4.3.5).

5.3.2 Duration of Search Stages

It was expected that stage 2 would be the most difficult and consequently the longest. In order to compare the duration of each stage overall and in each audio output condition, a two-way repeated measures ANOVA was performed.

Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of search stage, $\chi^2(2) = 7.70$, $p < .05$, therefore the degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity ($\varepsilon = .87$). Sphericity was not violated for the interaction effect between the type of audio output and the search stage, $\chi^2(9) = 14.87$, $p > .05$.

The mean duration of each search stage, regardless of audio output, was as follows: 21.90 seconds for the first stage; 26.03 seconds for the second stage; and 16.25 seconds for the third stage. There was a significant difference in the duration of the search stages when the type of audio output was ignored, $F(1.74, 60.76) = 23.07$, $p < .05$, $\omega^2 = .17$.

Repeated-measures t-tests (using a Bonferroni adjustment, $\alpha = .05/3 = .017$) showed that the three stages were significantly different from each other. The second stage was the longest and the third stage the shortest (stage 1 versus stage 2: $t(35) = -2.53$, $p < .017$, $r = .39$; stage 1 versus stage 3: $t(35) = 5.31$, $p < .017$, $r = .67$; stage 2 versus stage 3: $t(35) = 5.92$, $p < .017$, $r = .71$).

The interaction between the type of audio output and the search stages was not statistically significant, $F(4, 140) = 1.95$, $p > .05$. The graph in Figure 15 shows this interaction and reveals that the most drastic differences in stage duration occurred in the *Serial* output condition and the smallest differences in the *Parallel Output + Reduction* condition. All audio output conditions produced a similar pattern, where stage 2 is the longest and stage 3 is the shortest, but in the *Parallel Output + Reduction* condition the time spent in stages 1 and 2 was almost the same.

5.3.3 Steps per Search Stage

The mean number of steps taken in each search stage, regardless of audio output, was as follows: 3.34 for the first stage; 3.55 for the second stage; and 2.50 for the third stage. There was a significant difference in the number of steps taken per search stage when the type of audio output was ignored, $F(2, 70) = 21.16$, $p < .05$, $\omega^2 = .13$, with sphericity assumed, $\chi^2(2) = 5.30$, $p > .05$.

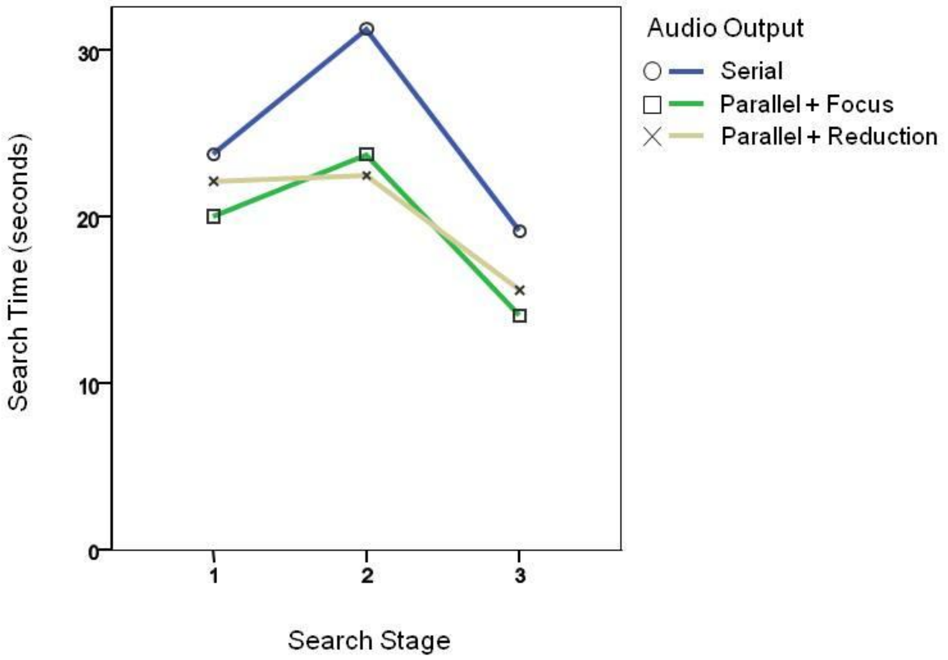

Figure 15 – Interaction graph for the duration of each search stage in the three types of audio output (experiment 1)

Repeated-measures t-tests (using a Bonferroni adjustment, α = .05/3 = .017) showed that the third stage required significantly less steps than the first two stages (stage 1 versus stage 3: $t(35) = 5.24$, $p < .017$, $r = .66$; stage 2 versus stage 3: $t(35) = 5.92$, $p < .017$, $r = .71$). Stages 1 and 2 were not significantly different from each other, $t(35) = -1.23$, $p > .017$. These results are consistent with the differences in the duration of the stages.

The interaction between the type of audio output and the search stages (Figure 16) was not statistically significant for the number of steps, $F(3.34, 116.95) = 1.95$, $p > .05$, with degrees of freedom corrected using the Huynh-Feldt estimates of sphericity (ε = .84), since Mauchly's test was significant, $\chi^2(9) = 19.61$, $p < .05$.

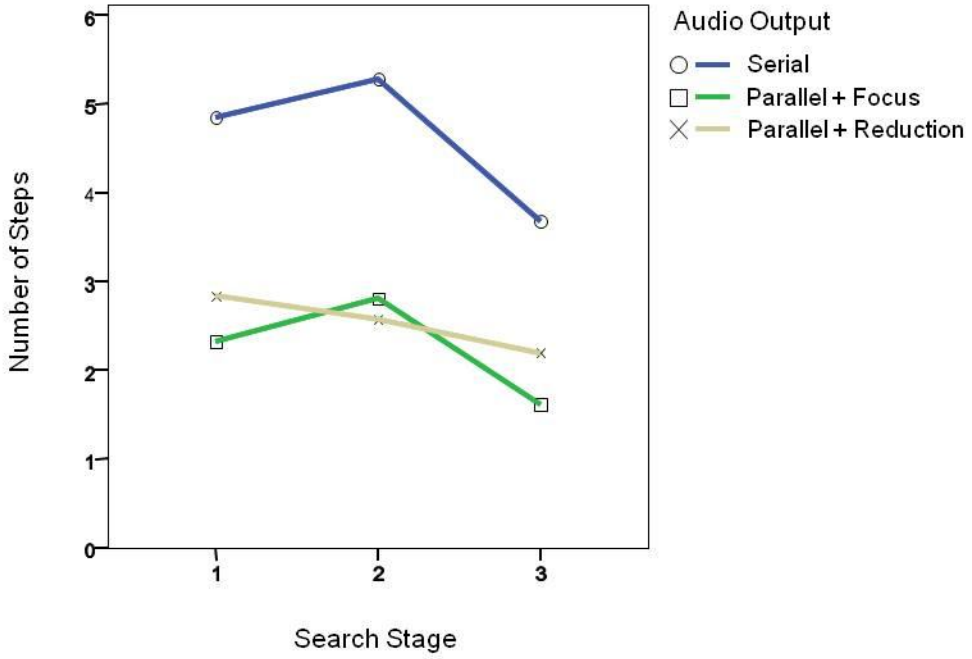

Figure 16 – Interaction graph for the number of steps per search stage in the three types of audio output (experiment 1)

5.3.4 Time before First Step per Search Stage

This analysis examines how long it took participants to take their first step in each search stage. The goal of this analysis is to detect if participants made an effort to listen to the parallel output before taking a step.

Note that each search stage begins with the announcement of the stage number, which takes about 2 seconds. Only after the stage announcement, participants can start listening to the audio recordings. The values reported in this section for the *time taken before the first step* were recorded from the beginning of each stage and include the time listening to the stage announcement.

The mean time taken before the first step was 3.69 seconds on the *Serial* form of audio output; 7.90 seconds on the *Parallel* + *Focus* form; and 8.04 seconds on the *Parallel* + *Reduction* form, when the search stage was ignored. The two-way repeated-measures ANOVA shows that these times are significantly different, $F(1.75, 61.33) = 33.61$, $p < .05$, $\omega^2 = .32$, with degrees of freedom corrected using the Huynh-Feldt estimates of sphericity ($\varepsilon = .88$) due to Mauchly's test of sphericity being significant ($\chi^2(2) = 7.26$, $p < .05$).

Contrasts revealed that participants waited significantly less time before taking a step in the serial form of audio output then in the parallel forms, $F(1, 35) = 62.34$, $p < .05$, $r = .80$. There was no significant difference between the *Focus* and *Reduction* forms of parallel audio output, $F(1, 35) = .06$, $p > .05$.

The mean listening times before the first step was 6.97 seconds in the first stage, 6.93 seconds in the second stage, and 5.72 seconds in the third stage. There was a significant difference between the number of seconds before the first step in each search stage when the type of audio output was ignored, $F(2, 70) = 8.25$, $p < .05$, $\omega^2 = .05$, with sphericity assumed, $\chi^2(2) = 3.50$, $p > .05$.

Repeated-measures t-tests (using a Bonferroni adjustment, $\alpha = .05/3 = .017$) showed that participants waited significantly less time before taking a step in stage 3 than in the first two stages (stage 1 versus stage 3: $t(35) = 3.19$, $p < .017$, $r = .48$; stage 2 versus stage 3: $t(35) = 4.09$, $p < .017$, $r = .57$). There was no significant difference between the first and second stages (stage 1 versus stage 2: $t(35) = .11$, $p > .017$).

The interaction between the type of audio output and the search stages was not statistically significant for the number of steps, $F(4, 140) = 2.36$, $p > .05$, with uncorrected degrees of freedom, since Mauchly's test of sphericity was not significant, $\chi^2(9) = 9.86$, $p > .05$. The graph in Figure 17 shows that the waiting time before the first step did not vary from stage to stage when the audio was available serially, and was much shorter than for the two types of parallel output.

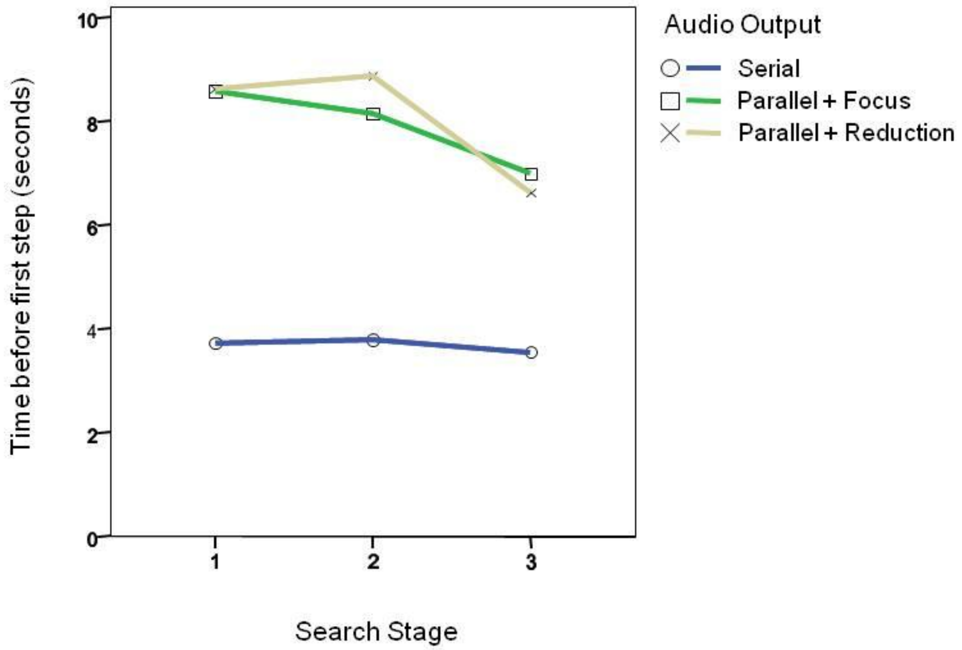

Figure 17 – Time to take the first step in each search stage per audio output (experiment 1)

5.3.5 Time Listening to the Target Sample per Search Stage

After the initial 30-second playback of the target sample, participants had to search for it going through the three stages of the search process. At any time during the search, participants were allowed to listen to the target sample again, as many times and for as long as they needed. The amount of time spent on the target sample during the search was recorded by stage, and analyzed for differences between the audio output conditions.

