Multichannel Sound Perception and Learning

An Interactive Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

2017

Submitted By:

Emmit Joyal

Christopher Madu

Jake Merdich

Nicholas Paganetti

Approved by:

Professor Frederick Bianchi, Department of Music, Project Advisor

Abstract

The aim of this paper is to introduce the natural abilities that humans have in the perception of multichannel audio environments, and explore ways that they can be utilized for comprehension and learning purposes. Background information about audio perception as a whole is explored to set up the foundation for this topic, and the potential increase in comprehension is demonstrated through two experiments related to multichannel audio. Recommendations for the implementation of these systems based on existing technologies and possibilities in the future are also explained to promote further research and development in the area.

Acknowledgements

We would like to thank professor Frederick Bianchi in aiding our efforts in completing this project. Without his insight and guidance the following project would not have come to fruition.

Executive Summary

The way that humans process information through sound is a very complex topic, and is something that many believe has potential to be improved. A large part of this topic is the way that humans learn through sound, and the effectiveness of audio input in a learning environment. Hearing information is useful when learning, and often can be used to enhance the connections made from just reading plain text. Educational videos, lectures, and simple group or individual discussion are all ways that sound had been used to foster learning, but most methods used do not take advantage of a certain property of the way that humans process information. This property is the ability of a human to take in and understand several sources of audio input at once. The methods used now to provide information are mostly focused on a single stream of audio input. Abilities such as these are topics that are still not fully understood, which leaves room for several improvements to be made. This paper lays out the background for these ideas and demonstrates their effectiveness in real experiments, while providing a discussion on the potential for implementation in comprehensive and learning environments.

In "Auditory Scene Analysis", Albert Bregman covers a vast amount of information on auditory perception, how it relates to different kinds of comprehension abilities, and the underlying functionality of human understanding when it comes to the topic. He highlights several important points about the way audio information is processed compared to visual information, especially in the case of distinguishing sources of information. In the case of visual perception, this topic has been concretely studied and analyzed, mainly because the distinguishing process is more intuitive. But the audio side of perception analysis has far more unexplored areas and opportunities for advancement that have yet to be discovered. One popular example of these abilities is the Cocktail Party Effect. During a cocktail party, there are several different conversations taking place between the attendees of the party, and the structure of the event means that these conversations are heard from several directions. When not focusing on any conversation, an observer would hear a very large amount of voices at the same time, which would be processed as one large input of people talking over each other. However, if the attendee focuses on one conversation in any direction from where they are standing, they are able to clearly process what is being said while the noise from the other conversations is still present. Like many of the examples Bregman presents, the Cocktail Party Effect raises many questions about how and why audio information can be processed that way, and whether or not it is something that can be taken advantage of in improving comprehension.

In an experiment done by the Air Force Research Laboratory, the effectiveness of a multichannel audio environment in comprehension was tested based on several real applications of existing multichannel environments. The experiment used 3 different audio setups: Monaural, Dichotic, and 3D audio. In the monaural setup, each source of audio was projected in a standard mono way, with no extra separation for the listener. The Dichotic configuration is similar, but also has two sources from each ear and three sources similar to the Monaural setup, giving the typical experience found in a stereo setup with 2 speakers. Lastly, the 3D audio display uses Head-Related Transfer Function to simulate several sources of input that are placed separately around the listener. The results of the experiment suggest several important themes that Bregman outlined regarding the comprehension of auditory scenes. In each audio setup of the experiment (Monaural, Dichotic, and 3D), as the distinction between each individual source of sound increases, the level of comprehension in the listener increases. The experiment was done in an attempt to find improvements to the existing communication systems that use the simple monaural setup, which raises the questions of how existing methods in any learning environment can be improved or replaced by more advanced multi channel setups.

Although the benefits of moving away from simple stereo and mono sound systems have not been widely implemented, some corporations have been using a multichannel approach to existing sound systems to add another layer to the user experience. A significant example of one of these companies is Dolby, who have recently introduced and been developing an audio setup system known as Dolby Atmos. The basic idea of the Dolby Atmos sound system is to simulate audio that can't be traced back to a speaker source, but instead is heard naturally by the user as if they were living the desired experience. To achieve this, the Atmos system uses as many as 7 to 34 different speakers strategically placed in the room, based on each speaker's functionality. The Dolby Atmos system provides an enhanced experience for movie viewers in a theater, but a large part of the potential of the system comes from the home speaker setup that can be used in any reasonable area.

The idea behind the home setup is important in the potential profit of the Atmos system, but also opens up several possibilities about using similar setups in learning and comprehension environments. Similar to the situation described under the lecture section of the paper, the Atmos home setup is a prime example of something that can be used in libraries, schools, and homes with supported and tested audio formats. Instead of educators needing a dedicated space the likes of a movie theater to use the multichannel learning environment, a similar setup to the home Atmos system can be used in any traditional classroom. The potential benefits of this are clear to see: saving money from having to construct a dedicated multi channel sound room, the ability to have several setups working at once, and simple integration with other specialized learning rooms such as labs are a few examples.

Improvements in comprehension in any environment are extremely beneficial, and any opportunity to create them should be given a high priority. In the past the opportunity that comes with multichannel audio environments may have been too expensive or not known well enough to be used, but with the new technologies and information set forth in the field the opportunity is a lot more concrete. With the continuation of research in the area, the addition of a multichannel audio component can become a staple component of learning environments and add a valuable method of improving comprehension.

Table of Contents

Abstract 1

Acknowledgements 2

Executive Summary 3

Table of Contents 7

List of Figures 8

List of Tables 9

Chapter 1: Introduction 10

Chapter 2: Background 12

2.1. Auditory Scene Analysis 12

2.2. The Cocktail Party Effect 17

2.3. Psychoacoustics Principles of Multi-channel Sound 18

2.4. Head Related Transfer Function (HRTF): 21

2.5. In Changing Techniques in Practical Surround Audio: 24

Chapter 3: Experiments 26

3.1. Air Force Research Laboratory Experiment 26

3.2. WPI Experiment 29

Chapter 4: Discussion 34

4.1. Effectiveness of Existing Learning Environments 34

4.2. Potential of multichannel technologies (Dolby Atmos) 37

Chapter 5: Conclusion 40

Appendix 42

List of Figures

2.1 Example of the advantage of separation in a visual scene..13

2.2 A line drawing of blocks for visual scene analysis……………………………………………..14

2.3 A spectrogram of a mixture of sounds containing the word “shoe”.......................................16

3.1 Results of the Air Force Research Laboratory Experiment...28

4.1 An example setup of the Dolby Atmos home sound system..38

List of Tables

A.1 Trial 1 (Stereo) Word Reference…………………………………………………………….42

A.2 Trial 2 (Stereo) Word Reference…………………………………………………………….43

A.3 Trial 3 (Surround) Word Reference…………………………………………………………44

A.4 Trial 4 (Surround) Word Reference………………………………………………………....45

Chapter 1: Introduction

The way that humans process information through sound is a very complex topic, and is something that many believe has potential to be improved. A large part of this topic is the way that humans learn through sound, and the effectiveness of audio input in a learning environment. Hearing information is useful when learning, and often can be used to enhance the connections made from just reading plain text. Educational videos, lectures, and simple group or individual discussion are all ways that sound had been used to foster learning, but most methods used do not take advantage of a certain property of the way that humans process information. This property is the ability of a human to take in and understand several sources of audio input at once. The methods used now to provide information are mostly focused on a single stream of audio input. In lectures there is only the sound of the lecturer being processed, and in discussions there is focus on a single topic or piece of information at once. These all have educational merit for their simplicity, but the possibility of improving on these methods by opening up several input channels at once is something to be considered.

