1-30-2024 12:00 PM

Validation of a virtual auditory space, and its use to investigate how pitch and spatial cues contribute to perceptual segregation of auditory streams

Nima Zargarnezhad, *Western University*

Supervisor: Ingrid Johnsrude, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Neuroscience

Abstract

The human auditory system can decompose complex sound mixtures into distinct perceptual auditory objects through a process (or processes) known as Auditory Scene Analysis. Pitch and spatial cues are among the sound attributes known to influence sequential streaming (Plack 2018). In this project, the fidelity of a virtual acoustic space (the Audio Dome) in reproducing precisely located sound sources with a 9th-order ambisonics algorithm was validated. The estimated horizontal Minimum Audible Angles aligned with previously reported values (Mills 1958) homogeneously across the space, and a robust low-frequency presentation was identified. Then, the Audio Dome was utilized to test van Noorden's (1975) ABA paradigm with displaced A and B sources on a continuum of locations and several pitch differences. A two-dimensional sigmoid function was utilized to model this two-dimensional psychophysical space and revealed that spatial and pitch cues are both essential to organize perception, with pitch cues perhaps being more influential.

Keywords

Higher-Order Ambisonics, Virtual Acoustic Spaces, Minimum Audible Angles, Auditory Scene Analysis, Perceptual Organization, Psychoacoustics

Summary for Lay Audience

In our daily lives, we are surrounded by lots of sounds that are most likely heard with other sounds. For example, when passing by a park, one might hear birds tweeting, children playing, people talking, bikes crossing, etc. These sounds all spread in the air, get mixed, and the mixture of sounds reaches our ears. Our auditory system can distinguish these sounds from the mixture and link them to the surrounding objects and events. This ability, referred to as "Auditory Scene Analysis", relies on several attributes of sounds to distinguish them. For example, the sounds that come from the right of the body are less likely to be related to some object that is heard from the left. Therefore, the human mind uses location information to assign the sounds to two different objects. Similarly, when a female and a male voice are heard in a radio show, their sounds come from the same location, yet we can tell apart their sounds based on other qualities of sound. One of these qualities is "pitch," the same quality that enables us to distinguish between different notes played on a piano. Now the question is, what would happen if one distinguished sounds based on their pitch but their location is not different? (And vice versa?) Would the human mind rely on only one attribute and neglect the other, or do they cooperate? The present work shows that location and pitch are both important for the mind to analyze auditory scenes, with pitch cues perhaps being more influential. These results help us understand the importance of sound attributes for audition, which is essential for designing functional hearing aids. A virtual auditory space (the Audio Dome) was used to manipulate sound source locations and create auditory scenes to accomplish this goal. Because the Audio Dome was not previously used for research purposes, the suitability of this newly installed device for auditory research with humans was established in the first study. The validation experiment results should reassure researchers who wish to use the Audio Dome for further auditory research experiments.

Co-Authorship Statement

The study described in each experimental chapter will likely be submitted for journal publication, with Nima Zargarnezhad as the first author for both. The co-authors for the first experimental chapter will be Bruno Mesquita, Ewan Macpherson, and Ingrid Johnsrude. Bruno Mesquita contributed to the conceptualization, design, and implementation of the experiment, as well as editing this thesis. Ewan Macpherson contributed to the conceptualization and design of the human experiment, the design and implementation of the Head and Torso Simulator measurements, and editing this thesis. The co-authors for the second experimental chapter will be Ezgi Coskun and Ingrid Johnsrude. Ezgi Coskun contributed to the data collection for this experiment. For both studies, Ingrid Johnsrude served as the principal investigator, provided funding, contributed to the conceptualization and design of the experiment, to the interpretation of the results, and to editing this thesis.

Acknowledgements

I wish to express my heartfelt appreciation to the individuals who played crucial roles in the completion of this thesis. Their consistent support, guidance, and encouragement have been invaluable throughout my academic journey.

Foremost, I extend my sincere thanks to my supervisor, Dr. Ingrid Johnsrude, for being a supportive mentor and a wonderful individual. Her insights into the intricacies of the human brain and behavior have been enlightening, and I feel honored to have learned from her and to have been her student. Her guidance has significantly contributed to my growth in various aspects.

I am grateful to my advisors, Dr. Ewan Macpherson, Dr. Blake Butler, and Dr. Jessica Grahn, for their continuous support and constructive feedback. Their expertise and encouragement have played a vital role in shaping the direction of my research.

I acknowledge the members of CoNCHLab, especially Bruno Mesquita, whose enthusiasm and curiosity served as a constant source of motivation. Additionally, I appreciate Ezgi Coskun, the dedicated undergraduate student whose assistance in data collection for my experiments was indispensable.

Special mention goes to Dr. Derek Quinlan for his support and assistance with research facilities and experiment setups, particularly with the Audio Dome. I also want to express my gratitude to Karsten Babin, who played a crucial role in initiating the programming aspect of this work. Without them, this project would not have reached completion.

My heartfelt appreciation extends to my family and friends, especially Ali Tafakkor, Chelsea (Bo-ra) Kim, Mansoure Jahanian, and Saba Charmi Motlagh. Their unwavering support and the unique relationships we developed have been a source of joy and strength. I am particularly thankful for their presence during the challenging past two years, helping me bridge the distance with my immediate family.

The completion of this thesis would not have been possible without the collective support, encouragement, and contributions of these remarkable individuals. I am genuinely grateful for their impact on both my academic and personal growth.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Appendices

List of Symbols, Abbreviations, and Nomenclature

β_x	Psychophysical model coefficient for variable x
Δf	Frequency difference between A and B tones in the ABA paradigm (F_B - F_A)
Δf^*	84.1% threshold value for Δf
$\Delta\varphi$	Azimuth angular difference between A and B tones in the ABA paradigm
$\Delta\varphi^*$	84.1% threshold value for $\Delta\varphi$
θ	Elevation angle coordinate
φ	Azimuth angle coordinate
ψ	Psychometric Function
ABA	van Noorden's (1975) paradigm
ASA	Auditory Scene Analysis
EV	Explained Variance
F_A	Frequency of A tones in the ABA paradigm
F_B	Frequency of B tones in the ABA paradigm
GUI	Graphical User Interface
HATS	Head-And-Torso Simulator
HOA	Higher-Order Ambisonics
HRTF	Head-Related Transfer Function
ILD	Interaural Level Difference

ITD	Interaural Time Difference
JND	Just Noticeable Difference
LED	Light-Emitting Diode
LSP	Loudspeaker
MAA	Minimum Audible Angle
MLS	Minimum-Length Sequence
OSC	Open Sound Control
PSD	Power Spectrum Density
PSE	Point of Subjective Equality
R^2	Coefficient of Determination
STD	Standard Deviation
VBAP	Vector-Based Amplitude Panning

Chapter 1

1 Introduction

Hearing is one of the basic human senses that registers sound as reflections of the outer world. Material objects and their physical interactions cause vibrations that propagate through the matter that fills the environment between the source and a listener's ear. The vibrations are relayed through the outer and middle ears and transduced into electrical signals, which propagate through the auditory nerves to subcortical structures and the cortex, initiating an auditory experience. Auditory perception facilitates awareness about the world by forming perceptual objects and establishing connections between the aural experience and (physical or perceptual) objects and events. Through time, humans have started to benefit more from this ability by producing purposeful sounds to communicate with each other and to express themselves via language and music.

Irrespective of sounds' origins or intended purpose, one of the fundamental challenges in audition is the formation and organization of perception to assign meaning to the auditory experience. This association is achieved by applying perceptual organization principles to physical and perceptual attributes of sounds. Within the diverse types of information, cues, and sound properties contributing to auditory perceptual organization, spatial and pitch cues are two major groups of cues provided in natural hearing experiences. Listening to live ensemble music is an excellent example. In this experience, some listeners perceive the music as a cohesive auditory object (music), while others (or the same listener at a different time) might be able to segregate so as to attend to individual instruments by virtue of their being at different locations or playing different pitches at any given moment. Conductors who (literally) orchestrate this experience are professionally trained to carefully listen to each individual and correct their performance.

In this thesis, I focus on how the human mind utilizes the combination of spatial and pitch cues to organize auditory perception. I specifically aim to identify mechanisms underlying the interaction of pitch and spatial cues when they are both provided to listeners. Studying these mechanisms (and potential individual variations) may help

researchers and educators develop strategies that improve skills related to the use of spatial and pitch cues for perceptual organization, such as those required for the art of conducting music ensembles.

1.1 Auditory Scene Analysis

Natural hearing experiences typically involve multiple environmental objects and events producing sounds at the same time. An illustrative scenario is the cocktail party situation where conversations, music, clinking glasses, and ringing phones form a complex auditory experience. With all these sounds experienced simultaneously, one might want to attend to a conversation or listen to music. Whatever the listener's preference is, the auditory system must be able to dissociate sounds from different origins by extracting them from the experienced sound mixture. Such situations, in which multiple objects of hearing are present, are called “Auditory Scenes,” and the process of segregating and grouping sound components to perceptually organize the scene into meaningful sounds is referred to as “Auditory Scene Analysis” or ASA (Bregman 1990).

The phrase “scene” is borrowed from the terminology of visual scenes, in which one or multiple objects are resolved from the perceptual background within the perceptual foreground. The perceptual organization principles of vision facilitate the segregation of features that belong to different objects and the grouping of features that belong to the same object. To describe scene analysis, Bregman uses the coloring analogy: if one is asked to paint each object in *Figure 1.1* with a different color, they[1] will probably end up using two crayons as two cuboid objects are perceived in this scene; one occluded by the other one. The lines in this scene specify the borders and edges, the intersection of those edges identifies the corners, and those edges and corners define surfaces. It might be clearer why surfaces A, B, and C belong to the same object: they share corners and

[1] For inclusivity reasons, the singular “they” pronoun will be employed when referring to an individual of unknown gender, replacing the traditional use of “he/she” or “(s)he”.

borders, but how does surface D associate with them? Human perceivers do not have any difficulty understanding why D and C are the same surface occluded by another object. This is because of the unconscious awareness of the visual perceptual principle of good continuation, which in this context translates to aligned smooth edges being more likely to be continuous and associated to belong to the same object rather than being abrupt and sharp. (Bregman 2005; Guzman 1969; Koffka 1935).