Mauchly's test indicated that the assumption of sphericity was met for the main effect of the search stage, $\chi^2(2) = 4.20$, $p > .05$, but was violated for the main effect of the type of audio output, $\chi^2(2) = 21.96$, $p < .05$, and the interaction effect, $\chi^2(9) = 52.33$, $p < .05$. Therefore, the degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity ($\varepsilon = .70$ for the main effect of audio output and .55 for the interaction effect).

The average time spent listening to the target sample was 0.89 seconds on the *Serial* form of audio output; 0.44 seconds on the *Parallel* + *Focus* form; and 0.45 seconds on the *Parallel* + *Reduction* form, when the search stage was ignored. The two-way repeated-measures ANOVA shows that these times are significantly different, $F(1.39, 48.62) = 4.67$, $p < .05$, $\omega^2 = .03$. Contrasts showed that participants spent significantly more time listening to the target sample in the serial form of audio output then in either parallel form $F(1, 35) = 5.52$, $p < .05$, $r = .37$. There was no significant difference between the *Focus* and *Reduction* forms of parallel audio output, $F(1, 35) = .01$, $p > .05$.

When the form of audio output was ignored, the average time taken listening to the target sample was 0.20 seconds on the first search stage; 0.72 seconds on the second stage; and 0.86 seconds on the third and final stage. These times were also significantly different, $F(2, 70) = 4.62$, $p < .05$, $\omega^2 = .04$. Repeated-measures t-tests (using a Bonferroni adjustment, $\alpha = .05/3 = .017$) showed that participants spent significantly more time listening to the target sample in the third stage of the search than in the first stage (stage 1 versus stage 3: $t(35) = -2.27$, $p < .017$, $r =.36$). The other differences were not statistically significant (stage 1 versus stage 2: $t(35) = -2.06$, $p > .017$; stage 2 versus stage 3: $t(35) = -.771$, $p > .017$).

The interaction effect between the type of audio output and the search stage was not statistically significant, $F(2.21, 77.24) = 1.06$, $p > .05$. The interaction graph in Figure 18 shows that participants rarely listened to the target sample in stage 1 (compared to the other stages), regardless of the type of audio output. In stages 2 and 3, participants spent more time on the target sample when searching serial output than when searching parallel output.

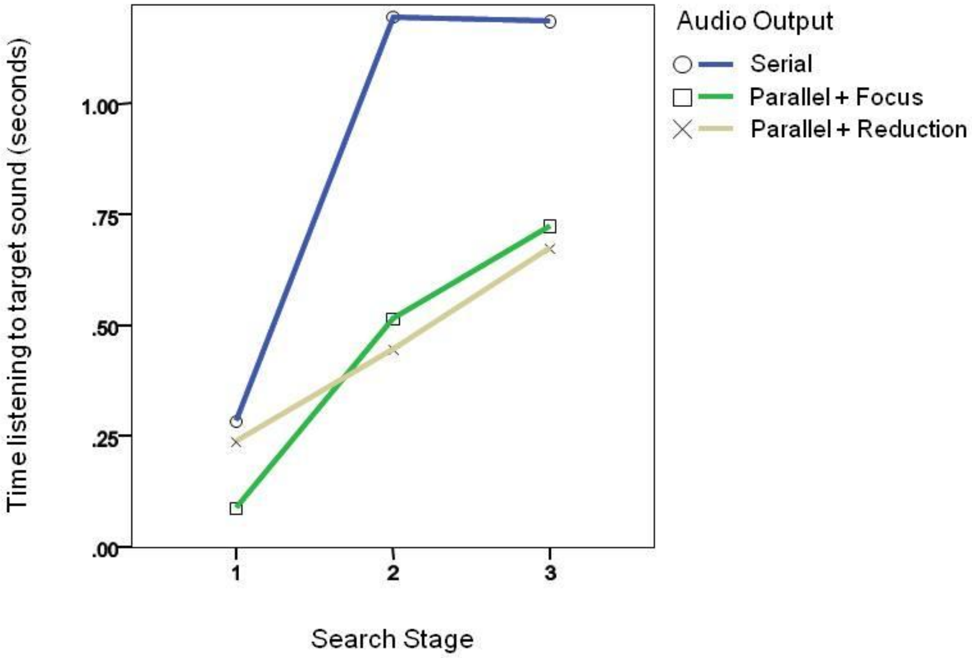

Figure 18 – Time spent listening to the target sample during the search (experiment 1)

5.3.6 Number of Incomplete Searches

Table 2 summarizes the total number of incomplete searches per type of audio output. All incomplete searches happened because of timeouts, that is, the time limit of 4 minutes was reached before the participant could complete the search. Eleven participants had at least one timeout and nobody had more than one timeout per type of audio output. The total number of timeouts represents less than 5% of all searches.

Table 2 – Number of incomplete searches per type of audio output (experiment 1)

Number of Searches	Serial	Parallel + Focus	Parallel + Reduction	Total
Incomplete	6	5	4	15
Complete	102	103	104	309
Total	108	108	108	324

5.3.7 Number of Backtracks

Out of 309 complete searches, 46 required participants to use the backtrack button. The maximum number of backtracks on a single search was four, but typically only one backtrack was needed for participants to find their way to the target sample. The total number of backtracks in each type of audio output is shown on Table 3. Searches using serial output had over 75% more backtracks than searches using parallel output, indicating that participants chose incorrect genres more often when searching serial output.

Table 3 – Number of backtracks per type of audio output (experiment 1)

	Serial	Parallel + Focus	Parallel + Reduction
Complete searches	102	103	104
Searches with backtrack(s)	21	12	13
Total number of backtracks	34	19	19

5.4 Experiment 1: Survey

Participants were asked to rate how easy it was to search using each type of audio output (*Serial*, *Parallel + Focus*, and *Parallel + Reduction*) on a scale of 1 to 5, where, 1 is very difficult, 2 is difficult, 3 is neutral, 4 is easy, and 5 is very easy. All 36 participants rated all three types of search. Nobody found any of the types "very difficult". The answers are presented in Figures 19, 20, and 21.

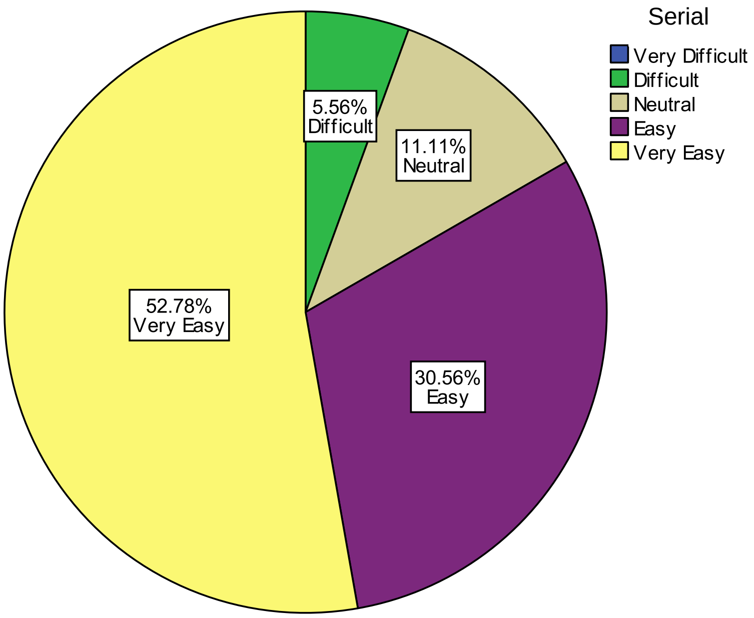

Figure 19 – Difficulty ratings for searching *Serial Output* (experiment 1)

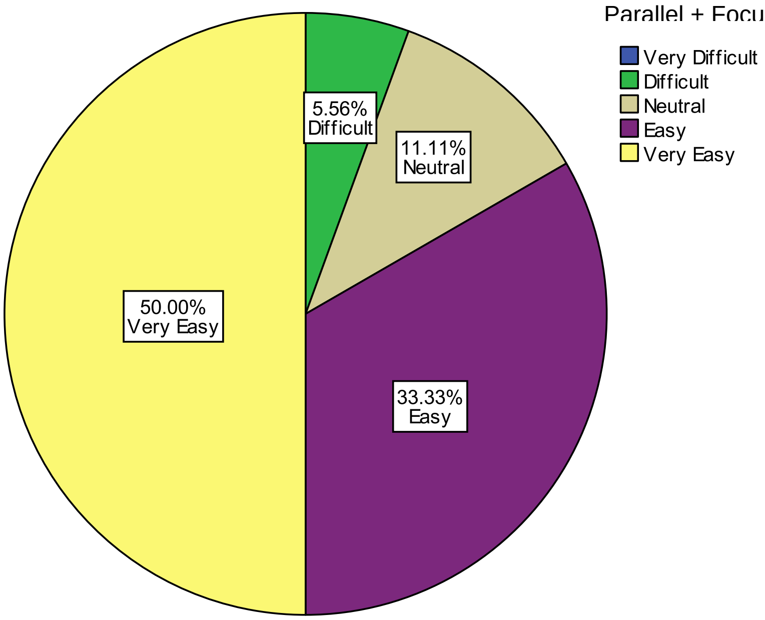

Figure 20 – Difficulty ratings for searching *Parallel Output* + *Focus* (experiment 1)

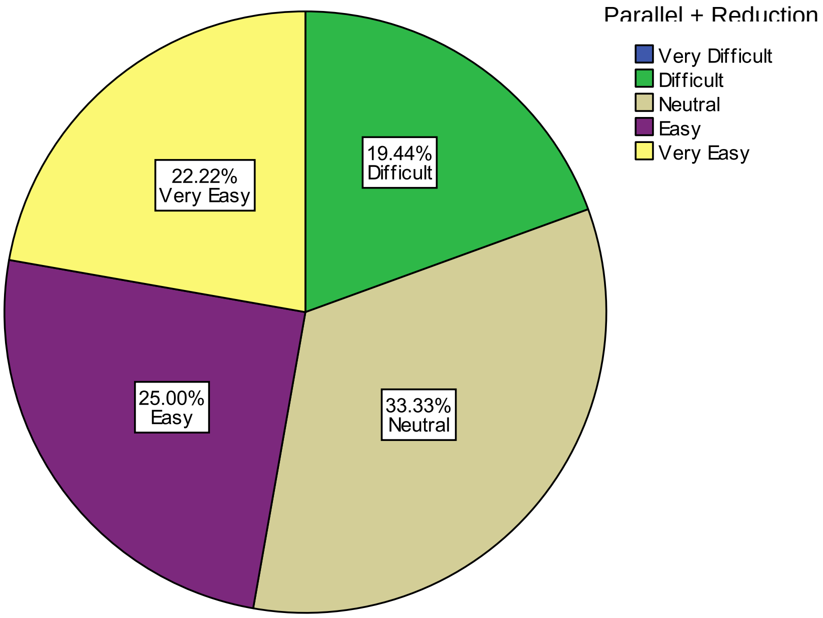

Figure 21 – Difficulty ratings for searching *Parallel Output* + *Reduction* (experiment 1)

Participants were also asked which type of search they would prefer to use and which one they perceived as being the fastest to find the target sample. Their answers are presented in Figures 22 and 23. Out of 36 participants, 24 correctly identified their fastest search type. Ten participants thought their preferred search type was also the fastest, when in reality it was not. The other two participants simply chose a slower search type for no apparent reason.