The creation of a virtual acoustic environment is possible with the existence of multi-channel sound systems, which represent a certain number of speakers that are installed around the listener in a certain manner. However, it should be taken into account that the human perception of sound reproduced by speakers differs from the perception of sound of natural origin. Natural sound source is not divided into components, it comes from a single place. However, while overcoming distance and obstacles in the propagation path, it becomes three-dimensional. In two-channel stereo setups there are two sources, and in Surround Sound there can be several more. The formation of a single sound image from multiple speakers then appears due to the psychoacoustic property of the human auditory system to perceive the phantom image between two coherent sound signals, which are arriving from different directions. (Howard 89)

The frequency changes in this phantom image compared to the original tone increase with the arrival of the signal from the speakers, which are different from the front. Although slight tonal changes are observed in the phantom image between the front speakers, the difference between the rear speakers has significant changes in tone by ear. Thus, an attempt to recreate the natural sound source via an increasing number of speakers will inevitably lead to some changes in the natural source of sound, but makes it more and more detailed for recreating the illusion of almost any area of human perception. The change of localization during natural sound listening is connected with changes in its frequency components. These frequency changes of sound are partially subtracted on the step of processing the signals, which come to the brain. As a result, a person perceives and identifies a “source of tone”, which remains unchanged when the angle of arrival is changed.

Abilities such as these are topics that are still not fully understood, which leaves room for several improvements to be made. This paper lays out the background for these ideas and demonstrates their effectiveness in real experiments, while providing a discussion on the potential for implementation in comprehensive and learning environments.

Chapter 2: Background

2.1. Auditory Scene Analysis

In “Auditory Scene Analysis”, Albert Bregman covers a vast amount of information on auditory perception, how it relates to different kinds of comprehension abilities, and the underlying functionality of human understanding when it comes to the topic. He highlights several important points about the way audio information is processed compared to visual information, especially in the case of distinguishing sources of information. In the case of visual perception, this topic has been concretely studied and analyzed, mainly because the distinguishing process is more intuitive. But the audio side of perception analysis has far more unexplored areas and opportunities for advancement that have yet to be discovered.

As an introduction, Bregman introduces an important concept used as a basis for his research known as a perception “scene”. In a basic sense, the purpose of human perception is to take information from one’s surroundings and represent it in a meaningful way. This process can then be broken down into a “two-part system: one part forms the representations and another uses them to do such things as calculate appropriate plans and actions” (Bregman, 3). The ability to perform the first of these two steps becomes complicated when presented with an environment where some inputs are important for comprehension while others are not. To go further, once these inputs are seen as important, there is still the task of organizing and combining the correct combinations of inputs to understand what is really happening. It is this environment which constitutes the perception scene, and provides the basis of scene analysis.

To demonstrate this idea, Bregman includes several audio and visual examples. When faced with a scrambled sequence of letters like in the image below, the human brain may not be able to distinguish any meaning from what they see, but certain properties of visual perception can be taken advantage of to give meaning to the letters.

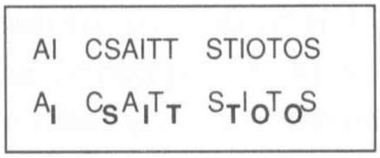

Figure 1.1
Top line: a string of letters that makes no sense because it is a mixture of two messages. Bottom line: the component messages are segregated by visual factors. (From Bregman 1981b.)

Figure 2.1 Example of the advantage of separation in a visual scene (Bregman, 4)

In this case, the step would include rearranging letters to form words, or shifting some letters up and down to form clear sentences. In other related visual examples, things like depth, relative position, and clearly defined labels also help to make up the more concrete package of visual perception. The example given by Bregman is that of the visual scene, where a human or a computer is given an image such as the one below and asked to distinguish each item in the image from the other.

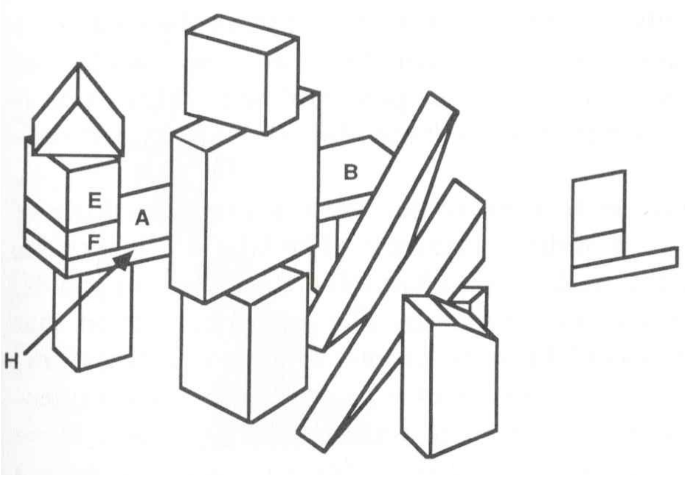

Figure 2.2 A line drawing of blocks for visual scene analysis (Bregman, 5)

In visual cases like these, the extra information provided with the image allows for a much simpler analysis of the scene. The sections labeled A and B are separated by lines in the original image, but because of human's ability to conceptualize objects in 3d space, the block labeled with A and B is seen as sitting behind the tower of blocks in front of it.

The comprehension from the audio side of perception may not be as clear cut, but it certainly does exist, with Bregman outlining the following example:

> Take the case of a baby being spoken to by her mother. The baby starts to imitate her mother's voice. However, she does not insert into the imitation the squeaks of her cradle that have been occurring at the same time. Why not? A physical record of what she has heard would include them. Somehow she has been able to reject the squeak as not being part of the perceptual "object" formed by her mother's voice (Bregman, 5).

The above scenario demonstrates one of the relatively unexplored areas of audio perception. The entire auditory scene is taken in by the baby, but somehow it is able to distinguish important parts and cut out unimportant parts based on the audio alone. This is not only seen when recognizing a person talking, but in several real life scenarios as well, such as the cocktail party effect also described further in the paper.

Another important distinction between audio and visual perception is the relationship between objects and streams. In the process of forming information from vision, light from one or many different sources bounces and reflects off of several objects that are in the world. What humans see is the light that is reflected off of these objects and into the human eye. Using this information, there are several things that humans then can comprehend about the object, like distance, size, shape, and maybe most importantly when it comes to scene analysis, a clear distinction from other objects in the three dimensional environment.

This process is very different in auditory perception, which is based around the idea of streams. The human ear takes in information in a simpler way than the eye-each eardrum vibrates in turn with the vibrations of its surroundings, which are interpreted by the brain as sound. But in this case, there are no apparent significant extra attributes of the vibrations that the eardrum can take advantage of. Instead, the audio information is processed in a stream. This concept is the basis of spectrograms, which display a picture of a certain sound based on the frequency at each moment in the audio stream. The figure below represents a spectrogram reading of the word “shoe” said with conflicting background noise.

