Joseph Bender

A Sonic Revolution

1. Introduction

This generation will forever be known as the one that experienced the internet explosion. Since the advent of computers and the internet, the new technology has been fully integrated into every aspect of society and daily life. Humans embraced the internet in the most brilliant way, and it now acts as a hub for utility and creativity. There have also always been traditional ways to interact with a computer. The screen is used to display visual information. Eyes absorb light waves and convert them to information in the brain. A mouse and keyboard are used to interact spatially and tactically with the graphical user interface. Using the sense of touch, an input device is manipulated and allows for the operation of a cursor. Lastly, there is a critical hardware component that can be easily forgotten. Speakers and microphones enable sound to be a fundamental factor in the way people use computers and the internet. Often overlooked, digital sound has evolved and contributed every step of the way in this recent technological revolution. It adds an entirely other dimension to human information processing. The innumerable applications for audio on the web have already begun to be explored, and will only improve in the future. The web is an unfathomably large community and contains countless files filled with sonic data. Having an understanding of how to manage and use this information will be paramount as the internet moves forward. Activities that would long ago be considered science fiction are now accessible to the everyday consumer. Web conferencing with real-time sound requires extremely fast processing power and a solid network connection. Online gaming implements realistic spatial sound to immerse one in a virtual environment. Streaming music gives users a massive collection of music instantly at a high quality. All of these applications not only require sound quality, but also an internet connection to transmit and receive the audio information. In most file types, quality is directly proportional to volume. The higher fidelity a sound is, the more memory it will take up on a hard drive disc or server. Much of the effort of sound innovation has been towards finding a balance that will work in most situations. In this paper, research is proposed that would help to further identify the most efficient way of listening to audio on the web.

In this **introduction (1)**, a brief synopsis was given describing the recent significance of audio as the internet as evolved. Speakers, sound, and music have been used in applications with every advancement of the computer. Any development or progress in the industry must keep web audio utilization as a priority. The **background (2)** begins to elaborate on previous research done in the field. The methods, materials, participant selection, experimental procedure, and results of these reports will all be summarized. Testing and examination of sound is not a new idea, but there is always room for expansion when studies become out of date due to new technology. The background section will also identify the shortcomings in past research, and how those errors affect the results. In the third section, proposed **experiments (3)**, a hypothetical experiment will be outlined. The technique will be prepared and described in detail. Sample scenarios will be formulated that will convey how the test could have gone in reality. Screenshots of a mock-up software used in the experiment will be illustrated to give an indication of what the interface may

look like (*Figure 3, p. 7*). As with prior studies, the methods and subject recruitment will be conceptualized. This experiment will expand on the background research and address an unresolved problem in relation to audio on the internet. The **expected results (4)** are speculated in section four and will attempt to explain the outcomes of the tests. They will be compared to the hypothesis to demonstrate the value of performing this particular experiment. Finally, in section five is listed the **references in APA format (5)**. It is important to give credit to the individuals who had already begun to investigate web sound in the past. Citations throughout the paper will be referenced to these historical studies.

2. Background

As mentioned before, sound has not been left behind in the technological renaissance. Computer and the internet evolved and so did audio. Quality and capacity of data is an integral part of human interaction with a workstation. Research has been conducted every step of the way to perpetually analyze and improve the present functionality.