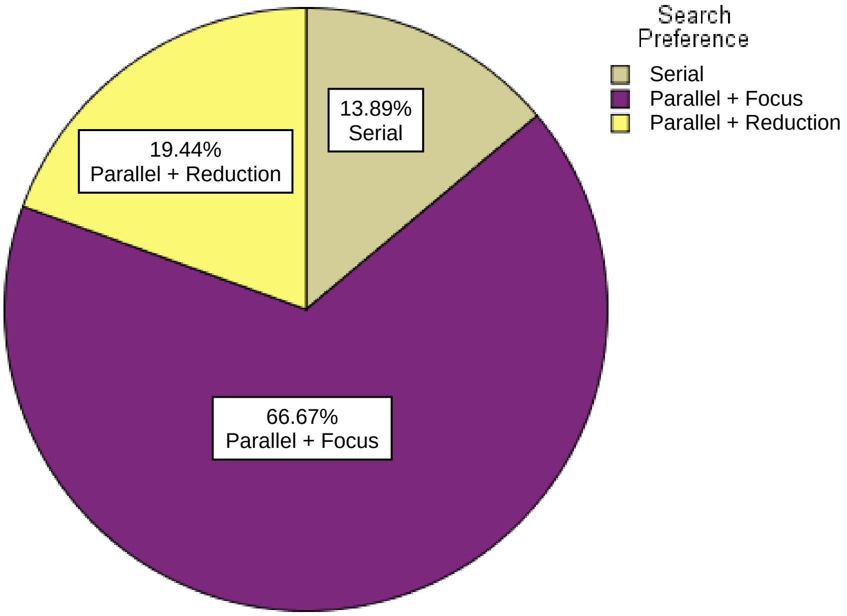

Figure 22 – Preference ratings for the three types of search (experiment 1)

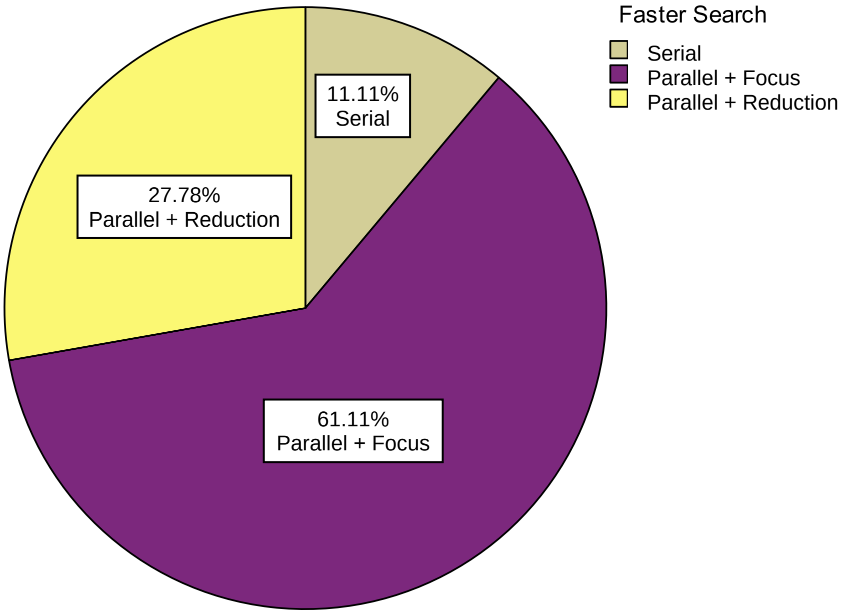

Figure 23 – Type of search that participants perceived as the fastest (experiment 1)

5.5 Experiment 2: Main Effects

The second experiment also compares the effort required to find a target sample in three different audio output conditions. In the *Serial* condition participants listened to one song at a time while in the two parallel conditions, they listened to six songs at the same time, with the option to either listen momentarily to any individual song in isolation (*Parallel Output* + *Focus* condition) or stop one or more songs from playing (*Parallel Output* + *Reduction* condition). The effort to find the target sample is represented by the two dependent variables: *search time* and *number of steps*. Data from 18 participants were used in this analysis.

A repeated-measures MANOVA was conducted and revealed, using Pillai's Trace, a significant effect of audio output on the effort required to find the target sample, $V = 0.45$, $F(4, 68) = 4.91$, $p < .05$. The results of univariate ANOVAs were then examined and are presented in the following sections.

5.5.1 Effect of Audio Output on Search Time

The means charted in Figure 24 suggest that the target sample was found faster when the search was performed with multiple songs being played simultaneously than with the songs played one at a time. A one-way repeated-measures ANOVA was performed to confirm this impression. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 0.83$, $p > .05$, therefore no correction was needed.

The results show that the *search time* was significantly affected by the type of audio output during the search, $F(2, 34) = 5.10$, $p < .05$, $\omega^2 = .07$. Planned contrasts revealed that the serial audio presentation produced significantly longer search times compared to the parallel conditions, $F(1, 17) = 7.48$, $p < .05$, $r = .55$. The *Parallel* + *Focus* type of audio output ($M = 76.18$, $SD = 33.47$) was on average faster than the *Parallel* +

Reduction type of output (*M* = 85.13, *SD* = 37.07), but that difference was not statistically significant, *F*(1, 17) = 1.34, *p* > .05.

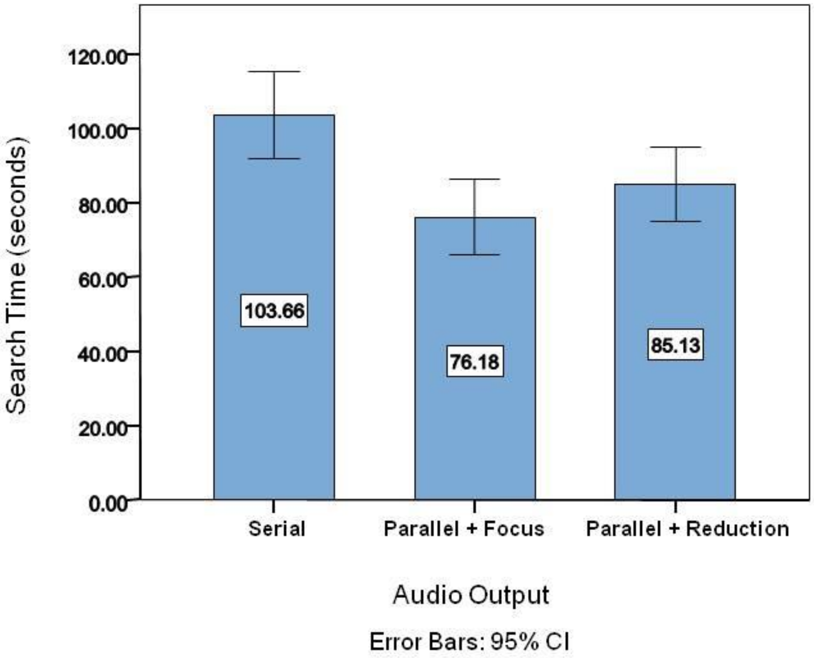

Figure 24 – Error bar chart of the mean duration of audio searches (lower is better) in three different forms of audio output (experiment 2)

5.5.2 Effect of Audio Output on Number of Steps

The number of steps required to find the target sample with the serial audio output was larger than with either parallel forms of audio output (Figure 25).

A one-way repeated-measures ANOVA was conducted on these data. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 3.01$, *p* > .05, therefore no correction was needed.

The analysis confirm a significant effect of the form of audio output on the number of steps taken to find the target sample, *F*(2, 34) = 13.40, *p* < .05, ω^2 = .28. Contrasts revealed a significantly larger number of steps when the audio output was serial as opposed to parallel, *F*(1, 17) = 30.06, *p* < .05, *r* = .80, but there was no significant difference between the two parallel conditions, *F*(1,17) = 2.68, *p* > .05.

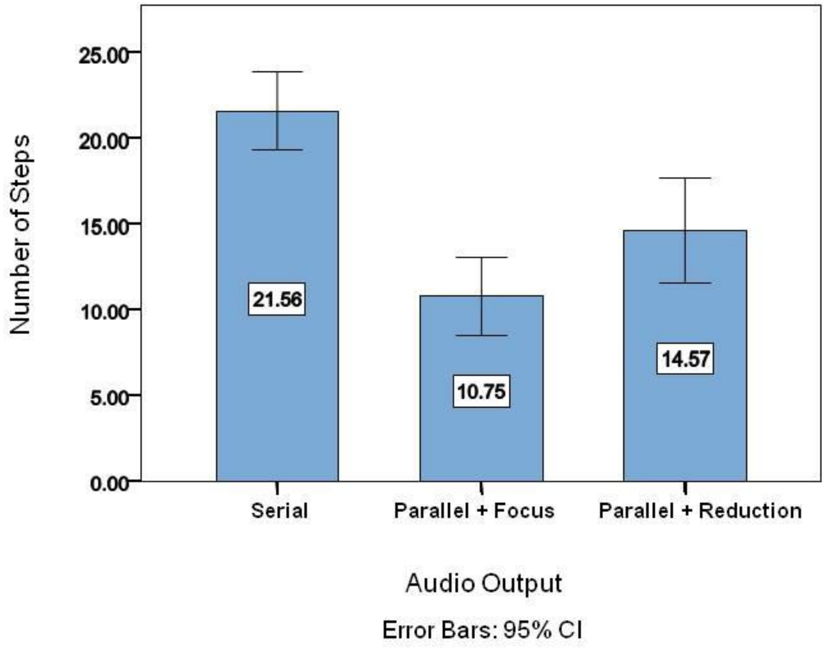

Figure 25 – Error bar chart of the mean number of steps taken during audio searches (lower is better) for three audio output conditions (experiment 2)

5.6 Experiment 2: Additional Effects

5.6.1 Duration of Search Stages

The following analysis compares the three search stages in the correct path towards the target sample. It was expected that stage 2 would be the most difficult and consequently the longest. In order to compare the duration of each stage overall and in each audio output condition, a two-way repeated measures ANOVA was performed.