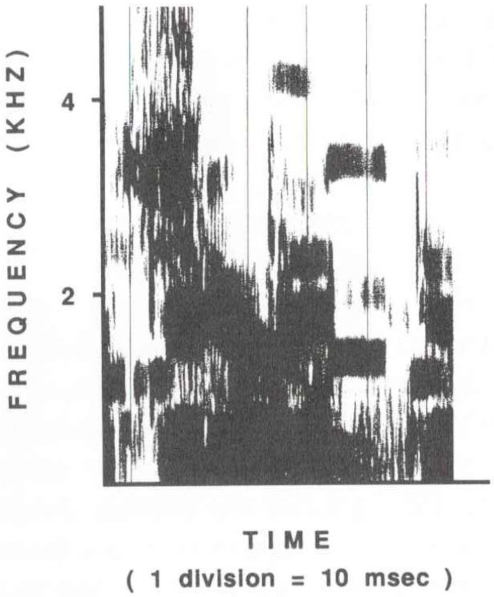

Figure 2.3 A spectrogram of a mixture of sounds containing the word "shoe" (Bregman, 8)

It is this "picture" that the human ear takes in and processes. This could cause some skepticism as to how much information can truly be extracted, but many of the same questions asked about visual perception can be answered in this way. How many people are talking at one time? Is one closer than the other? Are there any background noises that are distinctly separated from a human voice, or is the voice muffled or distorted in some way? The way humans can extract all of this information from the audio stream alone is a vital question in scene analysis.

One part of auditory comprehension that can be used to combat these problems is the ability of the brain to represent audio in a spatial sense. If this can be done, then important properties like distinct sources of sound, location from the listener, and motion of the sound source can be derived from the base auditory stream. In order for this to happen, the comprehension system that processes sound has to be able to be able to both pull information from a certain format and represent the information internally in the same format, known as a continuum. This continuum allows for an accurate representation of space that is used to provide additional information in the comprehension process. The physical world is in this format of a continuum, with the notion of space and the ability for objects to move from one location to another. This satisfies the first condition for spatial comprehension, that the source information is in the required continuum format, but the internal representation of that format is not guaranteed. In the case of visual perception and perception from the sense of touch, the process of defining space is more defined. This is due to a more complicated comprehension process, in which "the representations of space at the surface receptors and in the primary sensory cortex take the form of a topographical map" (Bregman, 74). With auditory perception, the audio stream that is taken in by the listener may at first seem to be too simplified to create a meaningful representation like the topographical map. However, studies have shown that the human brain has the capability to create something very similar in auditory comprehension. At higher levels of the brain and nervous system, the information that is taken in through the stream format is somehow coded topographically, without any of the extra information available in visual inputs. It is this ability that raises several questions about the potential of auditory comprehension and may offer new methods for utilizing the ability in learning and comprehension.

2.2. The Cocktail Party Effect

This ability to process information when faced with multiple channels of input is seen in an observation known as the cocktail party effect. During a cocktail party, there are several different conversations taking place between the attendees of the party, and the structure of the event means that these conversations are heard from several directions. When not focusing on any conversation, an observer would hear a very large amount of voices at the same time, which would be processed as one large input of people talking over each other. However, if the attendee focuses on one conversation in any direction from where they are standing, they are able to clearly process what is being said while the noise from the other conversations is still present. This ability is something that anyone with normal hearing capabilities is able to do, and raises several questions about how and why audio information can be processed that way. To further narrow down the explanation of this ability, Barry Arons summarizes this effect as two distinct problems. The first problem is sound recognition, namely recognizing when more than one source of audio input is present when hearing sounds. The second problem is the synthesis of certain cues in auditory perception. These cues could possibly be used to "enhance a listener's ability to separate one voice from another in an interactive speech system" (Arons). The experiment in question uses both of these principles as a starting point, and explores whether or not this cocktail party effect can be put to use in education.

2.3. Psychoacoustics Principles of Multi-channel Sound

While training a professional musician, it is not only important to bring knowledge from master to student hand to hand, but also to be able to put the student in a variety of situations in which he has to make his own decisions. These situations can be modelled by the teacher himself, and created with the use of technical means of training, resulting in an acoustic learning environment (Gulick 281). It is the basis of the virtual acoustic environment and gives the possibility to:

- simulate different listening environments, which are inaccessible for the creation of human verbal system;
- change the parameters and properties of the context, in accordance with the needs of dynamically changing conditions of formation of this context;
- simulate the response of the system to the actions of the musician, which is important when analysing the quality of the execution and evaluation of the reaction of the musician;
- perceive music in a holistic form.

However, it should be noted that the human perception of sound, which is reproduced by loudspeaker, is different from the perception of the natural origin of the sound, because it's formation in nature and in speakers are different (Werner 144).

The frequency changes in the phantom compared to the original tone increases with the arrival of the signal from the speakers, which are different than the front. Although, slight tonal changes are observed even in the phantom image between the front speakers, the difference between the rears has significant changes in tone by ear. Thus, an attempt to recreate the natural sound source via an increasing number of speakers will inevitably lead to some changes in the natural source of sound, but makes it more and more detailed for recreating the illusion of almost any area of human perception. In the development of learning systems, in which multi-channel audio is a major component, it is necessary to compensate changes in the sound associated with psychoacoustic characteristics of human perception (Larsen 88).

A change in source location during a natural auditory scene is connected with changes in its frequency components. These frequency changes of sound are partially subtracted during the step of processing the signals that the listener takes in. As a result, a person perceives and identifies a “source of tone”, which remains unchanged when the position relative to the listener is changed. It varies only in respect to the localization tools. As an example, take the scenario of a musician playing an instrument while moving around the room to different locations. Although the frequency response during the movement dramatically changes due to changes in the acoustic balance and the angle of perception to the listener, the instrument will still sound the same. If the musician were to switch instruments, the listener could easily tell the difference by their auditory perception alone. This remarkable ability, which allows one to isolate the ear from the sound changes associated with the acoustics of the room and any potential change in the source, helps to find the true source timbre (Zwicker 213). However, it works only with natural sources, and sound created by speakers is a subject of different laws. If you do not take this into account, it can be concluded that the speaker representation does not need to include this sound “correction” in cases where several audio channels are playing at once. Simply adjusting the speakers to the acoustics of the particular room would then be sufficient for recreating the full sound. In other words, the assumption is that a simple panning between speakers should lead to the same representation of sound that is heard from the physical instrument scenario, which is not always the case. To recreate this natural sound to the best quality possible, several different methods of auralization can be used.

Auralization is a recreation of some desired sound in a room using a speaker system, where the computer represented model of the sound source interacts with the room model. As a result the necessary sound is created, which is normally reproduced by a two-channel system with the exception of the interrelationship between the channels. These methods are used along with tracking the listener's position in the room to recreate the feeling of physical presence near the audio source. Many systems of this type also require a visual component as well, like in the case of a surround sound system for watching movies. An important part of the orientation of the system then becomes placing the speakers in a way where they do not interrupt the visual component. These cases have several methods to achieve this, such as placing the speakers on the corners of the screen, but this method is not always acoustically correct (Howard 163).[1] In both these scenarios and scenarios where access to auralization methods is limited, other methods have to be studied, such as Head Related Transfer Functions.