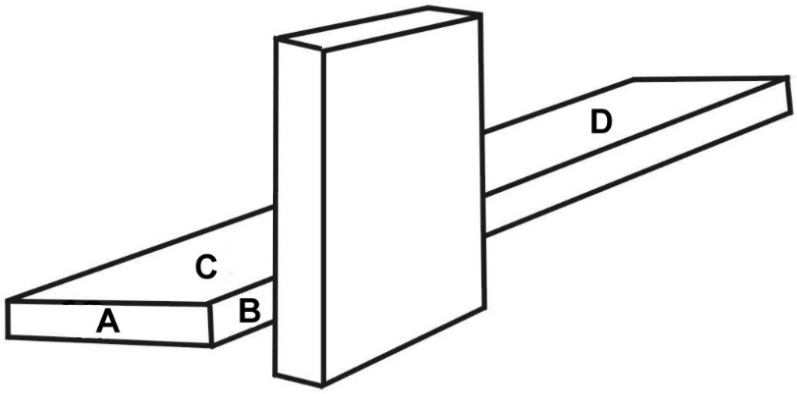

Figure 1.1 An example of a visual scene (Bregman 1990; Guzman 1969)

This perceptual "crayon" could be borrowed from vision to "paint" objects of audition: perceptual auditory streams. According to Bregman, an auditory stream is our perceptual grouping of [features of a] neural spectrogram that go together (Bregman 1990) as the object of audition, which, despite some differences in nature, is a parallel to visible objects. For example, suppose one is presented with repetitions of a sequence of 6 pure tones, illustrated in *Figure 1.2*. In that case, they will perceive two streams of sound: one composed of the high-pitched tones (1, 3, and 5 – shown in blue) and another one composed of the low-pitched tones (2, 4, and 6 – shown in red). In other words, tones 1, 3, and 5 are "integrated" or grouped together, and they are "segregated" from the grouping of tones 2, 4, and 6 (Bregman 1990). This organization of sound may not be very intuitive for the reader (especially since this example is not auditorily experienced and is just visually illustrated and described in written words). However, the question is, why don't we alternatively perceive a single stream of sound (or six different streams of

sound)? This is because of the principles of auditory perceptual organization, which govern the grouping and segregation of sound components in our mind.

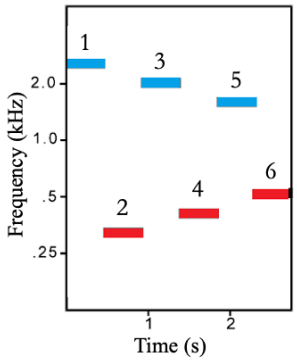

Figure 1.2 A visual illustration of an analyzed auditory scene (Bregman 1990)

Now imagine a scenario of two people with different voices (say one female and one male voice) speaking; one says, “I want coffee.” at the same time as the other says, “I hate zombies!”. Successful streaming is essential to hear these sentences as they are said instead of “I hate coffee!” and “I want zombies.” A successful scene analysis, in this case, needs to perceptually organize the sounds that overlap in time; for example, “I”s at the beginning of the sentences should be perceived with different voices (segregated) and be associated with two different sources (the same for “want”/”hate” and “coffee”/”zombies” pairs assuming the two people talk with the same pace). Additionally, the sources for the three word-pairs are related to each other through time (“want” and “coffee” are sequentially said by the same person). Discovering the correct sequential grouping of the tones interrupted with silence gaps but immediately following each other is another part of the successful scene analysis. The first type of analysis (segregating two things that happen at the same time) is known as “simultaneous grouping/segregation,” and the second one is referred to as “sequential streaming.” (Plack 2018; Bregman 1990; Darwin and Carlyon 1995) Although both types of analysis are required in most natural situations, the principles of perceptual organization translate differently to each of them, in term of the nature and importance of the information and cues each of them uses. Therefore, they are usually studied and discussed separately.

ASA benefits from different families of cues, including periodicity cues, spatial cues, temporal cues, harmonicity cues, spectral balance cues, level cues, etc. Among these cues, spatial and pitch cues seem relatively important for sequential streaming. Limiting the spatial cues to the location of the sound (not object movement cues and dynamic objects, for simplicity), the perceptual role that applies to them is that the sound components that come from the same location seem to be more likely to belong to the same object/stream and vice versa: the same location in space is more likely to form a stream than a sequence whose components are placed in different locations. Therefore, we tend to segregate sounds that are perceived from different locations (Plack 2018). Pure tones with similar frequency or complex harmonics with similar fundamental frequency yield pitch cues that promote grouping. Therefore, pitch cues are also considered strong cues in sequential streaming.

Contributions of spatial and pitch cues to ASA have mostly been studied with an isolated focus on one of these cue families in the literature. It is, therefore, unclear how they contribute to ASA when they both systematically vary in an auditory scene: do we rely on one set of cues more than the other? Do spatial cues dominate pitch cues and determine the organization of perception (or vice versa)? Or do spatial and pitch cues collaborate to form perception? If they collaborate, which one has a more critical role? Is the perceptual importance of spatial and pitch cues to ASA similar between different participants? I try to address these questions in this thesis as the first step to exploring a practical question: How do music conductors know how to locate errors and correct them in an ensemble rehearsal? We know that the instruments in ensembles are in different positions. Well-trained conductors are able to identify individuals (even within a section) who play a wrong note (which is identified by its pitch). In other words, they identify a different pitch at a specific location in a sequence of interrupted sounds. This suggests accurate sequential streaming with high spatial resolution: all the instruments have been segregated, and there is a one-to-one association between spatial locations and segregated streams. Although one might argue that the instruments are segregated based on other spectral differences (such as their timbre) or that maybe scene analysis in this situation is hierarchical and top-down attentional processes are involved, I keep the scope of this thesis to pitch and space for simplicity.

1.2 The Auditory System

The human auditory system consists of peripheral (*Figure 1.3.A*) and central parts. Sound pressure waves enter the peripheral auditory system through the outer ear (pinna and ear canal) to vibrate the eardrum. Eardrum vibrations are then amplified and transmitted by the middle ear (by auditory ossicles: malleus, incus, and stapes) to change the fluid pressure in the inner ear (cochlea). The cochlea is a spiral tube structure filled with fluid and divided into three parts by two membranes along its length. The fluid pressure changes in the cochlea translate to vibrations of the basilar membrane (one of the membranes along the cochlea). Because of its mechanical properties, enhanced by an electromechanical feedback system, the basilar membrane vibration is maximal to a specific frequency at each location - that location's *characteristic frequency*. The basilar membrane is tuned to lower frequencies towards the apex of the cochlea and to higher frequencies closer to the base (the tonotopic mapping, *Figure 1.3.B*).

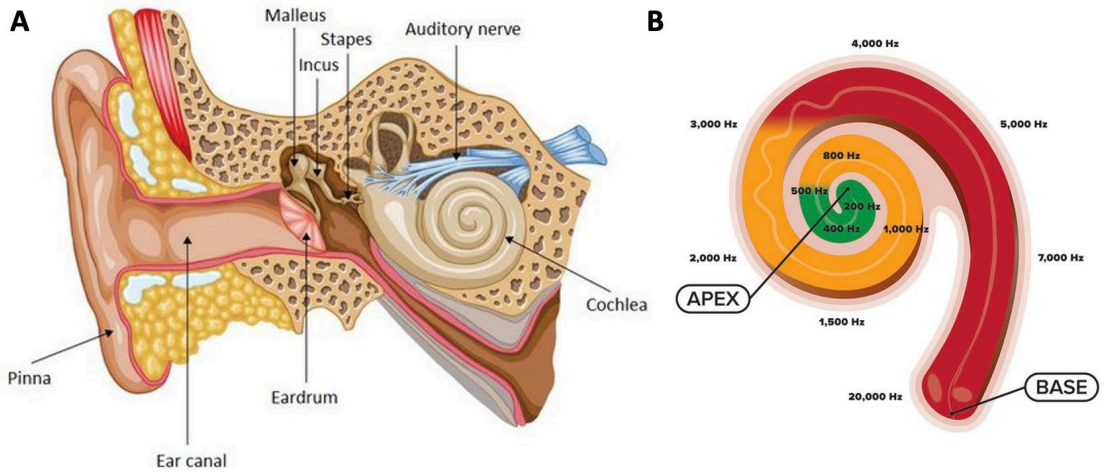

Figure 1.3 A. The peripheral auditory system. B. An illustration of the cochlea and its tonotopic characteristics. (adapted from Lahav and Skoe 2014)

The basilar membrane vibrations induce changes in the electrical potential of the sensory inner hair cells at the corresponding location, which leads to electrical activity in the nerve fibers attached to these cells. At this point, the mechanical vibrations are transformed into neural electrical activity. Furthermore, the electrical activity is modulated by the amplitude and phase of the stimulating waveform. The auditory nerve

carries this information to be further processed in the brainstem (through a chain of nuclei), thalamus, and the auditory cortex. It is worth mentioning that in addition to these ascending information pathways from the ears to the brain, there are descending connections which, in principle, enable higher processes such as attention to influence activity in lower centers (Plack 2018).

1.2.1 Perception of Auditory Space

A sound source in the physical world is located in a 3-dimensional space. While location could be specified in different 3-dimensional coordinate systems, one option is specifying direction, distance, and elevation as the three coordinate parameters (*Figure 1.4*). The human auditory system combines different types of information to estimate these parameters to locate objects based on their sound.