Mauchly's test indicated that the assumption of sphericity had been met for the main effect of search stage, $\chi^2(2) = 4.73$, $p > .05$, and for the interaction effect between the type of audio output and the search stage, $\chi^2(9) = 21.89$, $p > .05$, therefore no correction was needed.

The mean duration of each search stage, regardless of audio output, was as follows: 30.41 seconds for the first stage; 30.94 seconds for the second stage; and 20.73

for the third stage. There was a significant difference in the duration of the search stages when the type of audio output was ignored, $F(2, 34) = 24.88$, $p < .05$, $\omega^2 = .14$.

Repeated-measures t-tests (using a Bonferroni adjustment, $\alpha = .05/3 = .017$) showed that participants took significantly less time in the third stage than in any other stage (stage 1 versus stage 3: $t(17) = 6.39$, $p < .017$, $r = .84$; stage 2 versus stage 3: $t(17) = 7.79$, $p < .017$, $r = .88$). There was no significant difference between the duration of stages 1 and 2 ($t(17) = -.27$, $p > .017$).

The interaction between the type of audio output and the search stages was not statistically significant, $F(4, 68) = 1.75$, $p > .05$. The graph in Figure 26 shows this interaction and reveals that the most drastic differences in stage duration occurred in the *Parallel Output* + *Reduction* condition and the smallest differences in the *Parallel Output* + *Focus* condition. Stage 3 was the shortest in all forms of audio output. In the *Serial* condition, stages 1 and 2 lasted approximately the same time, suggesting that most participants had to listen to all six audio recordings before making a decision in both stages. In stage 3, as soon as the target sample was heard, the search was completed, without the need to listen to the other recordings.

5.6.2 Steps per Search Stage

The mean number of steps taken in each search stage regardless of the type of audio output was as follows: 5.52 on the first stage; 5.07 on the second stage; and 3.68 on the third stage. The two-way repeated-measures ANOVA shows that these numbers were significantly different, $F(2, 34) = 15.63$, $p < .05$, $\omega^2 = .21$, with sphericity assumed, $\chi^2(2) = 4.47$, $p > .05$.

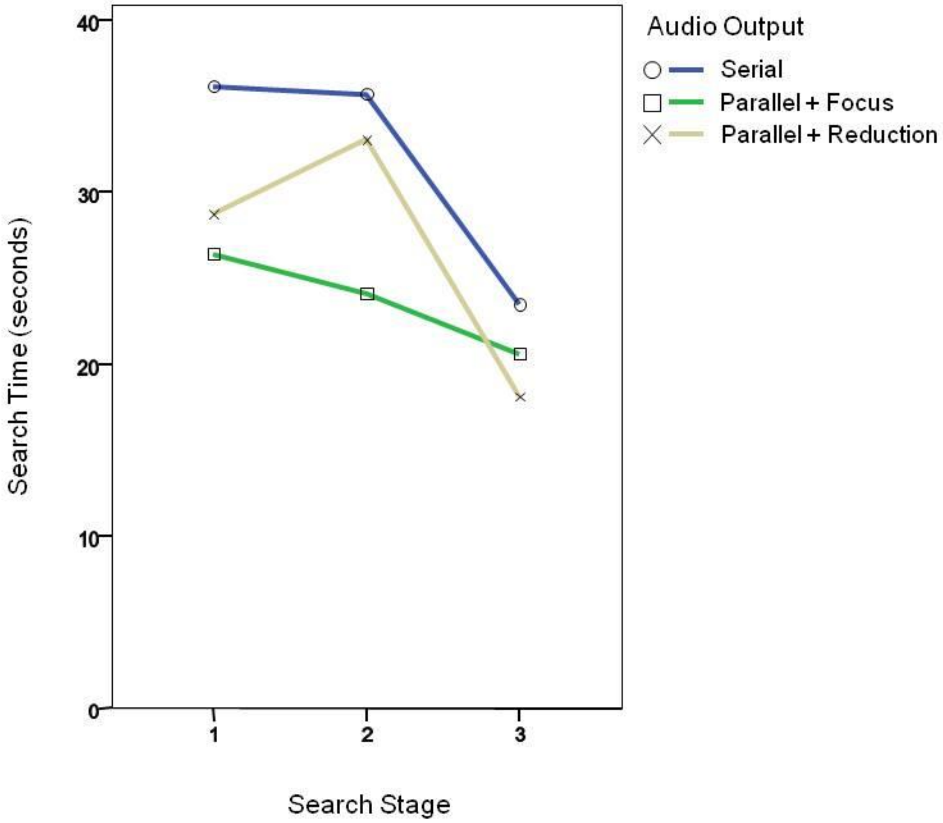

Figure 26 – Interaction graph for the duration of each search stage in the three types of audio output (experiment 2)

Repeated-measures t-tests (using a Bonferroni adjustment, α = .05/3 = .017) showed that participants took significantly less steps in the third stage than in any other stage (stage 1 versus stage 3: $t(17) = 4.44$, $p < .017$, $r = .73$; stage 2 versus stage 3: $t(17) = 5.23$, $p < .017$, $r = .79$). There was no significant difference between the number of steps taken in stage 1 and stage 2 ($t(17) = 1.33$, $p > .017$).

The interaction between the type of audio output and the search stages was not statistically significant for the number of steps (Figure 27), $F(4, 68) = 1.45$, $p > .05$, with sphericity assumed, $\chi^2(9) = 9.33$, $p > .05$.

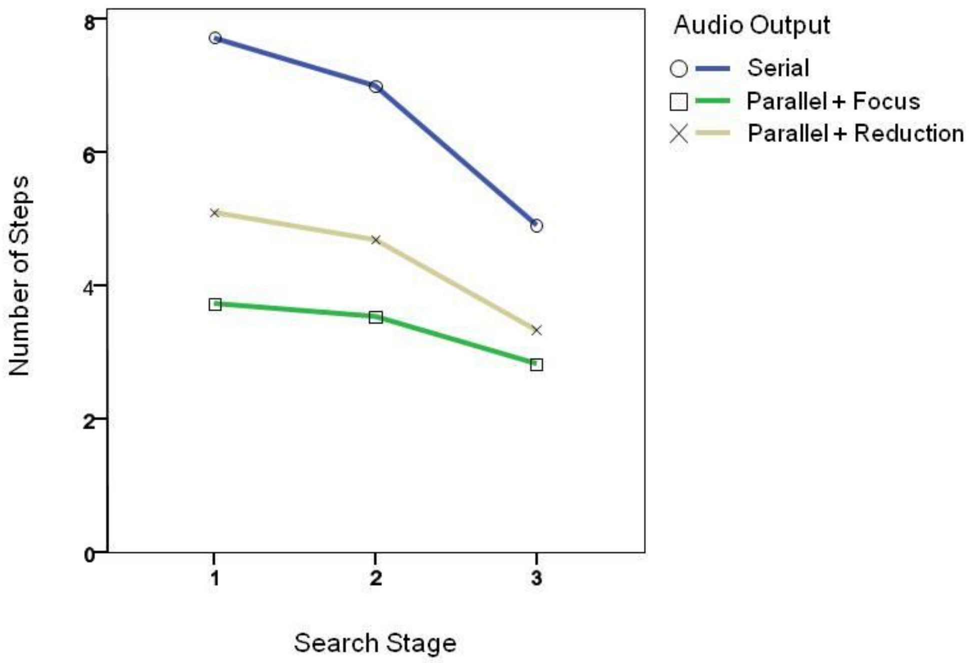

Figure 27 – Interaction graph for the number of steps per stage in the three types of audio output (experiment 2)

5.6.3 Time before First Step per Search Stage

This analysis examines how long it took participants to take their first step in each search stage. It was expected that participants would listen to the simultaneously playing songs for a few seconds before using the *Focus* or *Remove* buttons.

Note that each search stage begins with the announcement of the stage number, which takes about 2 seconds. Only after the stage announcement, participants can start listening to the audio recordings. The values reported in this section for the *time taken before the first step* were recorded from the beginning of each stage and include the time listening to the stage announcement.

The mean time taken before the first step was 4.73 seconds on the *Serial* form of audio output; 8.10 seconds on the *Parallel + Focus* form; and 8.09 seconds on the *Parallel + Reduction* form, when the search stage was ignored.

Mauchly's test indicated that the assumption of sphericity was met for the main effect of the type of audio output, $\chi^2(2) = 2.20$, $p > .05$, but was violated for the main effect of the search stage, $\chi^2(2) = 22.27$, $p < .05$, and the interaction effect, $\chi^2(9) = 17.86$, $p < .05$. Therefore, the degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity ($\varepsilon = .59$ for the main effect of search stage and .83 for the interaction effect).

The two-way repeated-measures ANOVA shows that the time before the first step was significantly different in the three forms of audio output, $F(2, 34) = 29.18$, $p < .05$, $\omega^2 = .24$.

Contrasts revealed that participants waited significantly less time before taking a step in the serial form of audio output then in the parallel forms, $F(1, 17) = 61.38$, $p < .05$, $r = .89$. There was no significant difference between the *Focus* and *Reduction* forms of parallel audio output, $F(1, 17) = .001$, $p > .05$.

The mean time taken before the first step was 6.75 seconds on the first search stage; 7.26 seconds on the second stage; and 6.92 seconds on the third and final stage, when the audio output was ignored. These times were not significantly different, $F(1.17, 19.89) = .42$, $p > .05$.

There was a significant interaction effect between the type of audio output and the search stage, $F(3.34, 56.74) = 3.79$, $p < .05$. This indicates that the audio output had different effects on the time before the first step, depending on the stage of the search. To break down this interaction, contrasts were performed comparing the two parallel conditions to the serial one and to each other, and all stages to stage 2, which was expected to be the longest one. These revealed significant interactions when comparing serial to parallel, both for stage 1 compared to stage 2, $F(1, 17) = 13.27$, $p < .05$, $r = .66$, and stage 3 compared to stage 2, $F(1, 17) = 7.32$, $p < .05$, $r = .55$. The remaining contrasts revealed no significant interaction term when comparing the *Focus* and the

Reduction forms of audio output, both for stage 1 compared to stage 2, $F(1, 17) = 1.91$, $p > .05$, and stage 3 compared to stage 2, $F(1, 17) = .02$, $p > .05$.

The interaction graph in Figure 28 shows that for serial searches, the waiting time before the first step was the shortest in stage 2 (compared to the other stages) and longest in stage 1. In contrast, for parallel searches, stage 2 had the longest waiting time.