2.4. Head Related Transfer Function (HRTF):

The development of modern digital technologies is largely dictated by entertainment industries due to the popularity of the fields of cinematography, television, gaming and music. In these markets, with the exception of music, the primary interest in improving existing technologies has been related to the visual component. However, from recent times the quality of sound has become an important aspect of the overall quality of the final product. As a result, auditory technologies are now also under rapid development and improvement. The head-related transfer function is one of these popular technologies, and is a method that digitally represents an auditory scene in real space using simple earphones or a pair of stereo speakers. The purpose of the HRTF is simple, but the technology is quite complicated and requires a deep understanding of physics and anatomy.

The purpose of the head-related transfer function is to accurately represent the process of how a person's ear receives sound from a particular place: both how the ears locate the space and the other factors that are relevant to the comprehension process. These factors include human characteristics related to the form of the ears and the head, as well as the human torso, all of which play their own part in the comprehension process. Altogether, the head, pinna and torso filter the sound wave by diffraction and reflection before it reaches the eardrum and inner ear that are responsible for transduction.

As Pec and Bujacz (2007) state, "measuring of HRTFs is a complex process as the transfer function must be calculated for a large number of directions relative to the head" (p. 2327). There are therefore multiple methods of the measurement, but only two general types: the direct and the reciprocal. According to Haraszy (2012), the direct method is used to measure the HRTFs by broadcasting the test signal at different directions in the 3D space and recording the received acoustic signal at the entrance of the two ears. The reciprocal method is based on Helmholtz' principle of reciprocity: the place of the broadcasting and receiving end are exchanged, which gives more opportunities for measurement. The reciprocal method is rarely used however, mostly because the direct method is cheaper, faster to perform, and can provide realistic results through the use of applied interpolation. Tise interpolation is a method of constructing new data points within the range of a discrete set of known data points. This in turn allows the measurements to collect all of the meaningful data using a small amount of control points as opposed to the large amount of individual measurements needed to cover the whole area. As Hacıhabibog˘lu, Gunel and Kondoz (2005) observe, “there exists a variety of methods applied to the HRTF interpolation problem” (p. 134), showing the method’s wide use.

When it comes to practical application, there are numerous fields where the head-related transfer function can be used: most of them are connected to entertainment industries such as in home gaming, music and cinema setups. In these case, there are approximate levels of the HRTF that are being used to relay the audio to the listener: this method is called the dummy head recording, and although it may not fully use everything the HRTF has to offer, it can certainly provide a consumer with a good quality product. It is mostly suitable for earphones, as there is a huge problem with stereo speakers since the sound cannot be separated into two independent parts: the left speaker is going to be heard by the right ear, and vice versa. To fix this problem, the so-called crosstalk cancellation technology has been developed which reduces the crosstalk effect by designing appropriate inverse filters for acoustic transfer functions.

In conclusion, the head-related transfer function, or HRTF, characterizes the process of a sound wave being filtered a by human’s head, pinna and torso through diffraction and reflection before the sound wave reaches the eardrums and inner ear. There are two methods of measuring HRTF – the direct and the reciprocal; the direct method is used more often since it provides realistic results through using interpolation methods. The HRTF is actively used for emulating real space in home entertainment through earphones or stereo speakers, which relate to the approximate levels of a given listener’s HRTF setup. With all of this in mind, the popular technology of HRTF is a very complex scientific method of defying the process of locating the sound by human’s ear, involving physics and anatomy knowledge. The HRTF is an important technology to consider in the topic of auditory comprehension and will be used throughout the paper.

2.5. In Changing Techniques in Practical Surround Audio:

With the advent of multichannel sound came differing techniques in how to achieve it, but ultimately they diverge into two categories: traditional and virtual. A traditional surround system (Ambisonics, Atmos, etc.) uses an array of speakers spaced around one or more listeners, while a virtual system (Oculus, Vive, etc) uses a pair of headphones, a motion sensor, and a mathematical model of the human head. From a practical standpoint, a traditional setup has a moderate fixed cost and zero cost per listener, while a virtual setup has a low fixed cost and a moderate equipment cost per listener (as each has separate hardware), but there are a couple other considerations to take into account before a recommendation can be made for classroom use.

In a physical scenario with a varying number of viewer, the typical solution has been to install a traditional surround environment, but this is generally only a “good-enough” choice, and not perfect. The key issue here is that an array of speakers can only be perfectly tuned and matched for one point in a room; if you're on the right side of the tuned point, the rightmost speakers will always be louder than the leftmost. This is obvious enough for most listeners to be a non issue, and the situation improves with distance from the speakers, to the point where experimental audio classrooms may use only a fraction of the room and leave the rest as dead space to improve their listening. However, even a perfect center is not perfect; listeners also have to deal with boundary interference and tuning the room in general. The sound will never be perfect in a physical room, and how the room affects the sound varies widely based on the position of the listener, right down to inches. All in all, however, these issues only crop up in audiophile-grade surround sound, but physical surround sound scales very economically, so it is considered 'good-enough'.

Chapter 3: Experiments

3.1. Air Force Research Laboratory Experiment

Introduction

In an experiment done by the Air Force Research Laboratory, the effectiveness of a multichannel audio environment in comprehension was tested based on several real applications of existing multichannel environments.[2] The main factor of these applications is the need for complex information to be displayed properly to a given listener, with little or no extra confusion or problems due to the way the information is represented. These scenarios partly arose from military, police, and emergency situations where a certain individual in the line of action would need to have critical information relayed to them over an auditory medium. In cases where a direct, single source of information was used, like a telephone, the process is simple and not prone to that many problems. However, when a scenario calls for several sources of input at once all overlaid on top of each other, like a police radio dispatcher, the listener in charge of comprehending the information is prone to a lot more mistakes.[2] It is in these scenarios where existing processes have the most room for improvement, and one of the major possibilities to bring these improvements is the implementation of a multichannel sound system.

Methodology

To show the potential of these multichannel systems, an experiment was done to test the correlation between a listener's comprehension when faced with several competing audio sources and the format with which those sources were presented. The experiment used 3 different audio setups: Monaural, Dichotic, and 3D audio, illustrated in the figure below. In the monaural setup, each source of audio was projected in a standard mono way, with no extra separation for the listener. The Dichotic configuration is similar, but also has two sources from each ear and three sources similar to the Monaural setup, giving the typical experience found in a stereo setup with 2 speakers. Lastly, the 3D audio display uses Head-Related Transfer Functions (Seen in section 2.4) to simulate several sources of input that are placed separately around the listener. In the experiment's HRTF setup, 5 of the sources were placed evenly in a semi-circle 1 m away from the participants, while the other two sources were placed 12 cm outside of the left and right ears. The main difference between the monaural, dichotic, and 3d audio setups used in the experiment is the way that the words are presented through each audio format. With monaural and dichotic systems, the audio representation passed to the listener is "flat" compared to the HRTF format. The words are simple streams of information either coming from both speakers (monaural) or alternating between the two (dichotic), but there are no extra characteristics of the sound that can be used for comprehension. In the HRTF format, the addition of a spatial representation of each sound allows the listener to hear each piece of information as if it were said at a certain point of space in the room they take the experiment in. Whether or not this extra level of distinction present in the HRTF format had any additional benefit was the major factor tested in the experiment.