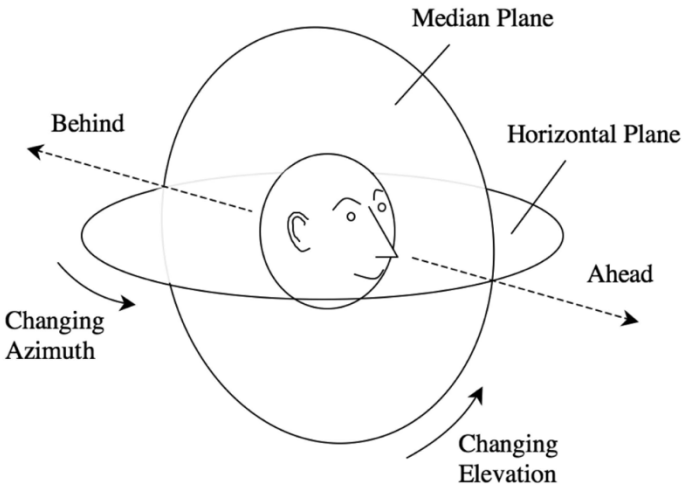

Figure 1.4 The coordinate system for sound direction (Plack 2018)

Two important cues that provide detailed information about the direction of the sound take advantage of having two ears. Cues that use information from two ears are called binaural cues. One binaural cue is the Interaural Time Difference (ITD), which is the delay in the sound arrival time between the ears. Another cue is the Interaural Level Difference (ILD), which indexes the difference in the sound intensity between the two ears. The logic behind using these cues is explained with an example: assume a sound source at the right side of the listener's body. The sounds from this source need less time

to reach the right ear than the left ear. Therefore, phase shifts between the sounds in the two ears are detected. Also, the sound that travels through a longer distance has less intensity because sound intensity decreases with the distance from the source. While both sets of cues are used in sound direction estimation, ITDs are more useful for lower frequencies (below 1000 Hz), whereas ILDs are more salient at higher frequencies (above 1500 Hz, Blauert 1997).

Distance as the second coordinate of a sound source location is estimated mainly by the intensity or level of the sound, timbre, and the ratio of direct to reflected sound. The assumption underlying the use of intensity to estimate the source distance is that quieter sounds usually come from sources that are further away, whereas louder sounds tend to come from closer sources. In reverberant spaces (most natural indoor spaces), the ratio of the energy of the sound directly coming to the ear over the energy of the reflected sound due to reverberation is another cue that helps estimate distance. The rule of using this cue is that the further away the sound source is, the greater the proportion of reverberant compared to direct sound energy at the ear.

Finally, the elevation of a sound source is mostly determined by the monaural spectral cues. The listener's head, upper body, and, most importantly, pinna have a particular shape that practically affects the sound spectrum as a mechanical filter. The frequency response of this filter, known as the Head-Related Transfer Function (HRTF), is modulated by the source elevation, which enables the auditory system, through experience, to estimate the elevation of the sound source.

However, humans have evolved high spatial acuity in the visual domain. Hence, the auditory spatial acuity has evolved to just be sufficient for approximating location; after the location of a sound is estimated, the head turns accordingly (if necessary), and vision is used to precisely locate the object. Therefore, the combination of auditory spatial cues is not perfect, and some ambiguity about the sound source location remains, especially about the direction of the sound determined by ITDs and ILDs if the individual cannot move their head. For sound sources laterally located to the side), the ITD (and sometimes ILD) cues are similar, which results in ambiguity in the precise location of the source.

This effect results in a set of locations that are physically different but produce the same ITD/ILD cues; hence, they are perceived at the same location. These locations which share ITD or ILDs are located on "cones of confusion" (*Figure 1.5*). These ambiguities are mainly resolved by additional cues such as information gathered by head movement.

Figure 1.5 An illustration of a cone of confusion. All sound sources located on the surface of this cone produce the same ITD and ILD cues. (Plack 2018)

The literature on spatial cues and their characteristics has led to the development of virtual auditory spaces. These virtual spaces, such as the stereo presentation of sounds over headphones, facilitate characteristics of ITDs, ILD, and approximations of pinna, head, and torso shape to manipulate the presented sound to be perceived from a desired location. (Plack 2018) Some of these technologies, their advantages and disadvantages, will be discussed later (*section 1.4* and Chapter 2) because of the crucial role of precise, systematic manipulation of sound source locations in this thesis.

1.2.2 Pitch Perception

One of the techniques to study and model signals is to represent them in the form of a combination of a basis function series that reflect the nature of the signal and study the signal in that representational space. Sound as physical changes of pressure in a medium is not an exception and is often represented and studied in the frequency domain. The building blocks of the frequency domain are periodic sinusoidal signals with a specific repetition rate. Any sound can be represented in this domain as a weighted sum of temporally jittered sinusoids. Sounds with just one frequency component are referred to as pure tones. The combination of multiple frequency components makes either complex

tones or noise, depending on the frequency and phase relations among the components. In the particular case where the frequency components are integer multiples of a particular value, the sound is called a complex harmonic tone, with a repetition rate equal to the fundamental frequency or the first harmonic of the sound (the integer multiplier for other components determines its index).

Pure tones and complex harmonics with the same fundamental frequency have the same periodicity or repetition rate (even without the first harmonic present in the complex harmonic). This common repetition rate is reflected in the perceptual quality called pitch. According to the American Standards Association (1960), pitch is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high." Pitch is a perceptual correlate of a sound's repetition rate - it is an attribute of sensation and perception, not a physical attribute of sound.

As described earlier, the amplitude and phase of the first six to eight harmonics (resolved components) are decoded in the cochlea. Therefore, information about the frequency decomposition of the sound (the frequency spectrum and phase) is represented in the auditory nerve activity. The resolution of this transform is higher in the lower frequency regions; therefore, the resolved lower harmonics form activity patterns that initiate the perception of pitch. In addition to that, higher unresolved harmonics interact in the basilar membrane and produce a complex wave that repeats the fundamental frequency. Although the neural mechanisms of pitch extraction are still a subject of study, pattern recognition models suggest that the combination of information in both resolved and unresolved regions builds the foundations of pitch extraction in the auditory nerve, brainstem, and cortex. (Plack 2018)

1.3 Experimental Auditory Scenes Models

Different paradigms have been developed to model auditory scenes and sequential streaming in experimental settings. Some of these paradigms present multiple concurrent naturalistic sounds (speech, animal sounds, music, etc.) and require listeners to attend to certain auditory objects or detect targets while some property of the objects (e.g., presentation location; Darwin and Hukin, 1999) or the scene (e.g., the number of objects;

Eramudugolla et al. 2005) is manipulated. While these paradigms are excellent candidates for studying higher-level processes, such as auditory attention, the complexity of their stimuli is not necessarily required to study lower-level features.

Another group of (less naturalistic) paradigms that are utilized to study the effect of physical sound properties on perception present sequences of tones to listeners. In these paradigms, listeners either actively adjust tones' physical parameters to maintain a certain percept or report their perception in different ways like pressing buttons, drawing what they hear, counting tones, etc. (Bregman 1990). One such paradigm, introduced by van Noorden (1975), is a simple model that enables researchers to simulate fully controlled auditory scenes. In this paradigm, a train of pure tones with the same frequency (F_A) and duration interleaved with silent intervals of the same duration forms the sequence A. Another sequence of pure tones with the same duration but a higher frequency (F_B) forms the sequence B. B tones overlap with every other silent interval in sequence A. Using "_" notation for silent intervals, "A" for tones of sequence A, and "B" for tones of sequence B, the model could be described as simultaneous presentation of "A _ A _" and "_ B _ _" repetitions. This sequence is also known as van Noorden's ABA triplet paradigm, as illustrated in *Figure 1.6*.

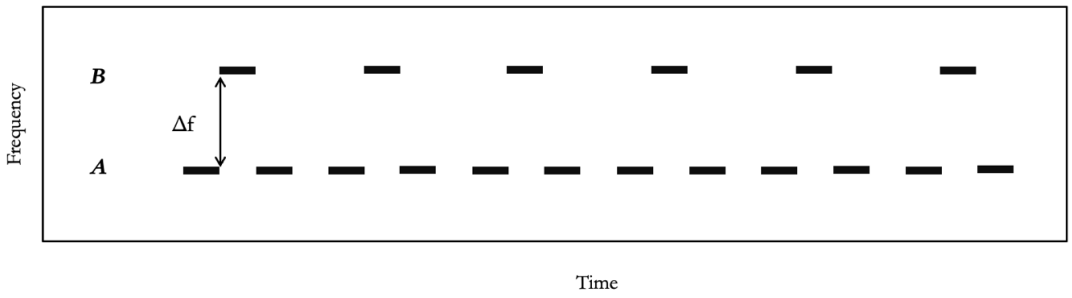

Figure 1.6 An illustration of van Noorden's ABA paradigm with the frequency difference of Δf between the two sequences.

When presented to listeners, the two sequences are either perceptually fused and heard as a cohesive stream, or they are perceived as segregated streams of sound (*Figure 1.7*). The first alternative sounds like a galloping rhythm known as the integrated or "horse" percept. The second percept is the segregated or "morse" percept, as it sounds like two

simultaneous pulsing rhythms like a morse code (with this analogy, this model is sometimes called the horse-morse paradigm).

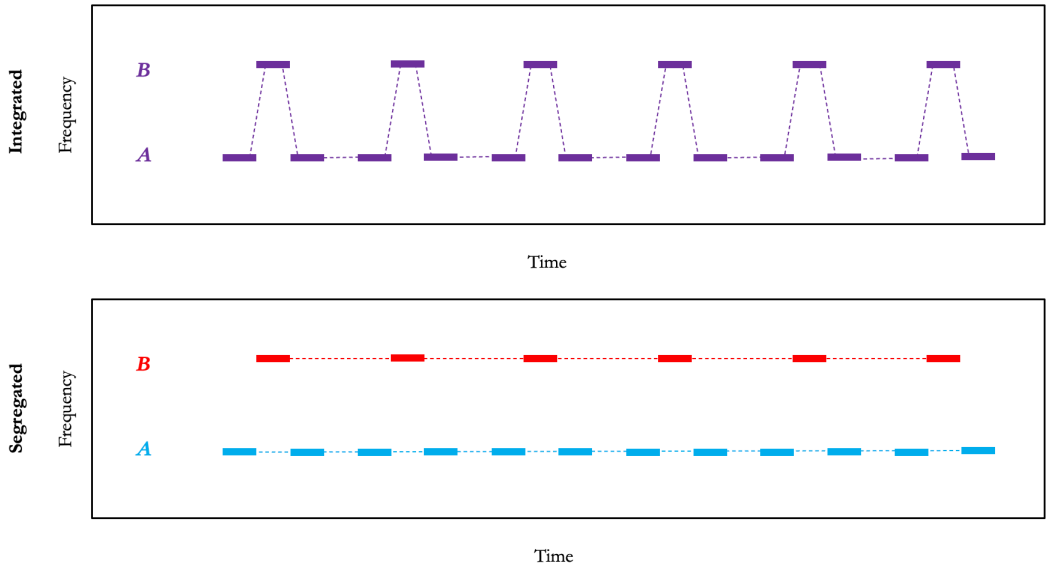

Figure 1.7 The alternative percepts (integrated on the top and segregated on the bottom) in the ABA paradigm are illustrated with dotted lines and colors that resemble the mental crayon analogy used by Bregman.