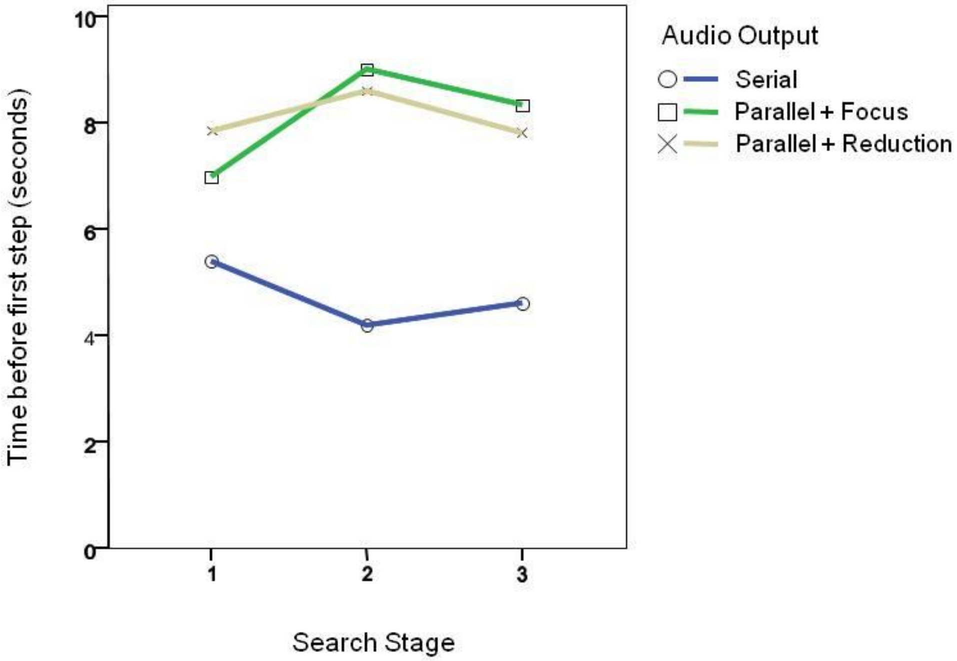

Figure 28 – Time to take the first step in each search stage per audio output (experiment 2)

5.6.4 Time Listening to the Target Sample per Search Stage

At any time during the search, participants were allowed to listen to the target sample again, as many times and for as long as they needed. The amount of time spent on the target sample during the search was recorded by stage, and analyzed for differences between the audio output conditions.

Mauchly's test indicated that the assumption of sphericity was violated for the main effect of the type of audio output, $\chi^2(2) = 7.52$, $p < .05$, the main effect of the search stage, $\chi^2(2) = 14.65$, $p < .05$, and the interaction effect, $\chi^2(9) = 20.47$, $p < .05$. Therefore, the degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity ($\varepsilon = .78$ for the main effect of audio output, .65 for the main effect of search stage, and .82 for the interaction effect).

The mean time spent listening to the target sample was 1.84 seconds on the *Serial* form of audio output; 1.42 seconds on the *Parallel* + *Focus* form; and 1.54 seconds on the *Parallel* + *Reduction* form, when the search stage was ignored. The two-way repeated-measures ANOVA shows that these times are not significantly different, $F(1.56, 26.44) = .81$, $p > .05$.

When the audio output was ignored, the average time spent on the target sample was 1.27 seconds on the first search stage; 1.66 seconds on the second stage; and 1.87 seconds on the third and final stage. That shows an increased need to replay the target sample as the search progressed, as expected. However, these differences between stage means were not statistically significant, $F(1.30, 22.13) = 1.45$, $p > .05$.

The interaction effect between the type of audio output and the search stage was not statistically significant either, $F(3.30, 56.04) = 1.65$, $p > .05$.

The interaction graph in Figure 29 shows that in the *Parallel Output* + *Focus* condition, the time spent on the target sample increased as the search advanced, as expected, due to the initial playback of the target sample being flushed out of memory with time and the other songs being played. Curiously, this increase did not happen in the other forms of audio output.

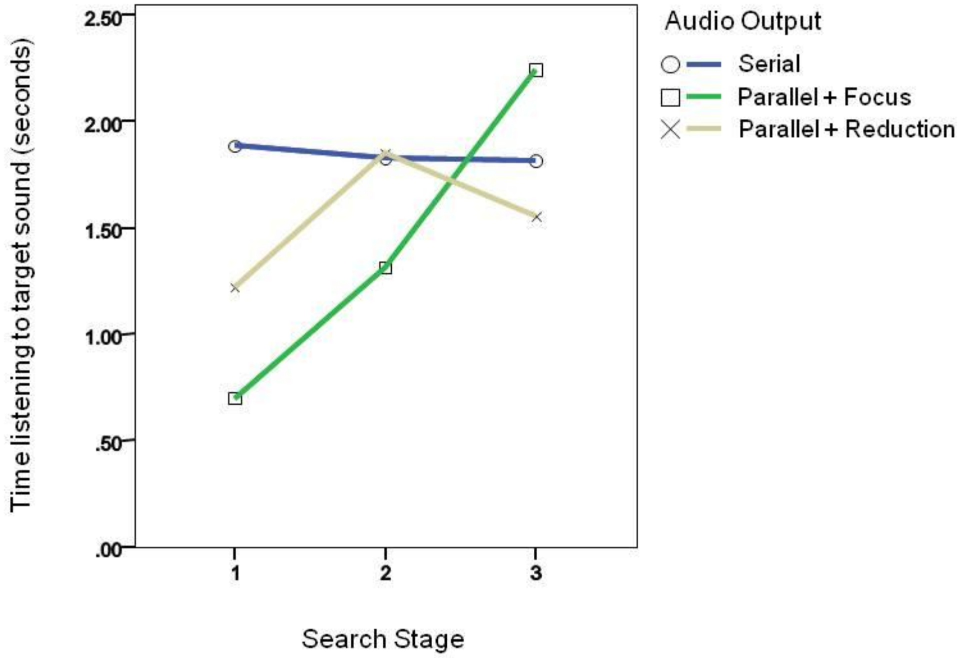

Figure 29 – Time spent listening to the target sample during the search (experiment 2)

5.6.5 Number of Incomplete Searches

Table 4 summarizes the total number of incomplete searches per type of audio output. All incomplete searches happened because of timeouts, that is, the time limit of 5 minutes was reached before the participant could complete the search. Four participants had at least one timeout and nobody had more than one timeout per type of audio output. The total number of timeouts represents less than 5% of all searches.

Table 4 – Number of incomplete searches per type of audio output (experiment 2)

Number of Searches	Serial	Parallel + Focus	Parallel + Reduction	Total
Incomplete	2	2	3	7
Complete	52	52	51	155
Total	54	54	54	162

5.6.6 Number of Backtracks

Out of 155 complete searches, 20 required participants to use the backtrack button. The maximum number of backtracks on a single search was four, but typically only one backtrack was needed for participants to find their way to the target sample. The total number of backtracks in each type of audio output is shown on Table 5. Searches using serial output had over 65% more backtracks than searches using parallel output, indicating that participants chose incorrect genres more often when searching serial output.

Table 5 – Number of backtracks per type of audio output (experiment 2)

	Serial	Parallel + Focus	Parallel + Reduction
Complete searches	52	52	51
Searches with backtrack(s)	9	5	6
Total number of backtracks	15	8	9

5.7 Experiment 2: Survey

Participants were asked to rate how easy to search using each type of audio output on a scale of 1 to 5, where, 1 is very difficult, 2 is difficult, 3 is neutral, 4 is easy, and 5 is very easy. All 18 participants rated all three types of search. Nobody found any of the types "very difficult". The answers are presented in Figures 30, 31, and 32.

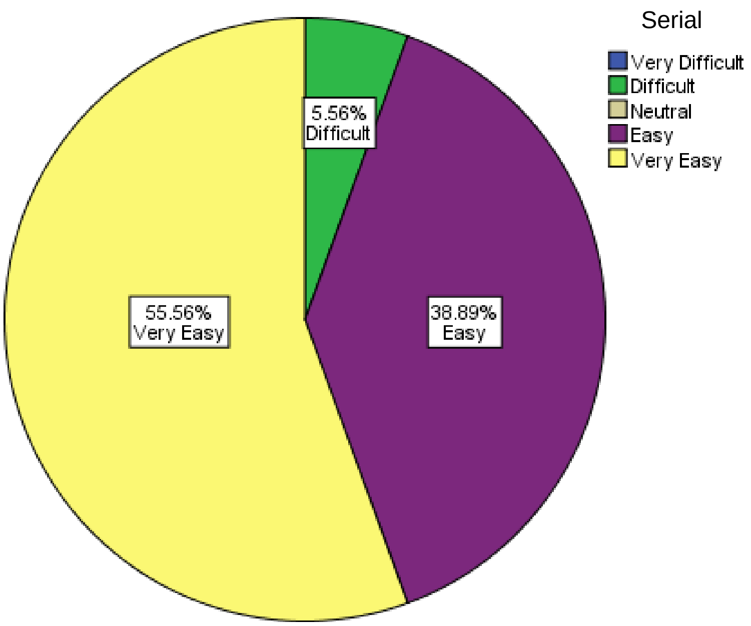

Figure 30 – Difficulty ratings for searching *Serial Output* (experiment 2)

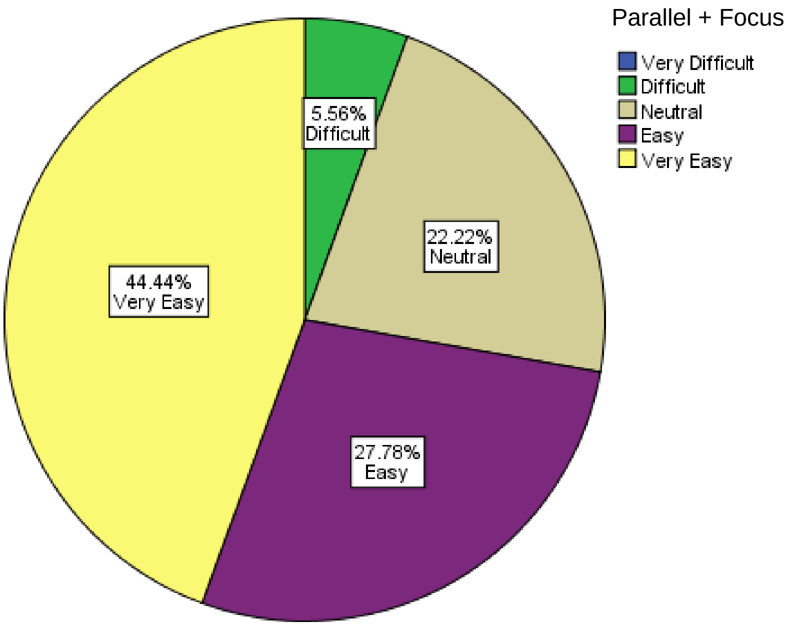

Figure 31 – Difficulty ratings for searching *Parallel Output* + *Focus* (experiment 2)

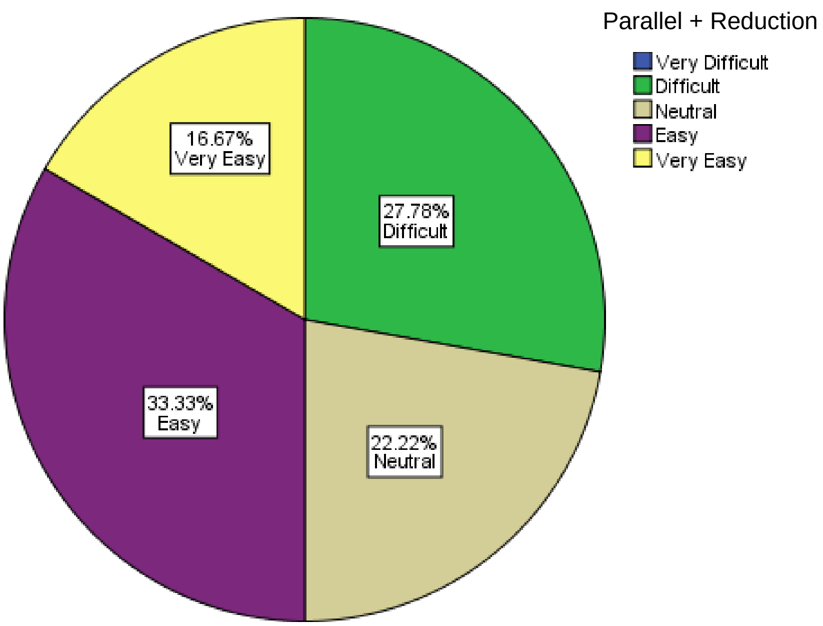

Figure 32 – Difficulty ratings searching *Parallel Output* + *Reduction* (experiment 2)

Participants were also asked which search form they would prefer to use and which one they perceived as being the fastest to find the target sample. Their answers are presented in Figures 33 and 34. Out of 18 participants, 13 correctly identified their fastest form of search. Three participants perceived their preferred search type as also being the fastest, when in reality it was not. The other two participants simply chose a slower search form for no apparent reason.