Each participant in the experiment was asked to identify a certain combination of a color and number that appeared alongside a phrase containing "Baron". This target phrase would also have up to six interfering talkers from the other specified audio sources, and was always presented 100ms prior to the masking phrases. This process was repeated several times for every possible number of interfering sources (2-7) and the results of each trial were compared.

Results

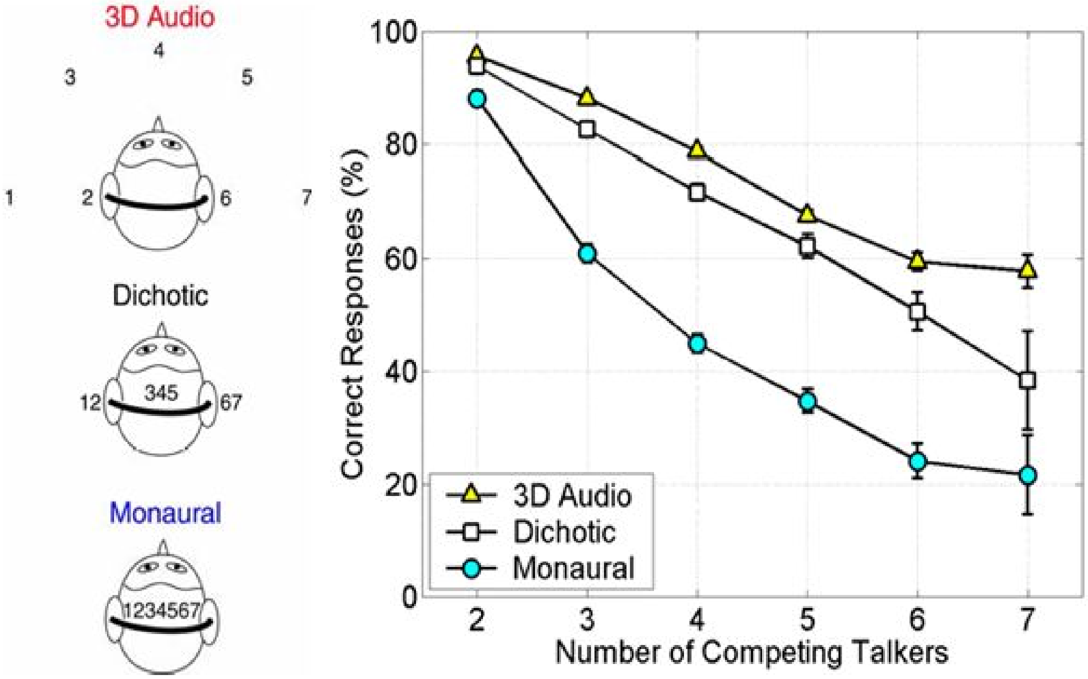

Figure 3.1 Results of the Air Force Research Laboratory Experiment (Brungart, Simpson, Iyer, 2007)

In the figure above, the symbol for each audio setup represents at least 300 trials with the error bars showing ± 1 standard error. When the number of competing talkers is only two, each of the three audio setups have similar results, with the more complex multichannel setups giving a slightly higher performance. However, as the number of competing talkers increases and the need for greater distinction arises, the Monaural and Dichotic setups begin to fail. In the case of the Monaural setup, the drop off is extremely steep at as low as 3 competing talkers, with a ~60% correct response rate compared to the 3D Audio's ~90% correct response rate. The Dichotic setup remains closer to the 3D Audio setup as the number increases, but always remains slightly lower until the larger number of competing talkers (6-7), where the correct response rate drops off significantly. At the highest number of competing talkers, the correct response rate for the Monaural, Dichotic, and 3D Audio setups are roughly 20%, 40%, and 60% respectively, with the 3D audio setup having the smallest standard error and the most consistency.

The results of the experiment suggest several important themes that Bregman outlined regarding the comprehension of auditory scenes. In each audio setup of the experiment (Monaural, Dichotic, and 3D), the level of distinction between the words increases. Instead of the words being played back directly on top of each other, they are represented as being in separate points in space. This property is mostly utilized in the 3D audio setup, but a less extreme version is also used when transitioning from the Monaural to Dichotic representations. The results of the experiment are in line with Bregman's ideas-As the distinction between each individual source of sound increases, the level of comprehension in the listener increases. Similar to the examples of a 3D visual scene, the spatial component of the sounds provides the listener with more information than just a single stream of audio. The information is still taken in as a stream, but the signatures of the words in the stream "picture" are separated enough to provide useful additional information for comprehension. The experiment was done in an attempt to find improvements to the existing communication systems that use the simple monaural setup, which raises the questions of how the existing methods can be improved or replaced by more advanced multi channel environments.

3.2. WPI Experiment

Introduction

The goals of the experiment that we performed are to further explore these comprehension enhancing possibilities. In particular, we want to observe subjects performance on a simple cognitive learning task with both a one audio input channel setup and a multi channel input setup, and analyze whether or not the multi channel learning environment resulted in increased performance. These results may also provide some useful information about what types of input are easier to process with one input or several, and open up topics for future research.

The major way this experiment will combine society and technology is the potential for enhanced learning using a multi channel audio system. As stated, the methods widely used in learning environments work to an extent but have potential to be greatly improved. Based on the effectiveness of learning with several channels, information that is learned from one source may be able to be learned far faster when modified to be presented in several channels. Just how much faster it could be learned is unclear, if possible at all, but the potential to process information is huge. If the processing time was reduced by an order of magnitude, then the benefit of the multi channel system would be clear, as heaps of information could be learned in a fraction of the time it would normally take. But even a small decrease in processing time, possibly around 25 to 40 percent, could still be enough to try to come up with systems to take advantage of this performance boost.

Methodology

To begin the experiment, the test subjects are first asked a few basic questions in a survey, covering their name, prior background in music, and their dominant hand. These questions are not directly related to the recognition of information in surround and stereo sound, but they could provide more interesting observations once the data has been collected. After answering these questions, the subject will listen to an audio file roughly 35 seconds in length consisting of 30 randomly chosen words. The first file will be played out of a stereo speaker system, and the words will randomly come out of either the left or right speaker, with a roughly even amount of words coming from each. There are small overlaps with several of the words, but no two words are played directly over each other to ensure every word can be heard clearly. In addition, the spacing of the words is not uniform, but scattered randomly to simulate a real life scenario of processing audio information. After listening to the file once, the subjects will continue the survey where they are asked with recognizing words that they heard. The survey has 30 total words, with roughly half of the words having appeared in the audio file and the other half not having appeared. Like the words that were put in the audio file, the half that were not in the file but listed in the survey were also randomly chosen, and were proofed to ensure that no words were repeated within individual trials and across trials. This process will be repeated again for the stereo system, and then the speaker setup will be switched to the 8 surrounding outputs. The process is still the same, but the audio files are played out of 8 channels evenly distributed 360 degrees around the subject. This is repeated again for a total of 2 stereo trials and 2 surround trials. At the end of all 4 trials 2 additional questions are asked. The first asks what audio setup the subject thought was better for remembering the words, and the second is an open ended question asking for potential suggestions in using the technology in an educational or learning environment.