As far back as 1979, researchers have tested the way different types of sound come out of an array of output devices. In particular, two Swedish men wanted to see how music, speech, and noise was perceived when emitted from a speaker, headphones, or a hearing aid. Hakan Sjogren and Alf Gabrielsson (1979), an audiologist and psychologist respectively, teamed up to conduct an extensive experiment and write the report *Perceived Sound Quality of Sound-reproducing Systems.* They separated subjects into three groups to listen to the sound from a speaker, headphones, and a hearing aid. These three hardware components were chosen because they represent different characteristics regarding basic audio principles, size, power, frequency response, distortion, and more. Tape-recorded sections of music, speech, and various noise from daily life were played back for 30 seconds. Each segment was “as homogenous as possible within itself with regard to sound level” (Gabrielsson & Sjogren, 1979, p. 1019). 20-42 subjects of typical hearing ability were used for each experiment, some going in groups of two or three at a time. The applicants were also from various backgrounds being either fans of high-fidelity music, musicians, or just general music listeners to give a wide array of familiarity with focused listening. The order was randomized, and the listeners were asked to judge with adjective ratings, similarity ratings, and free verbal descriptions. This provided comprehensive insight into the way subjects were experiencing the sounds. The results were variable, but presented intriguing insight into how humans feel when they hear things. The array of adjectives helped people to convey their thoughts. Words like clearness, brightness, darkness, sharpness, softness, nearness, fullness, thinness, or even feeling of space were all options when listening through the three mediums. As expected, when listening to the speaker the feeling of space increased because the soundwaves bounced around the room. With a hearing aid or headphones, the nearness factor increased and it felt like a little band inside the ear. While not providing too much hard, quantitative data, this experiment was groundbreaking in testing sonic frequencies from various mediums.

The previous study was done almost thirty years ago. Although it was advanced for its time, digital sound has come a long way from using simple adjectives to describe how music makes someone feel. When the transition from analog to digital arrived, sound compression became an integral way to keep music low capacity. Simply put, compression is attempting to represent the same set of analog information digitally, just using fewer bits. In the process, some of the original data is lost. This is not a bad thing necessarily, but it all depends on whether there is a noticeable change to the listener. The most popular compression method today is MP3, which has a default bit rate of 320 kilobits per second. This means that each second of music has 320 kilobits of data stored in it. On average, this reduces file size by 80% which is staggering when considering millions of songs stored on the internet. Since so much of the original analog data is lost, this is referred to as "lossy" compression. Conversely, using a "lossless" compression technique, very little of the data is lost. This is done with the FLAC (Free Lossless Audio Codec), and loses only 30% on the original analog data. However, it is controversial whether humans can perceive the difference between lossy and lossless compression techniques. Two Polish scientists Norbert Nowak and Wojciech Zabierowski (2011), collected some hard data comparing the two compression methods. Four systems of lossy compression and two of lossless were applied, and then compared back to the original uncompressed file derived from a compact disc. Adobe Audition 3.0 was the professional music program used for analysis. In particular, one graph of the spectrum of the acoustic signal, and the other graphing the

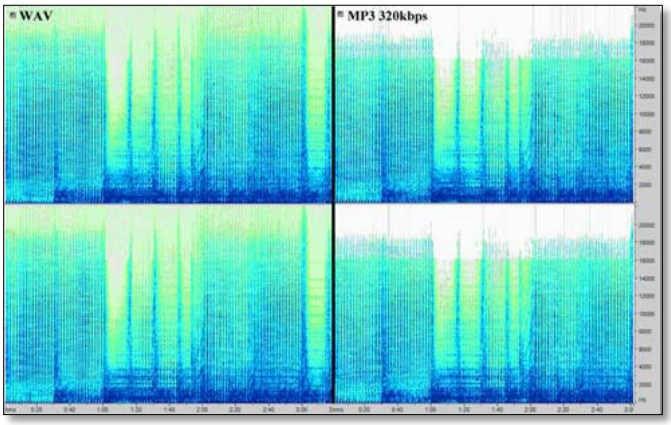

Figure 1: The spectrogram comparing lossy and lossless (Nowak & Zabierowski, 2011, p. 94)

spectrogram that is the signal amplitude spectrum diagram (*Figure 1)*. Only one song file was needed because although the size of the song can change, the rate of compression does not. The results produced were the percentages given above. The two researchers found that with lossy compression, 80% of the analog data was lost. With the more volume-consuming lossless compression method, only 30% of the original analog data was lost (Nowak & Zabierowski, 2011, p. 95). While contributing quantitative measures, this study still does not consider if the average human can tell the difference between the two methods whether they are familiar with the song or not.