Van Noorden (1975) had listeners report which alternative they perceived when presented with this experiment and observed that the listener's perception depends on the frequency difference between the two sequences (Δf). Segregation happened more readily at larger (inevitable after a certain point) and less readily at smaller Δf (impossible at some point) values. Increasing the tempo of the sequences had the effect of pushing the perception toward segregation as well. It was also observed that perception is bistable within an intermediate range of values for Δf and tempo (Moore and Gockel 2012), and listeners can switch between the two alternative percepts. Additionally, at faster tempi, lower Δf values facilitate segregation (Carlyon 2004). As shown by Cusack et al. (2004), listeners integrate the two sequences before they attend to them, and sequences must be attended to for a few seconds for the segregated percept to "build up." Therefore, attention plays a significant role in the perception of the ABA sequences.

Several variations of the ABA stimulus and response collection methods have been used to investigate different factors influencing sequential streaming. For example, (Micheyl et al. 2005) modified elements of one or both sequences and implemented “integrated easy” and “segregated easy” targets within the repetitions of ABA triplets to create an objective version of the task. Another variation was used by Boehnke and Phillips (2005) who presented the sounds dichotically (sequence A was presented to one ear and B to the other) via headphones instead of the diotic presentation in the original paradigm (both sequences presented to both ears) to study laterality differences.

In my variation of the ABA paradigm, to study the interactions between pitch and spatial cues, I not only had the A and B sequences differ in frequency, but they were also presented from different locations. This variation is similar to the one used by Boehnke and Phillips (2005) to the extent that the two sequences are in different locations, with correspondingly different spatial cues. However, a fundamental difference is that we test a continuum of locations in an open virtual auditory space rather than just testing the dichotic presentation. Our design will be elaborated more in Chapter 3 after the fundamentals of our spatial manipulations are justified in Chapter 2.

1.4 The Audio Dome

Loudspeakers (LSPs) and headphones are the basic tools to present audio to listeners. An LSP can present a very complex mixture of sound, but all the sound components will be perceived from where the LSP is located. Monaural headphones, which are speakers inserted inside the ear, can also present a complex sound wave to one ear and create the illusion that sound is located inside or very close to that ear. Except for the distance cues that signal amplitude manipulations can simulate, these methods are not very helpful in simulating the other aspects of the spatial hearing experience.

Virtual auditory space technologies expand these basic tools to simulate experiencing sounds from different locations. These technologies are essential for conducting experiments involving spatial auditory experience as a key variable; without them, it would be difficult for the researchers to change the location of sound sources precisely between the experiment trials or to move sound sources in an experiment about sound

movement perception. Virtual auditory spaces facilitate these manipulations by simulating spatial features of sound sources for the listeners.

The first step to overcome the spatial simulation problem of LSPs and monaural headphones is simply to increase the number of channels that deliver sound. Starting with the headphones, another channel could easily be added to make a set of binaural headphones. Presenting the same signal through both channels (mono display) only adds the perceived location of the object on both sides of the body or makes it sound like it is located on the median plane. On the other hand, presenting different signals to different ears (stereo display) makes the listener perceive each sound located on a different side of the body. If the recorded signal from some object is presented with different phases and amplitudes in the ears, the system simulates ITD and ILD cues that lead to the perception of a virtual location for the object. Also, with some additional filtering, the effects of the pinnae, head, and upper-body shapes could be incorporated into the simulation to make a more naturalistic experience of 'spatialized' sound. However, every listener has a unique combination of pinnae, head, and body physiques that might be different from the generic models used for these simulations. Therefore, because listeners do not hear with their "own ears," the hearing experience through stereo headphones will not be as natural. It is possible to customize the model parameters for each listener by estimating individual-specific HRTFs, but it is a time-consuming and somewhat unpleasant process (it usually involves the insertion of a microphone near the eardrum).

Alternatively, several loudspeakers could be placed at different locations to represent the sound sources at those locations. In this scenario, the sound sources are physically located in the desired locations, so all spatial cues are physically implemented; listeners hear with their "own ears." This experience is more natural, but such loudspeaker arrays are not as portable as headphones; they require dedicated physical space, and they are more expensive. But the most significant disadvantage of such loudspeaker arrays is their quantized and relatively low-resolution spatial coordinates that are set by the dimensions and number of the loudspeakers; for example, the best resolution of a thirty-LSP array with each LSP being 15 cm wide on the horizontal plane is 30 locations with 12° distance increments in between, on a circle with a radius of 72 cm around the listener's head (if

LSPs are located beside each other with no gaps in between them). In this example, all the sound sources should be assigned to one of the 30 coordinates, and their distance can only be integer multiples of 12°, which, although it might be enough for specific simulations, has a poor resolution compared to stereo headphones (or the real world!) where the space could be realistically considered a continuum.

The spatial resolution of loudspeaker arrays could be enhanced by distributing the manipulated copies of the sound source signal to multiple channels, similar to the basic idea used in engineering stereo headphones. Vector-Based Amplitude Panning (VBAP) is a popular robust algorithm that implements this idea. To simulate a virtual acoustic source between the two LSPs of a dual channel system illustrated in *Figure 1.8.A*, VBAP projects the vector from the center of the circle (participant's head) to the source location $\vec{\boldsymbol{p}}$ on the unit vectors from the center to the location of the two LSPs $\overrightarrow{\boldsymbol{I_1}}$ and $\overrightarrow{\boldsymbol{I_2}}$. Two coefficients of $g_1$and g_2 are calculated for the unit vectors through the projection process such that the weighted summation of the unit vectors makes the original vector to the virtual source ($g_1\overrightarrow{\boldsymbol{I_1}} + g_2\overrightarrow{\boldsymbol{I_2}} = \vec{\boldsymbol{p}}$). Finally, the virtual source's audio will be played from the two channels with amplitudes proportional to the coefficients ($g_1$and g_2) to create the illusion of a phantom source at the intended location (Pulkki, 1997).

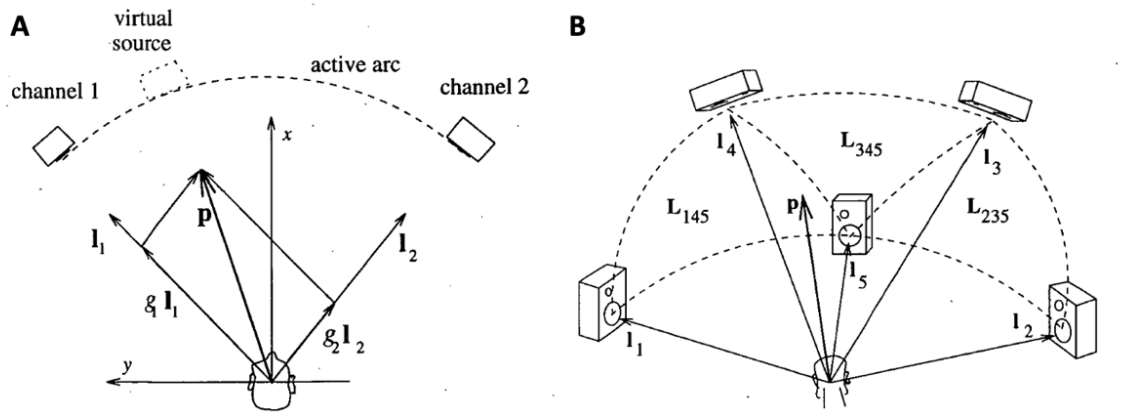

Figure 1.8 A. A dual-channel VBAP system configuration formulated with vectors. B. A configuration of an expanded multi-channel three-dimensional VBAP algorithm formulated with vectors. (Pulkki 1997)

The VBAP algorithm is generalizable to expanded configurations with more channels or to a three-dimensional setting (*Figure 1.8.B*). In such configurations, the smallest subset of required channels for projection is determined prior to rendering each sound source; if the source is intended to be at the location of one of the channels, then that channel will be the only channel used to render the source. If the sound source is intended to be on the arc that connects two adjacent channels, those two channels will be selected to render the sound. Finally, in a three-dimensional setting, if it is not possible to render the intended location with two channels, a subset of the three nearest channels is chosen to simulate the phantom source. For example, the virtual sound source in *Figure 1.8.A* is on the active arc of the two channels that are shown. Therefore, those channels will be used to render it. In *Figure 1.8.B* the space is divided into triangles (shown with $\boldsymbol{L}_{145}$, $\boldsymbol{L}_{345}$, and $\boldsymbol{L}_{235}$) that sound sources can appear on. If one tends to render a source right above source 5, then channels 3, 4, and 5 will be recruited, as $\boldsymbol{L}_{345}$ is the active triangle for that source (Pulkki, 1997). As described, the number of channels recruited to render each sound source with VBAP depends on the channel's layout and the location of the source. Therefore, the spread of energy for a sound source rendered by three LSPs in a three-dimensional setting is not only higher than a sound source rendered by one LSP, but the extent of the spread depends on how the channels are laid out.