There is not any designated rest period between the files, but the subjects are free to complete the task at their own pace. The total time of the experiment is approximately 10-15 minutes. Once the responses have been recorded, the main point of data to be analyzed is the percentage of correct answers that the subjects gave in the stereo and surround setups, and any possible conclusions that can be made to suggest that one works better than the other. Each of the four trials will have a total score of 30. Each correctly identified word will count as a correct answer, as well as each word that was correctly left unselected. If words that were not in the file were selected or words that were in the file were not selected, then there is no point deduction, but no points are awarded. The overall performance of the subjects will be based off of this percentage score, and other patterns in the data will be explored.

Results

As a whole, the results of the experiment did not show a clear distinction between the stereo and surround setups. The average percentage score for the stereo trials was 76.64 %, while the surround score was 72.44%. These two values are close enough to each other that a definitive statement about the setups is hard to make, but there were some interesting patterns when looking at individual trials for each participant. One of these patterns was related to the amount of words that the participants selected after hearing the file, and how this amount changed in the stereo and surround setups. In the data collected, the subjects either kept a consistent amount of words selected that they thought they heard in the file, or the amount slightly increased during the surround trials. For those subjects that selected the same amount of words, the scores in the surround sound trials stayed about the same or decreased, while those who selected more words in the surround trials scored better. This pattern is not definitive, but it could suggest something important about the surround audio setup, that the ability to confidently remember what was heard is enhanced.

As an example, in one particular subject's trial the majority of the incorrect options selected were from words that were in the file but not selected by the subject. Nearly all of the words selected by the subject were also in the file however, so the subject was often correct when selecting words but did not select them all. This trial also resulted in the highest score out of all other trials, 86.6% correct. As for why this happened, it is possible that the surround setup allowed for more distinction between certain words due to the amount of sources being increased. In the stereo setup, each speaker played 15 words as opposed to 3-4 words in the surround setup. When trying to remember if a word was in the file or not, the stereo setup would only have two additional pieces of information to factor into the subject's decision- whether the word was out of the left or right speaker. In the surround case, the subject may have distinctly remembered that they heard a certain word from behind them, in front of them, or from some other angle that made the word appear unique compared to one heard in the stereo setup, resulting in a higher percentage of correct words selected.

Chapter 4: Discussion

4.1. Effectiveness of Existing Learning Environments

An important thing to consider when exploring different educational options for multichannel audio sound is the effectiveness of existing education methods, and how they compare to the new methods proposed. The experiments in chapter 3 give some direct comparison between multichannel and stereo sound, but there are many other questions of where education could be improved. This topic spans across several different levels of education as well, and may provide different results in different scenarios.

Especially at the college and secondary education levels, standard lectures are one of the most popular and widely used methods of learning. The length of these lectures can vary, with many common lecture times being between 1-3 hours, sometimes more. In a study on the effectiveness of these traditional college lectures, several researchers found many benefits from an alternative approach deemed active learning. Active learning can be seen as a method of learning where students take a more involved approach than a traditional lecture experience. Some examples of active learning include problem based learning, group discussion, "clicker" polling techniques, and studio classrooms. According to the findings, the traditional lecture formats were not what many believed to be the most efficient way of learning new information, and that introducing concepts of active learning had significant benefits.[5] The study then states that "Active learning leads to increases in examination performance that would raise average grades by half a letter, and that failure rates under traditional lecturing increase by 55% over the rates observed under active learning" (Freeman). These improvements are very significant, and show a large window for improvement over the traditional lecture format widely used today.

Another important note made about traditional lectures is the unimportance of physically being at the location the lecture is given in. With a traditional lecture, a popular practice among educators is the use of recorded lectures and online courses to relay information to students. In some cases, entire degrees can be earned online without the need for a formally scheduled and attended in person lecture. The effectiveness of these online lectures has been a point of discussion for many educators and students, and opens up several possibilities to expand on the traditional lecture.[3] In a study By David Chandler, these online lectures are studied in more detail, taking into account several factors such as topic of the course and the experience of the students going into the online course. The study found that the online classes "really can teach at least effectively as traditional classroom courses… regardless of how much preparation and knowledge students start out with." (Chandler). These findings highlight another critical point when discussing online lectures, related to students who are learning the material online for the first time. At the collegiate level, this point is especially important due to the large variety of different educational backgrounds of the students in the class. For a course like a first year math course, where previous knowledge of the subject can range from nothing at all to nearly the entire course load, the fact that the online lectures would be effective for every student is extremely important. Even at a high school level, the prior knowledge students have when entering a new class can vary widely based on their aptitude in classes taken as prerequisites. If online lectures had a tendency to only work for those who have a firm understanding going into the course, then a student who may have struggled with previous courses related to the subject would fall further and further behind. But with the results found in this study, that is not normally the case, and the online lecture still allows students who may be trailing behind to not be left behind.

With the results of several studies highlighting the potential flaws of traditional lectures and benefits of online lectures, the question of where an audio based learning approach can be fit in still remains. Using the results of these studies, there are several different possibilities of using audio focused learning as both a supplement and a replacement. In the case of traditional lectures, these methods would most likely be better suited as a supplement. Many lectures require a visual component to them, and even if they do not, many people prefer to learn in a visual manner. However, if these multichannel methods did prove to increase comprehension, an example of supplementing the lecture would be changing the layout of the speakers in the lecture room to a surround setup. This would require a small upfront cost to change, but does not require any extreme changes to the traditional lecture setup and could have a benefit that far covers the cost. If these ideas were going to be implemented in a lecture hall that was under construction, then things like the layout with respect to how the sound moves throughout the room can also be considered.

In online lectures, the idea of adding multichannel audio as a supplement is the same, but there is also potential for a complete replacement. Both of these ideas mostly revolve around the use of the HRTF. Whether the lecture is streamed live from some location, or is an uploaded video, if the format of the audio supports the HRTF then there is an immediate benefit with very little additional cost. As long as the listener has some way to make use of the HRTF, which is more common with technological advancements, then the online lecture then has access to all of the benefits that come with the multichannel setup. Changing the audio in this way alone could be very beneficial, but the way that online lectures are created could be changed as well. For example, the lecture itself could be an audio recording, and the creator places certain information in a specific channel in a strategic way to maximize comprehension.

4.2. Potential of multichannel technologies (Dolby Atmos)

Although the benefits of moving away from simple stereo and mono sound systems have not been widely implemented, some corporations have been using a multichannel approach to existing sound systems to add another layer to the user experience. A significant example of one of these companies is Dolby, who have recently introduced and been developing an audio setup system known as Dolby Atmos. The basic idea of the Dolby Atmos sound system is to simulate audio that can’t be traced back to a speaker source, but instead is heard naturally by the user as if they were living the desired experience. To achieve this, the Atmos system uses as many as 7 to 34 different speakers strategically placed in the room, based on each speaker’s functionality. Some speakers may output audio in a more traditional way, but the “upward-firing Dolby Atmos speakers” and the “Dolby Atmos enabled sound bar” truly enable the Atmos system’s purpose.[4] The image below shows the recommended 7.1.2 speaker setup to be used in a home setting.