Back in the day of 8-bit, side-scrolling, arcade games, all that was needed was some generic music and a handful of sound effects coming out of some speaker tucked in the back of the machine. Today, much more sound design is needed to immerse gamers in the virtual reality. There is mono sound, which comes from a single generation point. Stereo has a right and left speaker, such as headphones or computer speakers. There is even 5.1 surround sound which maps audio to six different output devices! Spatial sound is extremely important in modern technology in relation to movies, music, gaming, and many other applications. Would a gamer perform better in a first person shooter video game if all the bullet sound effects were realistically generated from the point they were fired? Is the final scene in a thriller movie better if the monsters footsteps sound like they are getting closer and closer? Spatial audio has unlimited applications in virtual environments and media viewing. In 2013, five researchers from the University of Ontario ran subjects through a simulation to see if spatial sound helped them perform in a virtual surgery. In their paper *Spatial Sound and its Effect on Visual Quality Perception and Task Performance within a Virtual Environment*, an experiment is detailed in which unpaid volunteers from the university executed a modified total knee arthroplasty game. The subjects were three males and three females of average age 21, and none reported any hearing disabilities or defects. The "surgeon" was asked to grasp a virtual drill and make an insertion. Afterwards, task completion time was recorded and they were asked to rate the fidelity (realism) of the scene they were just immersed in on a scale of 1-7 (*Figure 2)*. The variable was the sound emitted from the drill. Depending on three types of drill sound characteristics (no

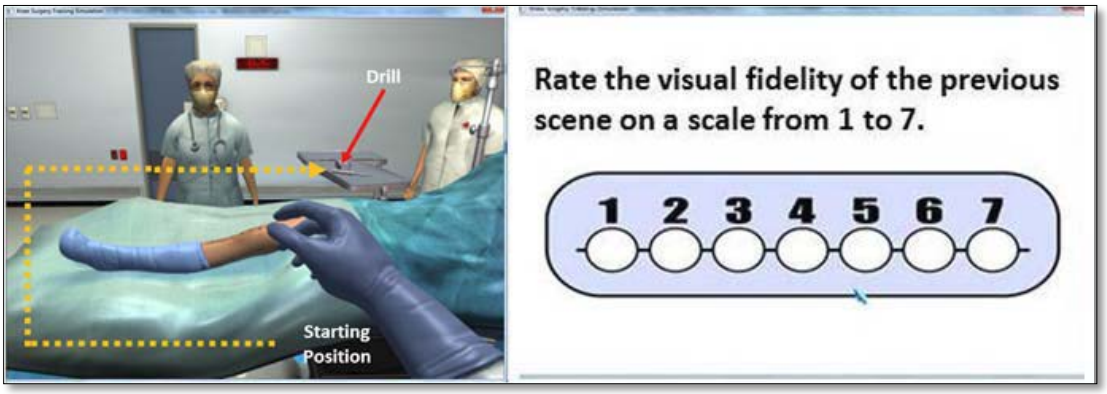

Figure 2: Virtual operating room used in experiment (Cowan, 2013, p. 4)

sound, monaural (non-spatial), and spatialized sound), the user would perceive the room and surrounding differently. The results were not significant at the $p<.05$ level for any of the conditions, which goes directly against the hypothesis that spatialized sound would make something more realistic (Cowan, 2013, p. 5). The reasons for this error will be discussed later.

In the recent years, streaming services and music in the cloud have been dominating the attention of listeners everywhere. Humans transitioned from physical mediums like vinyl and CDs, to MP3 players and iPhones full of songs. Now, the files are not even stored on the device! Apple Music, Spotify, TIDAL, and other companies house millions of songs on their servers for customers to access for a monthly subscription fee. Music storage should not be considered without bearing in mind what compression method is being used. These services offer mainly MP3 streaming at 256 or 320 kbps. That did not keep Pramod Kumar and Dinesh Goyal (2014),

two Indian researchers, from investigating other types of files that could stream from a cloud service last year. They used Windows Multipoint Server to stream a video on a network connection with three types of audio: MP3, WAV, and FLAC. With each of those audio formats they tested streaming from a cloud over a wired cloud network, cloud server, and wireless cloud network. The WMS software provided a graph showing the spectrum of the acoustic signal, and the spectrogram that is the signal amplitude spectrum diagram. These two figures were used to draw conclusions by contrasting the three compression techniques. They found FLAC is the worst in a wireless network, which should be obvious because it is the highest volume and bitrate. The performance of MP3 is also worst in the wireless network, but better than the other two compressions because it is the lowest bitrate. They concluded that FLAC is the "best audio coded available as on date for streaming, whilst WAV is the worst" (Kumar & Goyal, 2014, p. 103). This does not take into account that most consumers do not have the hardware capabilities to be able to listen to FLAC correctly. A special pair of headphones or speakers, and an amplifier, are required to boost and perceive the signal the way it was intended to be. The results of the MP3 streaming should be especially taken into account due to its predominance on the internet.