Ambisonic panning is another panning method for loudspeaker arrays that equalizes the spread of energy regardless of the sound source location and the loudspeakers' layout. In this panning method, the surrounding sound field is decomposed to spherical harmonics as a set of basis functions. These spherical harmonics represent information about the sound pressure. Therefore, information about the properties of the entire sound field is represented in these systems instead of source location and LSP layouts. Spherical harmonics represent the sound pressure, its velocity in different directions, and their derivatives. The order of the largest derivative used in a system's simulations identifies the system; for example, a 0^{th}-order ambisonic implementation only concerns the sound pressure, while a 1^{st}-order system includes the pressure velocity in different directions, and a 2^{nd}-order configuration involves the derivates of those velocities. (Higher order derivatives contribute to implementing more precise details, but they come at the cost of having more channels.) Finally, the spherical-harmonic representation is transmitted to all

channels (instead of just an audio signal and a modified gain to selected channels), so they each spread their share of plain waves that build up the intended signal (Zotter and Frank 2019).

In this thesis, I aimed to simulate a precisely controlled but natural auditory experience. Therefore, I used a state-of-the-art custom-built 95-channel loudspeaker array capable of 9th order ambisonic panning. This system, known as the "Audio Dome," consists of 4 dual-channel subwoofers and 91 LSPs geodesically arranged in a sphere with a radius of approximately 1.65 m in a sound-attenuating chamber (*Figure 1.9*). In addition to 9th-order ambisonics, this system is capable of VBAP and single-channel rendering of auditory scenes (sonible GmbH, Austria).

Figure 1.9 The Audio Dome in the sound attenuating chamber at Western University. (A moveable part of the frame rotates around a hinge to provide entrance for the listener. Once closed, the eight channels attached to it complete the spherical structure.)

This system was installed in November 2019 in the Western Interdisciplinary Research Building at Western University. This project is the first study that aimed to utilize this device to test human perception using the Higher-Order Ambisonics (HOA) method in the institute. Therefore, it's essential to ensure the stimuli presented with the technology

are accurate enough to represent the tested variables. Theoretically, it is possible to present virtual sound sources with 64-bit floating point precision angle parameters (sonible GmbH, Austria), but because of the blurring effects of the ambisonic method, these detailed differences may not be perceivable. Thus, we need to ensure the audio presentation system has a resolution approximating or exceeding human spatial acuity limits.

As the project's secondary but essential aim, an initial study was designed to validate that the perceived precision of the spatial details rendered by this 9th-order ambisonic system is sufficient to challenge the limits of human spatial acuity as a reassurance proof for further experiments.

1.5 Thesis Objective

Our auditory experiences are initiated by sounds reflecting the surrounding environment. The brain and mind analyze these reflections that enter our ears as a mixture of sounds in order to associate them with external events or auditory objects through a process known as auditory scene analysis. Several sound attributes provide cues for the auditory system to analyze auditory scenes according to the principles of perceptual organization. For example, sounds that are perceived to come from the same location or have a similar pitch are perceptually grouped together and are segregated from the other sound components (Bregman 1990; Carlyon 2004). Different cues can cooperate to accomplish perceptual organization; however, some may be more important than others in determining what is perceived. Spatial and pitch cues have been studied in the context of auditory scene analysis, each by itself. In this thesis, I aimed to characterize the relative perceptual importance of spatial and pitch cues in sequential streaming when they are both provided to the listeners.

I modelled auditory scenes with van Noorden's (1975) ABA paradigm. While this paradigm is most often studied with the sequences presented from the same location, I displaced sounds to introduce spatial cues to the listener. But before that, I assessed the fidelity of the utilized virtual acoustic space to spatial sound reproduction and ensured it was precise enough for human spatial acuity.

This thesis includes two experimental chapters; in Chapter 2, I describe the study held to ensure that using the 9th-order ambisonic technology in our virtual space is accurate enough for future studies, including the study described in Chapter 3. In Chapter 3, using the established technology in Chapter 2, I introduced a space dimension to the ABA paradigm to study how spatial and pitch cues interact in the perceptual space in the context of sequential streaming. In the end, a summary of the results, conclusions, discussions, and future directions for both studies are provided in Chapter 4.

Chapter 2

2 A 9th-order Ambisonic System's Spatial and Spectral Fidelity of Sound Reproduction

Psychophysical experiments aim to characterize the relationship between the physical stimuli and their evoked sensation or perception. The characterization process involves presenting stimuli with different levels of the physical attribute under study, recording the participants' sensory or perceptual responses, and quantifying the relationship between the physical attribute and the responses. Like every other measurement, several sources of noise in the stimulus preparation, stimulus presentation, response collection, and analysis processes increase the inaccuracies of such quantification. It is essential to identify and report these sources of noise before utilizing any new methods (device or algorithm) and characterize the limitations and inaccuracies they introduce to the measurement to consider when designing experiments and choosing effective methods. In this Chapter, the spatial and spectral fidelity of 9th-order ambisonics algorithm in the Audio Dome is assessed to ensure the reliability of this device for human experimentation.

2.1 Introduction

The Audio Dome is a custom-made device designed to simulate auditory scenes in auditory or multisensory perception experiments. This device is capable of single-channel, VBAP, and 9th-order ambisonic panning of sound sources. Of these methods, ambisonic panning theoretically has the highest spatial resolution and an equalized energy spread for all sound sources (Zotter and Frank 2019). Compared to headphones, no HRTF (generic or individual-specific) modelling is required to experiment with this device, and the free-field experience provides opportunities to study complex multisensory behaviors (such as audio-visually guided grasping) more naturally and more easily. However, for the lower-order ambisonic panning systems (1 to 4), a perceptual blurring effect has been reported that enhances as the order of the system increases (Bertet et al., 2013). Also, errors in ITD cue reconstruction from virtual ambisonic sources have been quantified (Neal & Zahorik, 2022), which adds to the potential risks of

this technology. Additionally, the Audio Dome LSPs distribution on the geodesic sphere is optimized for 9th-order ambisonic panning, which is not a uniform spatial distribution; the number of LSPs at some regions of the space (front and back) is higher than in some sparser regions. The manufacturer noted that some extra LSPs on the horizontal, median, and frontal planes were also added for better resolution across different panning methods (OWU 3D audio installation manual, sonible GmbH). Although, theoretically, the ambisonic panning resolution should be independent of the LSP layout, such a vague statement about resolution improvement in the device manual suggested that adding the extra LSPs increases the resolution in ambisonic panning as well. The non-uniform distribution, therefore, raises the concern that the Audio Dome may not reproduce source locations with a similar precision, which makes the device unreliable for accurate stimulus presentation.

In this study, the Audio Dome's 9th-order ambisonic panning was validated for human experimentation regarding the mentioned concerns; the system was expected to reproduce sounds precisely at the specified locations correctly perceivable by human listeners.

The Minimum Audible Angle (MAA) is the minimum angular separation of two sources that can be detected by a listener (Strybel and Fujimoto 2000). For pure tones in free-field experimentation, horizontal MAAs are reported to be about 1° right at the front (0° azimuth) and increase to about 3° at ±60° azimuth and more than 7° at above ±75°azimuth on the sides (Mills 1958; Blauert 1997).

In this study, listeners' horizontal MAAs for broad-band noise stimuli were estimated at the location of the nine horizontal LSPs and their midpoints (eight locations) spanning from -90° to 90° azimuth (17 locations in total; *Figure 2.1*). Perceptible blurring effects as a result of ambisonic panning (Zotter and Frank 2019) would lead to poorer spatial resolution that would be reflected in reduced spatial acuity (increased MAAs) compared to the literature. Additionally, to test the effect of LSP density on ambisonic panning the frontal MAAs, where listeners have the highest acuity, were also estimated when the listeners were facing the region in the dome with the sparsest LSP distributions (±50.41° azimuth). These frontal MAAs were then compared to the frontal MAA when the

listeners were facing a region with the highest density of LSPs (0° azimuth). This comparison would reveal any perceivable effects of the LSP density, layout, and geometry on the phantom sound sources rendered with abmisonics.

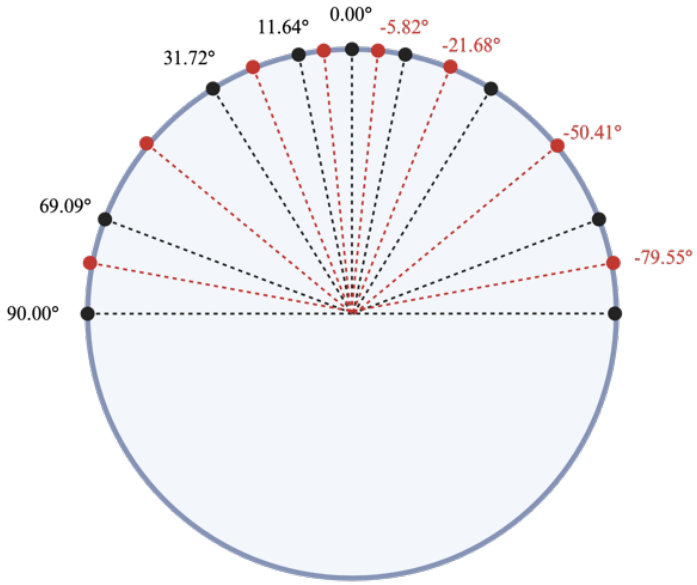

Figure 2.1 The layout of LSPs on the horizontal plane (elevation = 0°). LSP locations are shown in black, and the midpoints are in red. LSPs are located symmetrically around the midline plane (azimuth = 0°). Numbers specify the azimuth coordinate.

Secondly, the spectra of the sound sources rendered at LSP locations (indicated with black dots in *Figure 2.1*) were estimated to quantify potential spectral distortions. To do so, a Head-And-Torso Simulator (HATS, *Figure 2.2*; Brüel & Kjær, Denmark) was utilized. This device models the physical properties of the human ear canals, pinnae, head, and upper body that modify the external sound spectra. Microphones bilaterally implanted in its ear canals allow for precise measurement of received sound (relative to generated sound) that enabled me to evaluate systematic changes in spectral energy as a function of presentation method (single-channel/ambisonics) and location within the Audio Dome.

Figure 2.2 The Head and Torso Simulator (Type 4128C; Brüel & Kjær, Denmark).

The HATS was placed at the center of the Audio Dome, and a series of chirp signals presented from LSP locations (illustrated in *Figure 2.1*) were rendered with single-channel and ambisonic panning methods. The spectrum of the signals recorded by the two microphones inserted in the ears of the HATS estimated the spectrum of the signals as heard by average human adults. Estimated spectra for sources rendered with ambisonics were then compared with the spectra estimated for single-channel rendering (as the ground truth) to identify potential spectral distortions introduced by the ambisonic method in the Audio Dome.