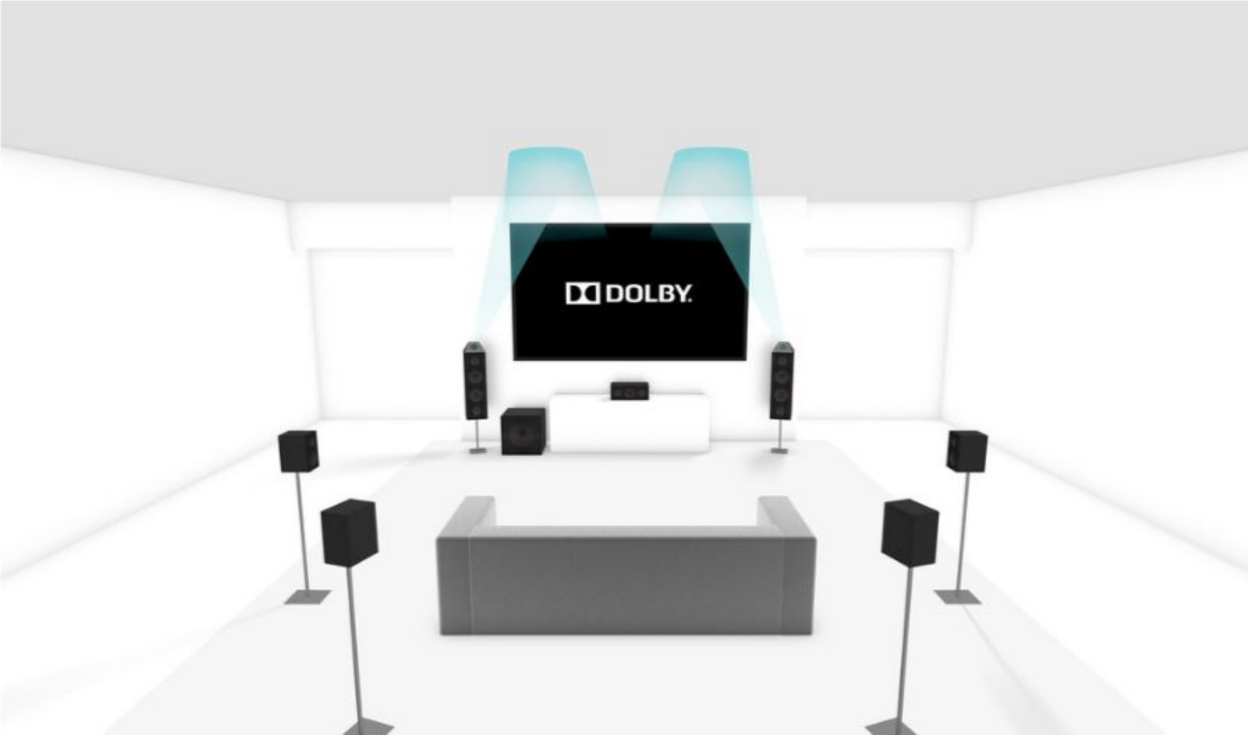

Figure 4.1 An example setup of the Dolby Atmos home sound system (Dolby)

These sources can be located anywhere in the room, including the floor and ceiling, and are able to take portions of the audio being played and disperse it around the room, effectively changing the perceived source and engaging the listeners to use the entirety of their surroundings. For instance, a certain movie scene may have a helicopter fly from a distance as seen by the viewer, and then pass over the camera as if the helicopter were flying overhead. With a single, or simply limited number of audio sources, it is very hard to represent the change in position of the sound of the helicopter in the audio format played to the listeners. In many cases, the listener would hear the audio in only that one dimension, and it can easily be traced back to the source speaker. But with the Dolby Atmos system, the sounds of the helicopter would be heard by the listeners as if it had just flown over the theater, giving a much more immersive and engaging experience.

The Dolby Atmos system provides an enhanced experience for movie viewers in a theater, but a large part of the potential of the system comes from the home speaker setup that can be used in any reasonable area. The home setup would normally have less speakers than the theater setup, but the idea behind the system is the same-speakers are set up all around a couch, chair, or other viewing area that take advantage of the multi channel audio format. The idea behind the home setup is important in the potential profit of the Atmos system, but also opens up several possibilities about using similar setups in learning and comprehension environments. Similar to the situation described under the lecture section of the paper, the Atmos home setup is a prime example of something that can be used in libraries, schools, and homes with supported and tested audio formats. Instead of educators needing a dedicated space the likes of a movie theater to use the multichannel learning environment, a similar setup to the home Atmos system can be used in any traditional classroom. The potential benefits of this are clear to see: saving money from having to construct a dedicated multi channel sound room, the ability to have several setups working at once, and simple integration with other specialized learning rooms such as labs are a few examples.

Chapter 5: Conclusion

To reiterate the ideas that Bregman set forth in *Auditory Scene Analysis*, the field of auditory comprehension when compared to the visual counterpart is one with a lot of room for growth. The properties of the two fields gives an advantage to visuals when representing and presenting information, but with advancements in auditory technology and the natural abilities that humans have in the perception of sound, this advantage is subject to change. The main situations where these abilities match or overtake the abilities of visual perception are those involving audio input from several sources at once, in a multi channel environment. In these scenarios, the natural abilities of auditory perception in humans are highlighted, resulting in things like the cocktail party effect, or a general ability to separate and represent meaningful information. As seen in the experiment done by the Air Force Research Laboratory, these abilities can be utilized in certain situations to provide a better level of comprehension in a multi channel environment.

However, much like the research side of the field, the implementation of these auditory abilities is not seen as much as it could be. A large industry where the multi channel setup is seen in is the entertainment industry, with technologies like the Dolby Atmos speaker setup and Head Related Transfer Function are used to enhance the immersion of the listener. These technologies worked well, and a great deal of development was put into them for use in a personal or home environment, as opposed to an expensive industrial setting. With this in mind, a potential area of development in the audio field as a whole then becomes applying this process for the learning and comprehension industry, namely schools, college lectures, or any training events. Expanding this area is a major argument point of this paper, as the results from the experiments and the available information on similar situations show the potential benefits.

Improvements in comprehension in any environment are extremely beneficial, and any opportunity to create them should be given a high priority. In the past the opportunity that comes with multichannel audio environments may have been too expensive or not known well enough to be used, but with the new technologies and information set forth in the field the opportunity is a lot more concrete. With the continuation of research in the area, the addition of a multichannel audio component can become a staple component of learning environments and add a valuable method of improving comprehension.