All four of these studies contributed immensely to the investigation and improvement of digital audio on the web. They were not without their errors and shortcomings however. The first study (comparing speakers, headphones, and hearing aids in 1979) was outdated and conducted when technical sound design for the web was just taking off. The sound was not transmitted over the internet with the possibility for noise or distortion. Also, subjects went in groups of two or three at a time and could have been biased by each other's responses. If a general listener was intimidated by one of the musicians, they could align their answers with them to seem knowledgeable. The perceptual adjectives were versatile, but not entirely comprehensive. Some sounds could have been described by words outside of the given selection. Although it was an interesting early paper in the field, it did not age well with the advent of the internet. The second study, in which the compression ratios of lossy and lossless compression techniques were compared, was informative but too specific. It gave only a quantitative measure of the compression ratio, but not a human's perception of quality. It proved that lossy music will sound "better" than lossless one-hundred percent of the time. It would mean nothing though if ears were not advanced enough to notice the difference. In addition, it was not streaming the music over a network connection and all playback was done locally. The simulation from 2013, in which spatial sound was tested in a virtual surgery, also overlooked some issues. It should be obvious that spatial sound in a surgery would assist in fidelity and task completion time. However, they only used six subjects and found no significant results. With a larger participant group, the data may have shown that on average, being able to hear the spatial location of the drill will help with its operation. They even claim in the conclusion that the results were "preliminary" (Cowan, 2013, p. 6). In the final and most recent research paper published only last year in 2014, three audio compression techniques were tested streaming from a cloud server. Although the produced spectrograph was informative, it was still up to the researchers to make deductions from the figures. It resulted in a lot of relative comparisons like "MP3 gives a better bit rate at low frequency than WAV", and not much real world application. Furthermore, its ultimate conclusion was that FLAC is the best audio available for streaming. Everyone knew that

already! Just because it is the highest quality does not mean it should be the industry standard on the web. Lossless audio requires an amplifier to boost the signal in order to be able to hear all of the information from the file. This is an audiophile setup that 99% of the population does not have, so MP3 remains the default compression method today.

In the next section, a hypothetical experiment will be proposed that will attempt to expand on historical research and overcome the shortcomings of the past. Whereas some of the past research was extremely technical and yielded overly-scientific results, this study will be investigating how a normal consumer would interact with the internet. The data will be a direct reflection of how humans perceive audio. Additionally, some of these studies waste time by exploring numerous compression techniques. MP3 is far and away the most used on the web, and therefore it should almost be designated as the control group of any audio investigation. The internet must also be used in the test because music is slowly being migrated onto cloud servers. The future of audio is not stored locally. With all this in mind, modern audio research must be done with consideration of the present technology and functionality.

3. Proposed Experiment

As mentioned, MP3 is the leading compression method on the internet. The standard MP3 file that anyone would listen to in iTunes or download from a blog is 320 kbps. It has been shown that this means 320 kilobits of information are in each second of the song. To most people, this is the “normal” quality of music that people are used to listening from their iPod or iPhone. However, in an attempt to reduce server space MP3 can also be compressed at 192 kbps. Massive community sites like YouTube and Soundcloud, which are typical places to hear new music, are encoded at this diminished 192 kbps bitrate. This is why people are deterred from using freeware to rip songs from the internet into their personal music library. The quality will be lower than any ripped CD or songs purchased off the web store. Is this a disservice to the artists who work hard to produce the song at the highest quality possible? The smaller the filesize, the more music that can be stored on a server. This allows these websites to house a larger variety of music instead of a smaller variety at a higher quality. This experiment will compare the quality of 192 kbps sound to 320 kbps sound streamed over the internet and hypothesizes that humans can indeed perceive a slight noticeable difference in quality between the two bitrates.