2.2 Materials and Methods

2.2.1 Experiment Setup

The Audio Dome is a 95-channel (four dual-channel subwoofers and 91 flat LSPs) array geodesically arranged in a sphere with a radius of approximately 1.65 m in a sound-attenuating chamber. The electronic parts of the device consist of a rendering server, seven 16-channel Digital to Analog Convertors, fourteen 8-channel amplifiers, one 8-channel controllable microphone preamplifier and Analog to Digital Convertor, power distribution units, and a network switch. The device is programmable with Spatial Audio Creator software that has a Graphical User Interface (GUI). Also, Open Sound Control (OSC) commands can be communicated to the software via the network connection to control the device. (sonible GmbH, Austria)

A version of the horizontal coordinate system is used to address the location of LSPs and sound sources in this system. The center of the coordinate system is the center of the sphere, with the horizontal plane as the fundamental plane (at 0° elevation). Azimuth angle (φ) is used to specify the angle deviation from the reference on the horizontal plane, within the ±180° range, with positive values indexing the counterclockwise and negative values indicating clockwise deviations from the reference (φ = 0°). Elevation angle (θ) specifies the altitude with respect to the fundamental plane, within the ±90° range, with positive values indexing higher elevation and negative values indexing lower elevation than the reference (θ = 0°). The schematic of this coordinate system is shown in *Figure 2.3*. The (φ, θ) notation will be used to specify coordinates in the thesis.

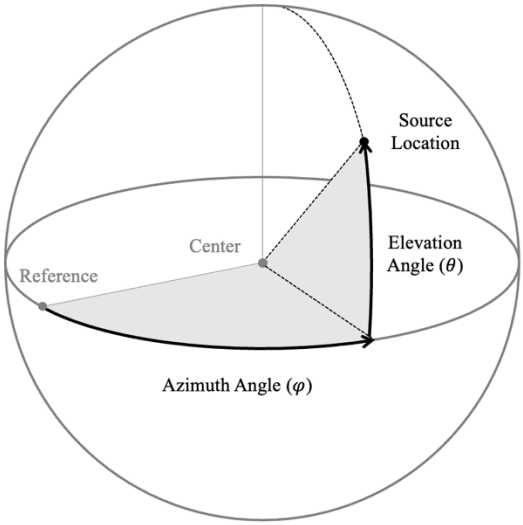

Figure 2.3 Audio Dome's coordinate system

Participants sat on an adjustable chair (with minimal acoustic shadows) that was set to face the reference at the center of the Audio Dome. Then, the height and position of the chair were adjusted such that the participant's ears were at the level of the reference plane aligned with φ = -90° to φ = 90° line, symmetrically around the central point.

To rule out the influence of vision on sound localization (Tabry, Zatorre, and Voss 2013), the experiment was held in total darkness, and all surfaces on the wall and speaker joints were covered with black Velcro to minimize any potential reflections. To ensure the

participants sat still during the experiment, a fixation Light-Emitting Diode (LED) at (0°, 0°) was red throughout each trial. Participants were instructed to fixate on the LED while it was on and try not to move as much as possible. (The position of the head was recorded with the placement of an 8 mm magnetic sensor attached to the back of the participants' heads during each trial. These recorded data were not analyzed, although participants reported anecdotally that the mere presence of the sensor reminded them to remain still and fixate on the LED.)

2.2.2 The Experiment Paradigm

2.2.2.1 Task and Stimuli

On each trial, two broad-band noise bursts (0-22050 Hz, with sampling frequency of 44100 Hz) were presented on the horizontal plane. Broadband noise bursts were used to provide listeners with all spatial cues available across the spectrum. The noise bursts were each 450 ms long, separated by a 200 ms silent gap, followed by a 1250 ms interval for response collection. Highest spatial acuities (i.e., lowest MAAs) were reported with these noise duration and onset asynchrony parameters (Strybel and Fujimoto, 2000). The envelope of each noise burst was shaped with attack and decay cosine^2 ramps (%5 of the beginning and the end of each noise burst duration). The noise bursts were generated in MATLAB 2022b (The MathWorks Inc., 2022) software using the white Gaussian noise function (wgn) with the noise power set to -25 dBW (73.1 dB SPL with background noise level of 25.2 dB SPL when the chamber is the quietest, both measured with a sound-level meter (Brüel & Kjær, Denmark) at the center of the Audio Dome), and were appropriately enveloped and zero-padded (*Figure 2.4.B*). Eighty different noise bursts were randomly generated and pseudo-randomly assigned to trials in the main experiment. A different set of noise bursts was used for the practice blocks.

The two noise bursts were either presented from adjacent (or occasionally the same) locations on the horizontal plane. Participants were asked to report their judgement of the direction of the second burst relative to the first via key press on a small wireless keypad connected to the experiment computer with Bluetooth. They were explicitly asked to judge whether the second burst was located in a clockwise or a counterclockwise

direction relative to the first one. The experiment paradigm and the stimuli used in a sample trial are illustrated in *Figure 2.4*. For each trial, the listener's response and reaction time were recorded.

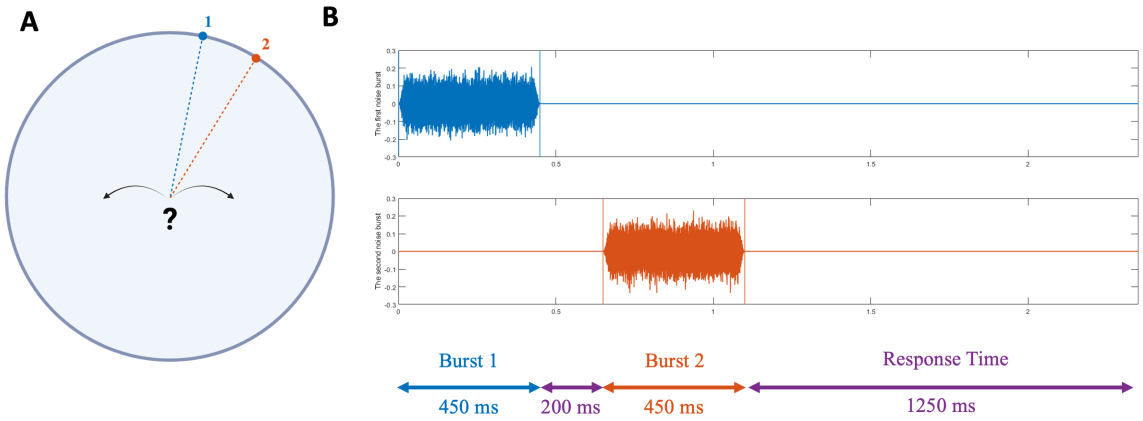

Figure 2.4 A. The experimental paradigm: Two noise bursts on the horizontal plane were presented to the listeners to judge their relative locations. B. The first and the second noise burst signals of a sample trial are illustrated in the top and bottom panels with trail progression below.

Each reference point (*Figure 2.5; Table 2.1*) was tested in an experimental block of 260 trials.

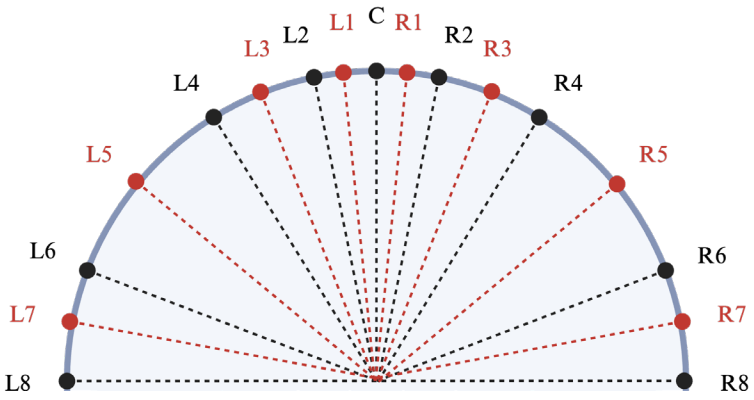

Figure 2.5 Reference point locations and labels on the horizontal plane. Reference points at the location of an LSP are shown in black, and the reference points at the midpoint of adjacent LSP pairs are shown in red.

Reference Point		φ_{ref} (°)		Location Type (LSP/midpoint)	Test Point Range (°)	Test Point Increments (°)
C		0		LSP	±6.00	1.00
L1	**R1**	5.82	-5.82	midpoint		
L2	**R2**	11.64	-11.64	LSP		
L3	**R3**	21.68	-21.68	midpoint		
L4	**R4**	31.72	-31.72	LSP		
L5	**R5**	50.41	-50.41	midpoint		
L6	**R6**	69.09	-69.09	LSP		
L7	**R7**	79.55	-79.55	midpoint	±9.00	1.50
L8	**R8**	90.00	-90.00	LSP	±12.00	2.00

Table 2.1 A list of the reference points, their coordinates, and associated test points. (In Reference Point, column C refers to the Center, L to the left, and R to the right. Location Type indicates whether that reference point is at the location of an LSP or it is at the midpoint between two LSPs. For each reference point, a range of test points was tested with different increments that are noted on the two last columns.)

In each block, a noise burst at the reference point was presented before (50%) or after the noise burst at each of 13 test points (6 to the left of the reference point, 6 to the right, and the reference point itself). Each pair was tested 20 times, but the order in which trials were presented in each block was randomized. Eighty different noise bursts were randomly generated and pseudo-randomly assigned to trials with an equal number of appearances across reference-test point pairs (a different set of noise bursts was used for the practice blocks).

2.2.2.2 Experimental Conditions

Participants completed the task in three conditions. In the first condition, they sat facing towards (0°, 0°) and performed the horizontal location discrimination task for 17 reference points spanning from φ = -90° to φ = 90°. Nine reference points were at the location of an LSP on the horizontal plane, and the other eight were at the midpoint between adjacent LSP pairs (*Figure 2.5*).