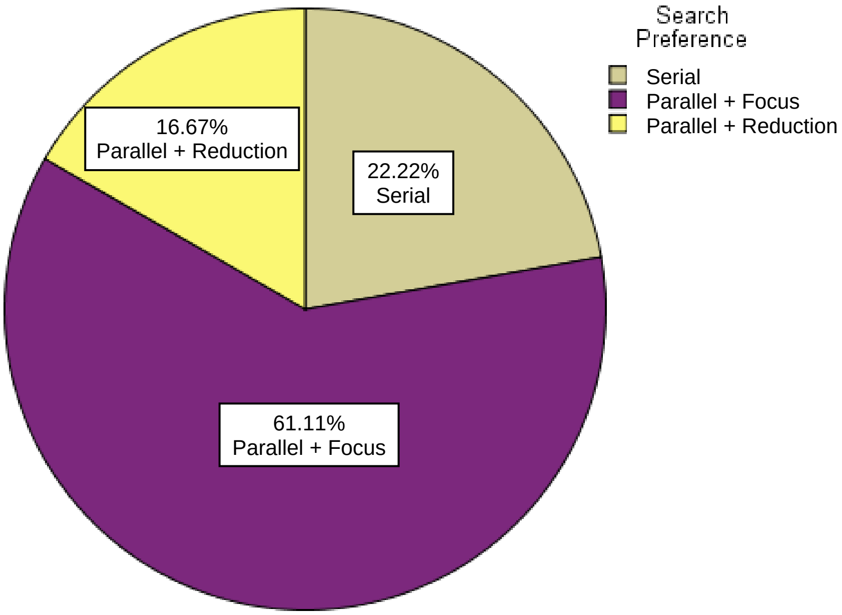

Figure 33 – Preference ratings for the three types of search (experiment 2)

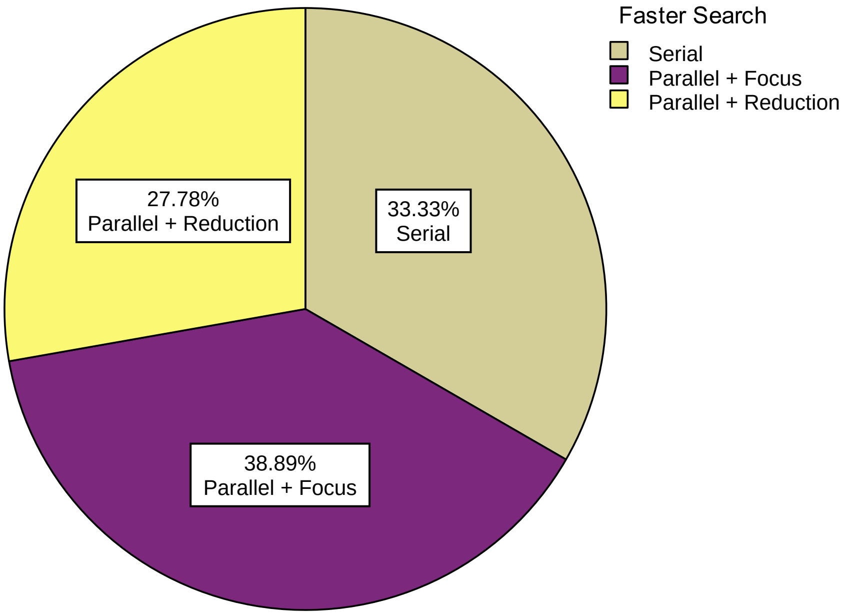

Figure 34 – Type of search that participants perceived as the fastest (experiment 2)

Chapter 6

Discussion and Future Work

6.1 General Discussion

The results of this research indicate that searching parallel output can be a valuable technique when compared to the current methods of audio search, which typically use serial output. The results confirm the study's prediction that an audio player that allows parallel audio presentation would produce more effective searches than an audio player that only allows serial presentation.

Searches were performed significantly faster using parallel output, in both the four-speaker configuration (experiment 1) and the six-speaker configuration (experiment 2). There was no significant difference between the two types of parallel output (*Parallel + Focus* and *Parallel + Reduction*), but they both produced faster searches than serial output. Moreover, the three search stages were faster when parallel output was used.

The total distance to the target sample, measured in number of steps, was significantly shorter using parallel output in both speaker configurations. All search stages required fewer steps when searching parallel output. Table 6 presents the mean number of steps per search stage. Note that these numbers do not include the steps taken "out-of-path" (see section 4.3.5). With serial output, the mean number of steps in stages 1 and 2 was greater than the number of audio recordings available per stage, indicating that some recordings were heard more than once. In the first two stages, the goal was to find the closest match to the target sample. In stage 3, where the goal was to find an exact match, the mean number of steps was slightly less than the number of recordings, because as soon as an exact match was found there was no need to listen to the other audio files. With parallel output, all stages required fewer steps on average than the number of options, indicating participants made decisions without the need to listen to each song in isolation.

Table 6 – Mean number of steps per stage

	Four-speaker configuration			Six-speaker configuration		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Serial	4.85	5.28	3.68	7.72	7.0	4.90
Simul+Focus	2.32	2.81	1.61	3.72	3.53	2.82
Simul+Reduction	2.84	2.57	2.19	5.10	4.69	3.33

Participants demonstrated an effort to listen to the parallel output in each stage before taking a step in the search. In searches using serial output, it was expected that participants would take a step – start listening to the songs – very quickly, since otherwise there was only silence. In the parallel output conditions, the music started automatically, and the first step was taken when participants focused on a sound (*Parallel* + *Focus*), muted a sound (*Parallel* + *Reduction*), selected a sound and moved to the next stage, or backtracked to the previous stage. It was expected that participants would listen to the simultaneously playing recordings for a few seconds before using the *Focus* or *Remove* buttons. On both experiments, participants took their first step approximately four seconds later, on average, when searching parallel output than when searching serial output. This difference was statistically significant with a large effect size. As predicted, participants listened to the simultaneously playing sounds for a few seconds before taking any action in each search stage.

For the most part, searches were successfully completed within the allowed time limit. The number of incomplete searches was less than 5% of all searches and was approximately the same in all types of audio output, indicating that the search tasks were not too hard to be completed in the time given, and that the type of audio output did not affect participants' ability to complete the search. Most incomplete searches happened because participants chose an incorrect audio category in the first stage of search, and spent too much time exploring its sub-categories in stages 2 and 3 without realizing they needed to backtrack all the way to stage 1 and choose another category. It is expected that this problem can be reduced with the use of better audio categorization.

In the first experiment, out of the 324 searches, 61 required at least one backtrack, that is, participants chose at least one incorrect audio category in about 19% of searches. Only 25% of those searches ended incomplete, which means participants were able to recover from choosing an incorrect category and complete the search successfully 75% of the time. The second experiment confirmed this finding with very similar results: Participants used the backtrack feature in 17% of the searches and successfully completed 74% of those searches.

Both experiments demonstrated an increased use of the backtrack feature in searches in serial output when compared to searches in parallel output, indicating that participants made more mistakes choosing audio categories when listening to them serially. This could be because they chose an audio category without listening to all of the options, or because it was more difficult to compare multiple audio samples while listening to them one at a time. According to Brown, Brewster, Ramloll, Yu, & Riedel (2002), audio comparisons can be more easily made when sounds are presented concurrently rather than serially.

Most incorrect category selections happened in the second stage of the search. This stage was expected to be the most difficult because the audio categories are more similar to each other than in stage 1: each stage has audio recordings that belong to sub-categories of the selected category from the previous stage. In stage 3, the audio recordings are even more similar, but since the target sample is expected to be found in that stage, participants' goal became to recognize the exact recording, which was an easier task than matching a category, which was the goal in stages 1 and 2.

In the first experiment, there was a significant difference in the duration of the search stages. The second stage was the longest and required the most steps and the third stage the shortest with the least number of steps, as expected. The second experiment also revealed the third stage to be significantly faster and require fewer steps, but showed no differences between the first and second stages. This could be attributed to the increased memory requirements of experiment 2, where each stage has six categories from which to choose instead of four. The extra categories seem to have made the first two stages comparable in terms of difficulty. The third stage is still the easiest because it does not require matching a sound to its category, but matching two identical sounds.

Even though the third stage was the fastest of the three search stages, participants spent more time listening to the target sample on the third stage than in any other stage. Since the sounds in the experiments were unfamiliar musical pieces, participants would naturally forget the target sample at some point in the search, especially as they were listening to other music samples throughout the process. It was predicted that in the first stage there would be no need to listen to the target sample because participants would be able to recall it (from the initial exposure), or at least remember some of its characteristics, in order to match it to a category. In the second stage, the need to listen to the target sample before making a selection was expected, as the categories were more similar to each other and shared many characteristics. In the third stage, it was expected that participants would be able to recognize the target sample, even if they could not recall exactly how it sounded, and consequently would not need to listen to it before making a selection.

As anticipated, very little time was spent on the target sample during the first stage in the first experiment (*M* = 0.20 seconds). However, the second experiment showed a greater need to listen to the target sample earlier in the search (*M* = 1.27 seconds in the first stage). This can be attributed to the fact that the second experiment's participants had to listen to six category samples in each stage, which took longer. The two additional categories made the participants more hesitant to choose the best match without listening to the target sample another time. As predicted, in both experiments participants spent more time listening to the target sample in the second stage than in the first. Curiously, the third stage had participants listening to the target sample just before selecting the correct sound and completing the search. It seems that they recognized the target sample but wanted to confirm their choice before making a selection. This most likely happened because many samples sounded very similar on the third stage.

A survey was given to the participants after they completed their experiment's search tasks. Overall, a very positive experience with the search interface was reported. With both speaker configurations, searching serial audio output was found to be slightly easier than searching parallel output, but over 86% of participants in the first experiment and 77% in the second experiment preferred searching parallel output. Participants also felt that the parallel output allowed them to find the audio recordings faster. Most participants preferred the focus technique rather than the reduction technique for parallel output, since the reduction technique was found to be more difficult to use than the focus one.

Effect sizes were medium to large for most of the statistically significant results. Both experiments were very similar in terms of effect sizes. Parallel output produced on average 25% faster searches than serial output for both experiments. This suggests that there was no decline in performance when more audio recordings were presented simultaneously (experiment 2), which leads to the belief that users may be able to handle more than six simultaneous recordings when searching parallel output.