Methods:

The test will be conducted in a university lab, and the subjects will be selected from research volunteers from the school. This will produce participants that are academically inclined, but not trained audiophiles. 100 randomly selected subjects will be led into a room with a computer terminal and headphones by the lab assistant. Headphones were selected because they produce binaural stereo audio and are a good middle-ground between mono and surround sound. When the subject is ready they will put the headphones on and click a button to begin the test. The first slide will be a quick description of the two levels of quality, and instructions for completing the experiment. A series of ten screens will be displayed. Each will contain a listening test for a twenty second snippet of a single song, randomized for each test. Twenty seconds will be enough to absorb the song, but not long enough to make the test fatiguing to the subject. There will

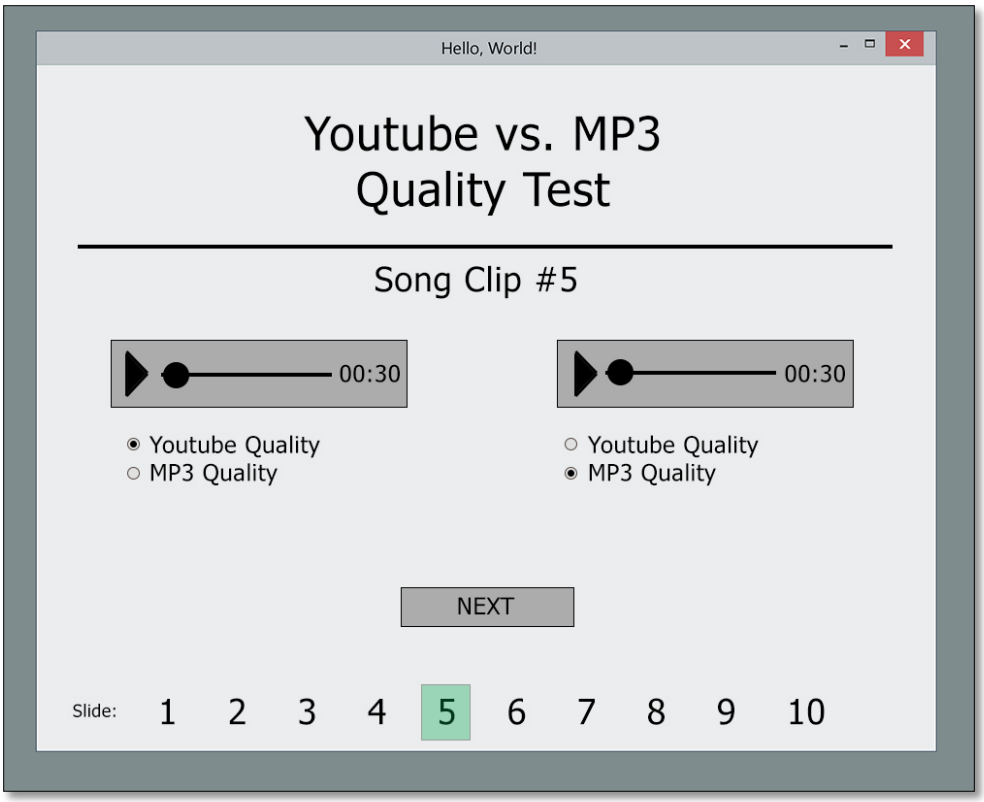

Figure 3: Mockup of graphical user interface

be two songs each from five separate genres: electronic, classical, rock, country, and spoken word. Having an assortment of genres helps eliminate bias that could arise from songs within genres being similar. On each screen will be two content players with a play button to begin audio playback. Underneath will be two radio buttons labeled "YouTube Quality" and "MP3 Quality." This phrasing is used because most people will recognize this terminology and understand the difference between these two levels of fidelity. Most of the public has listened to

a 192 kbps song on YouTube and understand that it is diminished quality, and also has an MP3 player which plays 320 kbps files. On the bottom will be an indicator of where the participant is in the ten page test. This interface and all the audio files will also be hosted on a remote server and accessed through a web browser. This is because most of the public accesses music through a streaming service, and audio compression must take this into account. In addition, the computer terminal will have an average network connection and central processing unit to eliminate hardware bias.