Each trial tested one reference point and an associated test point around it. Test points for each reference point were 13 locations, consisting of the reference point's location and six symmetrical locations around it on the horizontal plane. A larger range of test points was examined for the two most lateral reference points on each side (L7, L8 and R7, R8 in *Figure 2.5*) to capture the larger discrimination range expected to be required for these reference points. The reference points, their labels, and their associated test points are listed in *Table 2.1* (symmetrical reference points are listed in the same row with the positive φ values for the locations on the left and the negative φ values for the locations on the right).

The LSPs are most dense around (0°, 0°) coordinates (three LSPs on the horizontal plane within a ±12.00° range), whereas they are sparsely distributed around locations such as (50.41°, 0°) where the closest horizontal LSPs are located 18.68° away. To assess the homogeneity of the simulation (ambisonic panning) precision across regions with different densities of LSPs, in two more conditions, listeners sat facing L5 and R5 locations and repeated the task (just for one reference point at each location, in front of them). Measurements of these "rotated" conditions are comparable with the measurements of the center-fixated condition at reference point C because the listeners' spatial acuities are similar and at their highest in these conditions. Hence, behavioral differences in these conditions would reflect 9^{th}-order ambisonic panning dependency on the system's geometry (LSP density).

2.2.2.3 Blocks and Sessions

The 19 blocks (17 for the center fixated and 2 for the rotated conditions) were split into two sessions: 4 symmetrical pairs of reference points in session 1 and the other four pairs with the rotated conditions and reference point C in session 2. The order in which the reference points were tested was randomly assigned to both sessions before the experiment began for each listener. Also, the order in which reference points C, 5L and 5R in the rotated conditions were tested was pseudorandomized across participants. At the beginning of each session, listeners completed a practice block of 20 trials with noise bursts at the central or the left LSP locations. Participants were supposed to correctly indicate the direction of the second noise burst relative to the first when these bursts were presented from locations C vs L6, C vs L8, and L2 vs L8 since these are relatively large discriminable separations. Accurate performance on all these practice trials demonstrated that they had learned the task before they continued.

At the beginning of the first session, listeners completed a questionnaire about their demographics, musical training, and hearing and neurological backgrounds after their participation consents were obtained. Then, an audiometry test assessed the listeners' pure tone hearing thresholds at 125 Hz and at octave frequencies up to 8 kHz (exclusion cut-off threshold $\geq$ 25 dB HL for any frequencies at any ear), followed by recording their head measurements. Participants completed a post-experiment questionnaire about their motivation and attention during the task at the end of the second session.

2.2.2.4 Instructions and Feedback

A sample pair of noise bursts at widely separated sources (C vs. L8) was first presented to each participant. After they confirmed they had understood the task, they practiced a few more samples with moderate and large distances and with a sample of collocated bursts. Verbal feedback (correct/incorrect) was provided on these trials. The purpose of the trial with collocated stimuli was to remind the listeners that they should always respond, even if they find the condition challenging (feedback for this condition was always provided as incorrect). When they confirmed they had understood the task, they completed the practice block. No feedback was provided after they passed the practice

block assessment, except they were told if they missed responding to any trials at the end of each block.

The listeners were reminded to keep their eyes open, fixate on the LED, and try not to move at the beginning of each block. Their head position and seat height were adjusted at the beginning of the session and before the beginning of each block (especially before and after the rotation blocks). Also, the listeners were told that all noise bursts would be presented at locations on the horizontal plane that might not necessarily match an LSP's location. Finally, listeners were allowed to change their response on the current trial up until the next trial started. In case multiple responses were acquired, only the last one was kept and used in the analysis.

2.2.3 Participants

Six young (aged 18-25 years), normally hearing (tested with audiometry), right-handed listeners (four female) with no reported hearing or neurological abnormalities participated in this experiment. Three participants had not had any musical training, and the other three had some basic musical experience but were all out of practice for at least two years at the time of their participation. (One other participant marginally passed the audiometry test, and their performance in the first session's practice block was not acceptable after three tries. Hence, they were excluded from the study.)

2.2.4 Analysis

2.2.4.1 Preprocessing

For each block, the trials with no response were removed. Blocks with more than 5% missed trials in total (i.e., 13 missed trials in a block of 260 trials) were excluded from further analysis.

2.2.4.2 Psychophysical Modelling and MAA Estimations

The response (clockwise/counterclockwise) and the reaction time were recorded for each trial. For all reference-test point pairs, the ratio of the responses that indicated that the test point was perceived counterclockwise compared to the reference point was used as an

estimation of the probability of such judgement P($\varphi_{test} > \varphi_{ref}$). This probability ideally should be equal to 1 for the test points counterclockwise from the reference and 0 for the other test points (including those collocated with the reference point). However, like other psychophysical behaviors, this transition is not as sharp as the ideal scenario; it follows a gradual trend, and its smoothness depends on the participant's sensitivity or, in this case, spatial acuity. This probability measure was defined in such a way that aligned with the psychophysical functions and hence was modelled by a sigmoid function as the psychometric model (ψ) for each reference point:

$$\psi_{ref}(\varphi) = \frac{1}{1 + e^{\frac{-(\varphi - \alpha_{ref})}{\beta_{ref}}}}$$

Equation 2.1 The psychometric sigmoid model

In this model α_{ref} was the estimated Point of Subjective Equality (PSE or the %50 threshold, ideally equal to φ_{ref}) and β_{ref} estimated the spread of the psychometric curve around the PSE. The Just Noticeable Difference (JND), or the difference between the estimated 75% and 50% thresholds of the modelled psychometric curve, which is an indirect measure of β_{ref}, was used to estimate MAA at the location of the model's reference point (Gescheider 1997). These model variables are illustrated in *Figure 2.6*.

2.2.4.3 Model Fitness Quality Measure

When the observed ratios followed a general descending trend (as a function of test point location), the model fitted as a flat line, and it was not possible to estimate the MAA (no unique 50% and 75% thresholds were possible to estimate). Such blocks were excluded in the group average estimation of MAAs.

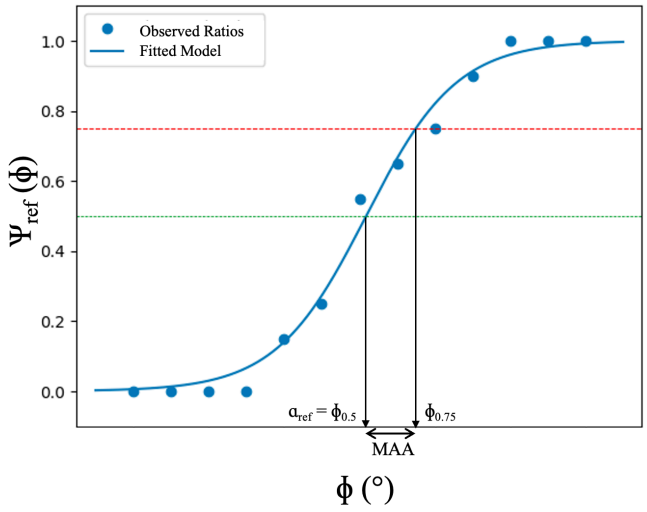

Figure 2.6 An example of a psychometric model fitted to some observed data with 50% and 75% thresholds and the estimated MAA.

For the defined logistic psychometric model, an explained variance index ($EV_{\psi_{ref}}$) was inspired by the coefficient of determination (R^2) for linear models as below to evaluate the validity of the psychometric functions and the model fitness quality.

$$EV_{\psi_{ref}} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{P}\left(\varphi_{test_i} > \varphi_{ref}\right) - \psi_{ref}\left(\varphi_{test_i}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{P}\left(\varphi_{test_i} > \varphi_{ref}\right) - \overline{\mathrm{P}\left(\varphi_{test_\iota} > \varphi_{ref}\right)}\right)^2}$$

Equation 2.2 The model fitness quality measure

The fitness quality measure is formulated in *Equation 2.2* in which i indexes the n test points around the reference point, $\mathrm{P}\left(\varphi_{test_i} > \varphi_{ref}\right)$ is the ratio of responses indicating test point i being located at a higher azimuth level that the reference point (the observed probability), $\overline{\mathrm{P}\left(\varphi_{test_\iota} > \varphi_{ref}\right)}$ is the average $\mathrm{P}\left(\varphi_{test_i} > \varphi_{ref}\right)$ across observations, and $\psi_{ref}\left(\varphi_{test_i}\right)$ is the probability value estimated by the fitted model. $EV_{\psi_{ref}}$ values closer to the upper limit (1) indicate high-quality sigmoid model fitness to the behavioral data, whereas the smaller values reflect poorly fitted models. Lower values of $EV_{\psi_{ref}}$ identify random responses or wrong responding strategies. Therefore average $EV_{\psi_{ref}}$ across

participant was used to identify challenging reference point locations. (All listeners reported a moderately high to highly maintained concentration during the task in the post-experiment questionnaire. Hence, lack of attention was not supposably the reason for apparently random or incorrect responses.)

2.2.4.4 Group Analysis

Individual MAAs were averaged for each reference point in the first condition. The MAAs for reference points L5 and R5, between the original and the rotated conditions, were compared with a paired-sample t-test. The differences between these MAAs and the MAA for reference point C in the first condition were statistically tested with repeated-measures analysis of variance (Repeated Measures ANOVA).

2.2.5 Estimation of the Spectral Effects of 9th-order Ambisonic Simulations

2.2.5.1 Measurement Setup

Head-And-Torso Simulator (HATS) type 4128C (*Figure 2.2 & Figure 2.7*; Brüel & Kjær, Denmark) is a manikin that simulates the acoustic properties of an average adult's upper body, head, and ears. These effects and the effects of the panning method on sound reproduction could be characterized by a set of cascaded linear systems that vary with the location of the sound source. The frequency characteristics of the combined system would reflect the phase and amplitude changes that form the monaural and binaural location cues. These system characteristics were estimated by calculating the HATS's frequency response to a broad-band signal at different locations produced by activating a single LSP (channel) and by ambisonic panning methods. The HATS model was placed on a tripod at the center of the Audio Dome, facing (0°, 0°), with the ears aligned on the φ = -90° to φ = 90° line on the horizontal plane, symmetrical with respect to the median plane and responses were recorded using two ½" microphones in its ears (*Figure 2.7*).