As a recommendation for implementing search in parallel output, a combination of both the focus and reduction techniques should be made available. Users will have the option to remove some incorrect choices from the search set and then focus on the remaining options, if needed, before selecting the correct one. Another scenario is when users focus on one option, decide it is not the correct one, and then remove it from the set, so that it will not interfere with the remaining options. Having both techniques available will make it more practical to increase the number of audio samples presented in parallel in each stage.

6.2 Limitations and Future Work

This study has some limitations that could be overcome in future studies to make the results more likely to generalize to broader samples. One limitation is that no demographic information was collected on the participants. The experimenter observed participants of various ages and technical skill levels. Individual differences should not affect the results of each within-subjects experiment, but if collected and analyzed, that information could show certain groups having a preference for search mode (*Serial*, *Parallel* + *Focus*, *Parallel* + *Reduction*), or a significantly better performance in the four-speaker configuration over the six-speaker configuration, for example. It will be interesting to collect demographic data in future studies and look for any differences in performance.

The time commitment required of the participants was another limitation. In order to keep each experimental session to at most one hour, only a few search tasks per condition were conducted. Ideally, both experiments from this study would be combined into one factorial repeated-measures experiment with *audio output* and *speaker* *configuration* as independent variables. This would allow both speaker configurations to be used and compared by the same participants.

The audio recordings used in the experiments were unfamiliar songs. It is believed that the benefits of parallel output will be even greater for other types of audio, since music search in parallel output is considered a more complex task than a search for sound effects or speech. However, further studies should examine whether the recent results generalize to other types of audio. It is also believed that the use of familiar songs would increase the benefits of parallel presentation. A study that uses the participants' own music collections, pre-categorized by an automatic genre classification, such as that proposed in Tzanetakis and Cook (2002) could reveal potential benefits to parallel presentation, even with more than six simultaneously presented recordings.

Finally, the Parallel Audio Player is a prototype created to satisfy the specific requirements of the designed experiments used in this study. However, it can be modified and used in several applications where simultaneous presentation of sounds can be beneficial, including:

- future studies in audio browsing;
- memory studies;
- obtaining relevance feedback for recommendation systems;
- sound identification applications, such as finding the type of noise a car is making before calling the mechanic, or identifying noises in search and rescue situations;
- monitoring multiple sound sources simultaneously;
- applications to use while driving, such as browsing several radio stations simultaneously, for a quicker and more informed selection of a pleasing station;
- user interfaces for the visually impaired.

References

Amazon.com Inc. (2011). *Amazon.com.* Retrieved from http://www.amazon.com

Apple Inc. (2011). *Apple - iPod - compare iPod models and find the right one for you.* Retrieved from http://www.apple.com/ipod/compare-ipod-models

Baeza Yates, R., & Ribeiro, B. d. A. N. (1999). *Modern information retrieval*. Reading, MA: Addison-Wesley Longman.

Bausell, R. B., & Li, Y. F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences*. Cambridge, England; New York, NY: Cambridge University Press.

Bederson, B., & Shneiderman, B. (2003). *The craft of information visualization [electronic resource] : Readings and reflections*. Amsterdam, The Netherlands; London, England: Morgan Kaufmann.

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia*. Boston, MA: AP Professional.

Birmingham, W., Dannenberg, R., & Pardo, B. (2006). Query by humming with the VocalSearch system. *Communications of the ACM, 49*(8), 49-52.

Bregman, A. S. (1990). *Auditory scene analysis : The perceptual organization of sound*. Cambridge, MA: MIT Press.

Brown, L., Brewster, S., Ramloll, R., Yu, W., & Riedel, B. (2002). Browsing modes for exploring sonified line graphs. *Proceedings of the British HCI Conference, 2*, 6-9.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE, 96*(4), 668-696.

CBS Radio Inc. (2011). *Mp3.com.* Retrieved from http://www.mp3.com

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975-979.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Colburn, H., & Kulkarni, A. (2005). Models of sound localization. In R. R. Fay, & A. N. Popper (Eds.), *Sound source localization* (pp. 272-316). New York: Springer.

Comparisonics Corporation. (2011). *FindSounds - search the web for sounds.* Retrieved from http://www.findsounds.com

Dannenberg, R. B., Birmingham, W. P., Pardo, B., Hu, N., Meek, C., & Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the MUSART testbed. *Journal of the American Society for Information Science and Technology, 58*(5), 687-701.

Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *The Journal of the Acoustical Society of America, 91*(6), 3381-3390.

eMusic.com Inc. (2011). *eMusic.com.* Retrieved from http://www.emusic.com

Fernström, M., & McNamara, C. (2005). After direct manipulation---direct sonification. *ACM Transactions on Applied Perception, 2*(4), 495-499.

Ghias, A., Logan, J., Chamberlin, D., & Smith, B. C. (1995). Query by humming: Musical information retrieval in an audio database. *MULTIMEDIA '95: Proceedings of the Third ACM International Conference on Multimedia,* San Francisco, CA. pp. 231-236.

Google. (2011). *Google images.* Retrieved from http://images.google.com

Grantham, D. W. (1995). Spatial hearing and related phenomena. In B. C. J. Moore (Ed.), *Hearing* (pp. 297-345). New York, NY: Academic Press.

Johnson, B., & Shneiderman, B. (1991). Tree-maps: A space-filling approach to the visualization of hierarchical information structures. *Visualization '91: Proceedings of the Second Conference on Visualization.* San Diego, CA. pp. 284-291.

Johnston, I. D. (2009). *Measured tones : The interplay of physics and music* (3rd ed.). Boca Raton: CRC Press.

Knees, P., Schedl, M., Pohle, T., & Widmer, G. (2006). An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web. *Proceedings of the 14th Annual ACM International Conference on Multimedia.* Santa Barbara, CA. pp. 17-24.

Kuhn, G. F. (1987). Physical acoustics and measurements pertaining to directional hearing. In W. A. Yost, & G. Gourevitch (Eds.), *Directional hearing* (pp. 3-25). New York: Springer-Verlag.

Last.fm Ltd. (2011). *Last.fm.* Retrieved from http://www.last.fm

Levy, M., & Sandler, M. (2009). Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia, 11*(3), 383-395.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications, 2*(1), 1-19.

Lindman, H. R. (1974). *Analysis of variance in complex experimental designs*. San Francisco, CA: W. H. Freeman.

Little, D., Raffensperger, D., & Pardo, B. (2007). A query by humming system that learns from experience. *Proceedings of the Eighth International Conference on Music Information Retrieval*. Vienna, Austria. pp. 335-338.

McGookin, D. K., & Brewster, S. A. (2004). Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception, 1*(2), 130-155.

Melodis Corporation. (2011). *Midomi.* Retrieved from http://www.midomi.com

Microsoft. (2011). *Bing.* Retrieved from http://www.bing.com/images

Moore, B. C. J. (1989). *An introduction to the psychology of hearing* (3rd ed.). London, England; San Diego, CA: Academic Press.

MP3Gain Development Team. (2009). *MP3Gain.* Retrieved from http://mp3gain.sourceforge.net

Neumayer, R., Dittenbach, M., & Rauber, A. (2005). PlaySOM and PocketSOMPlayer, alternative interfaces to large music collections. *Proceedings of the Sixth International Conference on Music Information Retrieval.* London, England. pp.618-623.

Northwestern University Interactive Audio Lab. (2010). *Tunebot - music search.* Retrieved from http://tunebot.cs.northwestern.edu

Pachet, F., & Cazaly, D. (2000). A taxonomy of musical genres. *Proceedings of the Content-Based Multimedia Information Access Conference.* Paris, France. pp. 1238-1246.

Pampalk, E., & Goto, M. (2006). Musicrainbow: A new user interface to discover artists using audio-based similarity and web-based labeling. *Proceedings of the Seventh International Conference on Music Information Retrieval.* Victoria, Canada. pp.367-370.

Pampalk, E., Pohle, T., & Widmer, G. (2005). Dynamic playlist generation based on skipping behaviour. *Proceedings of the Sixth International Conference on Music Information Retrieval.* London, England. pp.634-637.

Pandora Media. (2011). *Pandora internet radio.* Retrieved from http://www.pandora.com/corporate

Peek, B. (2011). *WiimoteLib - .NET managed library for the nintendo wii remote.* Retrieved from http://www.brianpeek.com/page/wiimotelib.aspx

Perkins, D. (1994). Transfer of learning. In T. Husen, & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed.). Oxford, England: Pergamon.

Rovi Corporation. (2011). *All Music.* Retrieved from http://www.allmusic.com

Risi, S., Mörchen, F., Ultsch, A., & Lehwark, P. (2007). Visual mining in music collections with emergent SOM*. Proceedings of the Workshop on Self-Organizing Maps.* Bielefeld, Germany.

Russolo, L. (2001). The Art of Noise. In Umbro Apollonio (Ed.), *Futurist Manifestos*. New York, NY: Viking.

Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction, 7*(3), 353-383.

Schmandt, C., & Mullins, A. (1995). AudioStreamer: Exploiting simultaneity for listening. *CHI '95: Conference Companion on Human Factors in Computing Systems,* Denver, Colorado. pp. 218-219.

Selfridge-Field, E. (2000). What motivates a musical query? *Proceedings of the First International Conference on Music Information Retrieval.* Plymouth, MA. pp. 23-25.

Serrà, J., Gómez, E., & Herrera, P. (2010). Audio cover song identification and similarity: Background, approaches, evaluation, and beyond. In Z. Ras, & A. Wieczorkowska (Eds.), *Advances in music information retrieval* (pp. 307-332). Springer-Verlag Berlin / Heidelberg.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer-interaction* (3rd ed.). Reading, MA: Addison Wesley Longman.

Slaney, M., & White, W. (2007). Similarity based on rating data. *Proceedings of the Eighth International Conference on Music Information Retrieval,* Vienna, Austria. pp. 479-484.

Sodnik, J., Dicke, C., Tomažič, S., & Billinghurst, M. (2008). A user study of auditory versus visual interfaces for use while driving. *International Journal of Human-Computer Studies, 66*(5), 318-332.

SoundBible.com. (2011). *Free sound clips, sound bites, and sound effects.* Retrieved from http://www.SoundBible.com

SoundHound Inc. (2011). *SoundHound.* Retrieved from http://www.soundhound.com

Spence, R. (2001). *Information visualization*. Harlow, England: Addison-Wesley.

Tjondronegoro, D., & Spink, A. (2008). Web search engine multimedia functionality. *Information Processing & Management, 44*(1), 340-357.

Torrens, M., & Arcos, J. (2004). Visualizing and exploring personal music libraries. *Proceedings of the Fifth International Conference on Music Information Retrieval.* Barcelona, Spain. pp. 421-424.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293-302.

Unal, E., Narayanan, S. S., & Chew, E. (2004). A statistical approach to retrieval under user-dependent uncertainty in query-by-humming systems. *MIR '04: Proceedings of the Sixth ACM SIGMM International Workshop on Multimedia Information Retrieval,* New York, NY. pp. 113-118.

Vignoli, F., van Gulik, R., & van de Wetering, H. (2004). Mapping music in the palm of your hand, explore and discover your collection. *Proceedings of the Fifth International Conference on Music Information Retrieval.* Barcelona, Spain. pp. 409-414.

Vinciarelli, A., Suditu, N., & Pantic, M. (2009). Implicit human-centered tagging. *ICME 2009: IEEE International Conference on Multimedia and Expo.* New York, NY. pp. 1428-1431.

Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM, 49*(8), 44-48.

Warren, R. M. (2008). *Auditory perception : An analysis and synthesis* (3rd ed.). Cambridge, England; New York, NY: Cambridge University Press.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using non-individualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111-123.

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America, 85*(2), 868-878.

Xingquan Zhu, Elmagarmid, A. K., Xiangyang Xue, Lide Wu, & Catlin, A. C. (2005). InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia, 7*(4), 648-666.

Yahoo! Inc. (2011). *Yahoo! Image Search.* Retrieved from http://images.search.yahoo.com

Yost, W. A. (2007). *Fundamentals of hearing: An introduction* (5th ed.). San Diego, CA: Academic Press.

Appendices

Appendix A: Statistical Analysis of the Normalized Data

A.1 Background Information

Since the data collected for both experiments appeared to be non-normal, a log transformation was used to normalize each dependent variable, and another statistical analysis was performed. The results are similar to the ones reported in Chapter 5 and are presented in this appendix.

For the variable *Number of Steps*, 1 was added to each value before the transformation, since there were a few searches with zero steps in the parallel conditions (when participants browsed through the stages and found the target sample without needing to use the Focus or Remove buttons).

For each experiment, a repeated-measures MANOVA was used to detect any effect of the type of audio output (serial vs. parallel) on the audio search effort (measured by *search time* and *number of steps*). Significant MANOVAs were followed by separate repeated-measures ANOVAs on each of the dependent variables.

All statistical tests were conducted using $\alpha = .05$. Effect sizes are calculated following the equations described in section 5.1.1.

A.2 Experiment 1

The results of the Shapiro-Wilk test of normality for the original variables are displayed in Table A1. After the log transformation, the variables became more normally distributed as shown in Table A2.

The results of the overall MANOVA revealed a significant effect of audio output on the effort required to find the target sample, $V = 0.56$, $F(4, 140) = 13.71$, $p < .05$, using Pillai's Trace.

Table A1 – Normality test results (experiment 1)

	Shapiro-Wilk		
	Statistic	df	Sig.
Time: Serial	.86	36	.00
Time: Parallel + Focus	.89	36	.00
Time: Parallel + Reduction	.88	36	.00
Steps: Serial	.89	36	.00
Steps: Parallel + Focus	.93	36	.03
Steps: Parallel + Reduction	.86	36	.00

Table A2 – Normality test results for the log-transformed variables (experiment 1)

	Shapiro-Wilk		
	Statistic	df	Sig.
Log Time: Serial	.96	36	.29
Log Time: Parallel + Focus	.97	36	.43
Log Time: Parallel + Reduction	.96	36	.23
Log Steps: Serial	.95	36	.11
Log Steps: Parallel + Focus	.97	36	.49
Log Steps: Parallel + Reduction	.99	36	.96

A.2.1 Effect of Audio Output on Search Time

A one-way repeated-measures ANOVA was conducted on the log-transformed data. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 1.19$, $p > .05$, therefore no correction was needed. The time to find the target sample was significantly affected by the type of audio output during the search, $F(2, 70) = 9.58$, $p < .05$, $\omega^2 = .08$. Planned contrasts were used to follow-up this finding and showed that the search time was significantly higher when the audio output was serial as opposed to parallel, $F(1, 35) = 15.63$, $p < .05$, $r = .56$. However, there was no significant difference between the two parallel conditions, $F(1, 35) = 0.81$, $p > .05$.

A.2.2 Effect of Audio Output on Number of Steps

A one-way repeated-measures ANOVA was carried out. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 1.55$, $p < .05$, therefore the degrees of freedom did not require any correction.

The analysis revealed a significant effect of the form of audio output on the number of steps taken to find the target sample, $F(2, 70) = 38.89$, $p < .05$, $\omega^2 = .32$. Contrasts confirmed that the number of steps was significantly higher in the serial search condition than in the parallel conditions, $F(1, 35) = 63.11$, $p < .05$, $r = .80$, but the two types of parallel audio search were not significantly different from each other, $F(1,35) = 1.70$, $p >$ .05.

A.3 Experiment 2

The results of the Shapiro-Wilk test of normality for the original variables are displayed in Table A3. After the log transformation, the variables became more normally distributed as shown in Table A4.

A repeated-measures MANOVA was conducted and revealed, using Pillai's Trace, a significant effect of audio output on the effort required to find the target sample, $V = 0.51$, $F(4, 68) = 5.88$, $p < .05$.

A.3.1 Effect of Audio Output on Search Time

A one-way repeated-measures ANOVA was performed on the log-transformed data. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 0.18$, $p > .05$, therefore no correction was needed.

Appendix A (Continued)

The results show that the search time was significantly affected by the type of audio output during the search, $F(2, 34) = 3.95$, $p < .05$, $\omega^2 = .06$. Planned contrasts revealed that the serial audio presentation produced significantly longer search times compared to the parallel conditions, $F(1, 17) = 6.51$, $p < .05$, $r = .53$. The two parallel conditions were not significantly different from each other, $F(1, 17) = 0.84$, $p > .05$.

Table A3 – Normality test results (experiment 2)

	Shapiro-Wilk		
	Statistic	df	Sig.
Time: Serial	.92	18	.12
Time: Parallel + Focus	.93	18	.18
Time: Parallel + Reduction	.96	18	.65
Steps: Serial	.78	18	.00
Steps: Parallel + Focus	.94	18	.32
Steps: Parallel + Reduction	.88	18	.02

Table A4 – Normality test results for the log-transformed variables (experiment 2)

	Shapiro-Wilk		
	Statistic	df	Sig.
Log Time: Serial	.93	18	.19
Log Time: Parallel + Focus	.99	18	1.00
Log Time: Parallel + Reduction	.96	18	.57
Log Steps: Serial	.90	18	.05
Log Steps: Parallel + Focus	.97	18	.74
Log Steps: Parallel + Reduction	.98	18	.95

A.3.2 Effect of Audio Output on Number of Steps

A one-way repeated-measures ANOVA was conducted. Mauchly's test indicated that the assumption of sphericity was met, $\chi^2(2) = 3.89$, $p > .05$, therefore no correction was needed.

The analysis confirms a significant effect of the form of audio output on the number of steps taken to find the target sample, $F(2, 34) = 16.19$, $p < .05$, $\omega^2 = .33$. Contrasts revealed a significantly larger number of steps when the audio output was serial as opposed to parallel, $F(1, 17) = 52.49$, $p < .05$, $r = .87$, but there was no significant difference between the two parallel conditions, $F(1,17) = 2.84$, $p > .05$.

Appendix B: Statistical Analysis of Order and Sequence Effects

B.1 Background Information

In both experiments, each participant performed nine search tasks, being three in each of the audio output conditions. The order in which the conditions were presented varied from participant to participant following a counterbalanced design that balances out order and sequence effects. There were six sequences of conditions, as explained in section 4.3.

A mixed one-way MANOVA was conducted for each experiment to detect if any sequence was significantly different from the others. The type of audio output was the within-subjects factor, with three levels: *Serial Output*, *Parallel Output* + *Focus*, and *Parallel Output* + *Reduction*. The sequence number was the between-subjects factor, with six levels. The two dependent variables were the *search time* and the *number of steps*.

A one-way repeated-measures MANOVA was conducted for each experiment to detect any practice or fatigue effects, for example, if the last trial was significantly faster (or slower) than the first trial, regardless of the type of audio output and the sequence they were presented. The trial number was the within-subjects factor, with nine levels. The total *search time* and *number of steps* were the dependent variables.

All statistical tests were conducted using α = .05 and all results for multivariate tests are reported using Pillai's Trace.

B.2 Experiment 1

B.2.1 Sequence Effects

There were 36 participants in this experiment and six different sequences, therefore six participants per sequence. The results of the mixed MANOVA revealed no significant differences between the sequences, $V = 0.39$, $F(10, 60) = 1.45$, $p > .05$, meaning that the search effort, when we ignore the type of audio output, was not different for each sequence. The interaction between the types of audio output and sequences was also non-significant, $V = 0.70$, $F(20, 120) = 1.28$, $p > .05$, suggesting that the search effort for each type of audio output was not affected by the sequence in which it appeared.

B.2.2 Practice and Fatigue Effects

The one-way repeated-measures MANOVA used all the nine measures for *search time* and nine measures for *number of steps* per participant. Eleven participants had at least one timeout, which is a search that was incomplete because the participant could not find the target sample within the time limit of 240 seconds, resulting in missing values in the data. When the analysis ignored the missing values, data from the 25 participants without timeouts were used and the results showed that the *search time* and *number of steps* were not significantly different between the nine trials, $V = 0.66$, $F(16, 9) = 1.09$, $p > .05$.

When the missing values were replaced with the maximum values of 240 seconds and 43 steps, the results were also non-significant, $V = 0.40$, $F(16, 20) = 0.82$, $p > .05$, indicating the absence of significant practice or fatigue effects.

B.3 Experiment 2

B.3.1 Sequence Effects

There were 18 participants in the second experiment and six different sequences, therefore three participants per sequence. The results of the mixed MANOVA revealed no significant differences between the sequences, $V = 0.48$, $F(10, 24) = 0.75$, $p > .05$, meaning that the search effort, when we ignore the type of audio output, was not different for each sequence. The interaction between the types of audio output and sequences was also non-significant, $V = 1.29$, $F(20, 48) = 1.15$, $p > .05$, suggesting that the search effort for each type of audio output was not affected by the sequence in which it appeared.

B.3.2 Practice and Fatigue Effects

The one-way repeated-measures MANOVA used all the nine measures for *search time* and nine measures for *number of steps* per participant. Four participants had at least one timeout, which is a search that was incomplete because the participant could not find the target sample within the time limit of 300 seconds, resulting in missing values in the data. When the analysis ignored the missing values, data from the 14 participants that did not have timeouts were used and the multivariate test statistics could not be produced because of insufficient residual degrees of freedom. Univariate results showed that the search time and number of steps were not significantly different between the nine trials, $F(5.28, 68.64) = 0.53$, $p > .05$ (*search time*), and $F(3.59, 46.71) = 0.77$, $p > .05$ (*number of steps*). The degrees of freedom were corrected using the Huynh-Feldt estimates of sphericity, $\varepsilon = .66$ (*search time*) and $\varepsilon = .45$ (*number of steps*), because the assumption of sphericity was violated, $\chi^2(35) = 72.26$, $p < .05$ (*search time*) and $\chi^2(35) = 110.50$, $p < .05$ (*number of steps*).

When the missing values were replaced with the maximum values of 300 seconds and 50 steps, the multivariate results were non-significant, $V = 0.91$, $F(16, 2) = 1.20$, $p > .05$, indicating the absence of significant practice or fatigue effects.

About the Author

Isabela C.R.M. Hidalgo is a doctoral candidate in Computer Science and Engineering at the University of South Florida. She attended the Pontifícia Universidade Católica in Rio de Janeiro, Brazil, where she received a Bachelor's Degree in Information Technology and a Master's Degree in Computer Science. Before moving to Florida, Mrs. Hidalgo has worked as a software developer in Cupertino, California and Rio de Janeiro, Brazil. Her main research area is Human-Computer Interaction, with a particular interest in user interfaces for audio search and discovery.