Appendix

WPI Experiment

Correct and Incorrect Words for reference

Table A.1 Trial 1 (Stereo) Word Reference

Words in File	Correct Options	Incorrect Options
convenience guess recruiter diameter splice recombination jail fence lamp basin residue mode hospital personnel cube cuff acceptor revision stand mile buzz leak clock lap link alignment father canal relocation gland	mode link splice leak basin stand gland recombination father clock cuff relocation personnel residue cube	reduction gasoline troop stowage multiplex soils report safety grass misalignment hoses responsibilities directive trash mailbox

Table A.2 Trial 2 (Stereo) Word Reference

Words in File	Correct Options	Incorrect Options
restraint	bandage	commendation
calculator	chill	shears
props	kettle	churns
recruiter	skill	particle
electronics	feed	basin
bandage	morale	try
zip	height	rotation
skill	zip	drip
morale	assignment	hoops
spiral	chemical	guilt
shape	act	swamp
assignment	drug	clock
kettle	toss	labors
validation	restraint	accountability
chemical		configuration
jig		ribs
malfunction		
splice		
track		
trod		
feed		
drug		
height		
affair		
bracing		
lamp		
gland		
act		
toss		
chill		

Table A.3 Trial 3 (Surround) Word Reference

Words in File	Correct Options	Incorrect Options
trod glue bail hole welder chill classification maneuver sprayer whisper delivery triangle cloud helmsman painting pulse trailer curve freedom group leap stretcher designator economy half peak label sight engine drink	label economy cloud welder curve triangle whisper chill glue classification maneuver stretcher designator delivery	block admiralty measures try states sale conventions piles captains torpedoes gap bypasses guess entrapment annex official

Table A.4 Trial 4 (Surround) Word Reference

Words in File	Correct Options	Incorrect Options
grown	flowchart	adjustments
restraint	oscillation	medium
citizen	jail	powder
mile	mode	worries
star	link	baths
photodiode	freedom	record
specialty	glue	realignments
oscillation	travel	transit
zip	longitude	raps
glue	strobe	distribution
mode	foam	dents
guideline	property	accessories
link	weld	age
grip	photodiode	hose
weld	zip	minutes
bristle		
area		
waste		
jail		
flowchart		
function		
freedom		
column		
foam		
travel		
fraction		
longitude		
property		
strobe		
propose		

Collected Data

Response 1

Year of graduation: 2018

Dominant hand: Right

Music background: Only a little

Self evaluation of memory: Neither good nor bad

Trial 1 results: 12 selected, 18 unselected, 23/30 correct (76.6%)

Trial 2 results: 11 selected, 19 unselected, 21/30 correct (70.0%)

Trial 3 results: 11 selected, 19 unselected, 17/30 correct (56.6%)

Trial 4 results: 12 selected, 18 unselected, 23/30 correct (76.6%)

Multichannel compared to stereo: Somewhat better

Ideas for learning implementation: Possible changes in lecture format

Things to note: In multichannel results, many of the incorrect words were those the subject thought they heard but were not in the file

Response 2

Year of graduation: 2016

Dominant hand: Right

Music background: Only a little

Self evaluation of memory: Somewhat good

Trial 1 results: 11 selected, 19 unselected, 24/30 correct (80.0%)

Trial 2 results: 12 selected, 18 unselected, 26/30 correct (86.6%)

Trial 3 results: 7 selected, 23 unselected, 19/30 correct (63.3%)

Trial 4 results: 11 selected, 19 unselected, 20/30 correct (66.6%)

Multichannel compared to stereo: Somewhat better

Ideas for learning implementation: Immersive education allows for a more bodily interactive education experience. As more parts of the body are involved, the more that is retained

Things to note: In the first multi-channel trial the number of words the subject thought they heard was fairly low, but increased in the second multi-channel trial

Response 3

Year of graduation: 2016

Dominant hand: Right

Music background: None

Self evaluation of memory: Somewhat good

Trial 1 results: 9 selected, 21 unselected, 24/30 correct (80.0%)

Trial 2 results: 8 selected, 22 unselected, 20/30 correct (66.6%)

Trial 3 results: 10 selected, 20 unselected, 26/30 correct (86.6%)

Trial 4 results: 10 selected, 20 unselected, 23/30 correct (76.6 %)

Multichannel compared to stereo: Somewhat worse

Ideas for learning implementation: For scenarios that require students to visualize themselves inside a part of an object being studied in class

Things to note: Although the subject thought that the multichannel setup was worse for remembering the words, their results in the multi channel setup were better.

Response 4

Year of graduation: 2019

Dominant hand: Right

Music background: Music is my life!

Self evaluation of memory: Somewhat good

Trial 1 results: 7 selected, 23 unselected, 22/30 correct (73.3%)

Trial 2 results: 8 selected, 22 unselected, 24/30 correct (80.0%)

Trial 3 results: 10 selected, 20 unselected, 20/30 correct (66.6%)

Trial 4 results: 11 selected, 19 unselected, 26/30 correct (86.6%)

Multichannel compared to stereo: Somewhat better

Ideas for learning implementation: None

Things to note: The subject often did not select many options, but the options selected were usually in the file, so the errors were from not being able to correctly identify a word.

General Data

Average Stereo Score: 76.64 %

Average Surround Score: 72.44 %

Works Cited

1. Bregman, Albert S. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Mass.: MIT, 2001. Print.

2. Brungart, Douglas S., Simpson, Brian D., Iyer, Nandini. “Maximizing Information Transfer in Auditory Speech Displays” *Air Force Research Laboratory.* 2007

3. Chandler, David L. "Study: Online Classes Really Do Work." *MIT News*. Massachusetts Institute of Technology, 24 Sept. 2014. Web. 27 Oct. 2016. <http://news.mit.edu/2014/study-shows-online-courses-effective-0924>.

4. “Dolby Atmos for the Home” Dolby Laboratories, n.d <https://www.dolby.com/us/en/technologies/home/dolby-atmos.html>

5. Freeman, Scott, Sarah L. Eddya, Miles McDonougha, Michelle K. Smith, Nnadozie Okoroafora, Hannah Jordt, and Maty Pat Wenderoth. "Active Learning Increases Student Performance in Science, Engineering, and Mathematics." *PNAS*. Proceedings of the National Academy of Sciences of the United States of America, n.d. Web. 27 Oct. 2016. <http://www.pnas.org/content/111/23/8410>.

6. Gulick, Lawrence. Hearing: Physiological Acoustics, Neural Coding, and Psychoacoustics. New York: Oxford University Press, 1989.

7. Hacıhabibog˘lu, H., Gunel, B., & Kondoz, A. (2005). *Head-Related Transfer Function Filter Interpolation by Root Displacement.* In 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 134-137). Guildford, UK: University of Surrey.

8. Haraszy, Z. (2012*). Personalized Head-Related Transfer Function Measurement for Acoustic Virtual Reality Development.* U.P.B. Sci. Bull., 74(3), 110-122.

9. Howard, David. Acoustics and Psychoacoustics. 4th ed. Oxford: Focal Press, 2009.

10. Larsen, Erik. Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. Chichister: John Wiley and Sons Ltd., 2004.

11. Pec, M. & Bujacz, M. (2007). *Personalized Head Related Transfer Function Measurement and Verification through Sound Localization Resolution.* 15Th European Signal Processing Conference (EUSIPCO 2007), 2326-2330.

12. Werner, Lynne. Developmental Psychoacoustics. Washington DC: American Psychological Association Developmental, 1992.

13. Zwicker, Eberhard. Psychoacoustics: Facts and Models. 2nd upd. ed. New York: Springer, 1999.References