Stimuli:

On each slide, the subject will have to listen to both clips and attempt to discern and label correctly which is the higher quality 320 kbps "MP3" file, and which is the lower quality 192 kbps "YouTube" file. They do this by listening to either clip as many times as they want, allowing for close direct juxtaposition of the two compression methods. If one is playing and the play button on the other is clicked, the one currently playing pauses so the other can be heard. The subject may not go back to pages once they have locked in an answer and hit "NEXT", and the button cannot be clicked if a label has not been selected for each snippet. On the last slide will be a "Complete Test" button, and once pressed the test is over. No more answers may be given, and the lab assistant will be informed and will enter to escort the participant from the room. Each snippet is twenty seconds. Anticipating that the user will go back and forth between clips twice to really absorb the music, each slide will have 80 seconds of music playback. Adding 20 seconds for clicking and 20 seconds for making a decision, that will bring each slide to 120 seconds. Multiplying by the 10 slides in the test, the total experiment time comes out as 20 minutes per participant.

4. Expected Results

The most important result will be the total number out of ten correct answers given by each person. Their score is directly proportional to the amount that humans can perceive of the difference between 192 and 320 kbps audio. If there is no perceptual difference, then the number correct will be similar to that of a completely random distribution. Since there are two options, the expected value would be five labels correct based purely on statistical averages. The hypothesis presented predicted that human ears do have the ability to discern between the 192 kbps and 320 kbps bitrate. If this is to be proven valid, the number correct would have to be closer to ten. If a subject answered ten correctly, this would indicate that they can easily tell the difference between the two. Even if the score was seven, eight, or nine, it would still indicate an above-random probability that there is a perceptual difference. In addition, the researchers could see whether any specific genres tended to have an above average number of correct responses. This would indicate that quality mattered more in the perception of that particular genre. In general, if anyone were to hear a 192 kbps file without any context whatsoever it might be difficult to tell what compression technique was used. However, this experiment gives the

subject the ability to listen to the two qualities back to back. This should allow a human to tell the difference and correctly label one as 192 kbps bitrate and the other as 320 kbps bitrate. The fact that an average CPU and network connection were used will give this research validity in a basic consumer setting such as a household or school. The study would be comparing two different bitrates of MP3 which is the most universally used at the moment. This makes the results relevant and applicable in software development today unlike some of the historical studies. The most important factor of this hypothetical tests' results is that they directly reflect real-life human necessity. Listening to and downloading 192 kbps and 320 kbps MP3s is something almost every internet user experiences. Proving a perceptual difference in quality between the two could be the first step in disputing the widespread use of the diminished 192 kbps bitrate. So much new music is first released to Soundcloud and YouTube channels. However, every one of these streams is first compressed to 320 kbps with the 80% data loss as proved by the lossy research study in 2011 (Nowak & Zabierowski, 2011, p. 94). Then, it is compressed even more to the 192 kbps bitrate, resulting in a 90% data loss. Is it fair that this music, that artists labored long hours to create, should be premiered at an inferior lossy quality? MP3 may be the prevailing compression method today on the internet, but it must remain at the 320 kbps bitrate in order to preserve the musical quality that the artist intended.

5. References in APA Format

Collins, K., Cowan, B., Dubrowski, A., Kapralos, B., Rojas, D. (2013). Spatial sound and its effect on visual quality perception and task performance within a virtual environment. The Journal of the Acoustical Society of America, 133(5), 1-7.

Gabrielsson, A., & Sjogren, H. (1973). Perceived sound quality of sound-reproducing systems. The Journal of the Acoustical Society of America, 65(4), 1019-1033.

Goyal, D., & Kumar, P. (2014). Performance analysis for audio streaming in cloud. IOSR Journal of Computer Engineering, 16(5), 98-104.

Nowak, N., & Zabierowski, W. (2011). Methods of Sound Data Compression – Comparison of Different Standards. Journal of Rating and Investment Information, (4), 92-92.