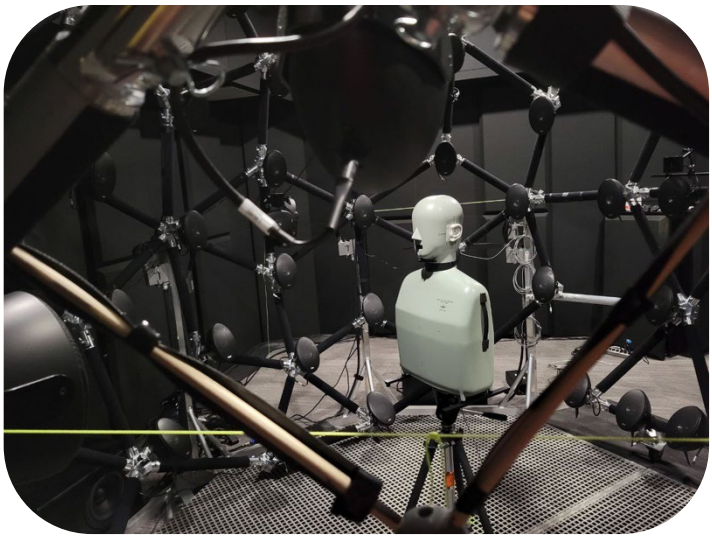

Figure 2.7 The HATS recording position in the Audio Dome

To quantify the effects of the spatial panning methods on the spectra of the sounds and the binaural information, a series of chirp signals were presented, and the sound in the HATS ears was recorded to represent the model's responses. The chirp signals were linear frequency sweeps starting from 0 Hz to 22050 Hz in 22.05 seconds with a constant amplitude of -20 dB (*Equation 2.3*). The signal was generated and presented at 44100 Hz sampling rate.

$$x(t) = 0.1\sin(1000\pi t^2)$$

Equation 2.3 The chirp signal formulation

The sweep was rendered at the location of LSPs on the frontal half of the horizontal plane (*Table 2.1*, type "LSP" reference points). For each rendering method, a sound source was placed at φ = -90°, the sweep audio was presented four times (with pauses in between), and the sound source was moved to the next location. This process was repeated for all listed locations (in ascending order). Signals associated with each panning method were recorded in one take without interruption. Because of the technical difficulties of communicating synchronized pulses between the recording microphones and the Audio Dome server, a 5 kHz pure tone enveloped with cosine^2 attack and decay ramps was used as the trigger signal. These triggers were always presented at φ = 0° for 250 ms, followed

by 750 ms of silence before the beginning of the chirp audio presentation (*Figure 2.8*). The trigger onsets were later detected by finding the peaks of the cross-correlation function between the trigger audio file and the average of the recorded signals in the HATS two ears. Using the trigger onsets, the onset and offset of the chirp presentation were calculated and used to trim the audio files for each repetition.

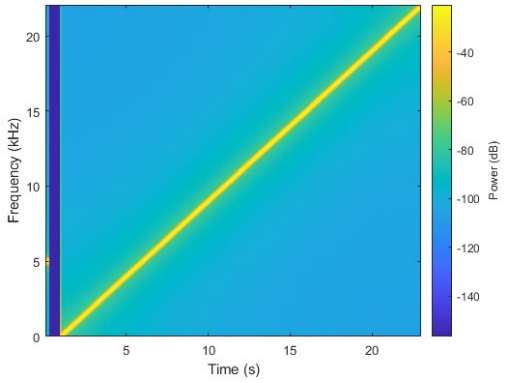

Figure 2.8 The spectrogram of the presented audio files in each trial: a 250 ms long 5 kHz pure tone was played at (0°, 0°), 1 s before the chirp sweep onset.

2.2.5.2 Frequency Response Estimation

The "pspectrum" function, with parameters set to *fs = 44100*, *type = 'power'*, *FrequencyLimits = [0, 22050]*, *Leakage = 0.5* (default value was used for the other parameters) in MATLAB 2022b (The MathWorks Inc., 2022) was used to estimate the Power Spectrum Density (PSD) function of the chirp signal and all recorded responses. Using these PSDs, the amplitude of the system frequency response was estimated using *Equation 2.4*, in which $|Y_{\varphi}^{i}(f)|^2$ is the PSD of the i-th recording for the source located at φ, $|C(f)|^2$ is the chirp signal's PSD, and $\left|H_{\varphi}(f)\right|$ is the system's frequency response amplitude.

$$\left|H_{\varphi}(f)\right| = \frac{1}{4}\sum_{i=1}^{4}\sqrt{\frac{|Y_{\varphi}^{i}(f)|^2}{|C(f)|^2}}$$

Equation 2.4 Frequency response amplitude estimation equation based on the input and recorded signal power spectra.

2.3 Results

2.3.1 9th-order Ambisonics Horizontal Minimum Audible Angles

Excluding the flat estimated psychometric functions, average MAAs are as small as 0.95° at the frontal reference points and diverge to higher values as large as 21.78° moving to the sides. The MAA values are illustrated in *Figure 2.9* and summarized in *Table 2.2* with the number of participants included in each estimation.

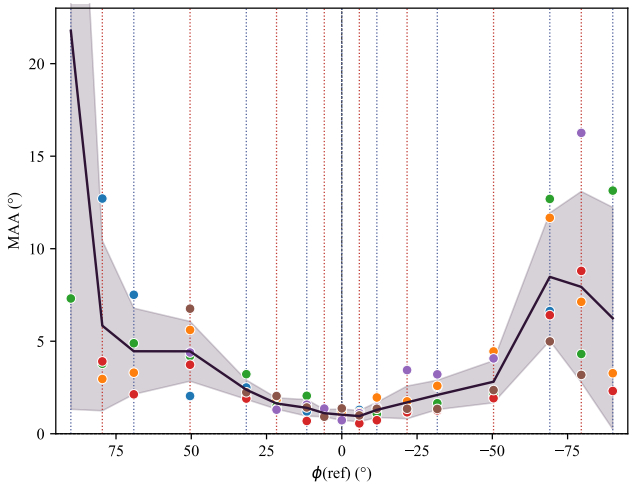

Figure 2.9 Horizontal MAAs for sources rendered with 9th-order ambisonic panning at LSP (blue dashed lines) and midpoint locations (red dashed lines). Individual listeners' data are shown with color-coded dots. The black line shows the average MAAs, with the shade illustrating standard deviations.

As shown in *Figure 2.9* and *Table 2.2,* the variability of the MAA estimates is very low at the frontal locations and increases as the reference point is further away from the center. (The brown color-coded listener's MAA at φ = -90° was 36.24° that is not visualized in *Figure 2.9* for spacing reasons.) Also, the number of listeners with flat psychometric functions (hence, it was not possible to estimate an MAA for them) increases towards the lateral locations. The average explained variance for each of the models is illustrated in *Figure 2.10* (all listeners included). This demonstrates that not only do the models fit better in more central locations, but variability across participants is dramatically reduced relative to more lateralized locations.

Reference Point		MAA ± *SE* (°)		N	
C		1.03 ± *0.12*		6	
L1	R1	1.11 ± *0.08*	0.95 ± *0.12*	6	6
L2	R2	1.40 ± *0.18*	1.29 ± *0.17*	6	6
L3	R3	1.64 ± *0.12*	1.69 ± *0.36*	6	6
L4	R4	2.38 ± *0.23*	2.10 ± *0.32*	5	6
L5	R5	4.46 ± *0.66*	2.81 ± *0.47*	6	6
L6	R6	4.46 ± *1.16*	8.48 ± *1.55*	4	5
L7	R7	5.84 ± *2.30*	7.94 ± *2.31*	4	5
L8	R8	21.78 ± *14.47*	6.24 ± *3.46*	2	3

Table 2.2 Average MAAs, their standard errors (SE), and number of listeners used to estimate each (N, max = 6)

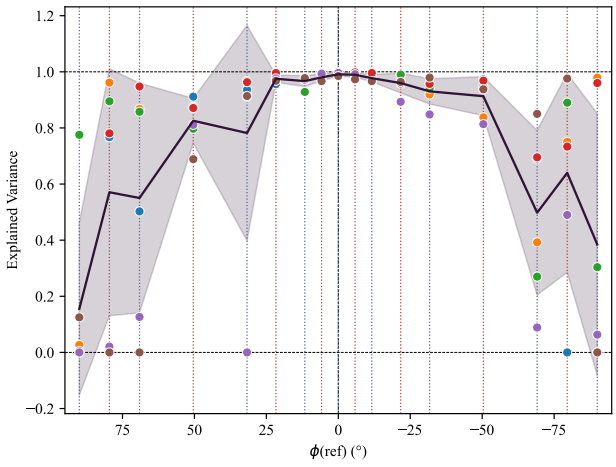

Figure 2.10 Modelling explained variances at LSP (blue dashed lines) and midpoint locations (red dashed lines). Individual listeners' data are shown with color-coded dots. The black line shows the average EVs, with the shade illustrating standard deviations.

Surprisingly, all listeners reported that sometimes they heard noise bursts coming from well above or below the horizontal plane when they performed the task for (some or all) lateralized reference points, namely L6, R6, L7, R7, L8, and R8. This was one reason why I conducted the second half of this study, examining whether ambisonic panning artefactually introduced spectral shaping consistent with pinna cues for elevation at these or other locations.

2.3.2 Frontal Minimum Audible Angles at Different Locations with Variable LSP Densities

Average horizontal MAAs at reference points L5 ($1.20° \pm 0.08°$, $N = 6$) and R5 ($1.17° \pm 0.12°$, $N = 6$) in the rotated conditions were smaller than their respective MAAs in the original condition, in which participants were facing towards the front (one-tail paired-sample T-tests; L5: $t(5) = 5.13$, $p < 10^{-2}$ & R5: $t(5) = 3.63$, $p < 10^{-2}$). These values and the MAA at reference point C (in the original condition) were not statistically different (three-condition repeated-measures ANOVA; $F(2,10) = 0.8275$, $p = 0.4650$).

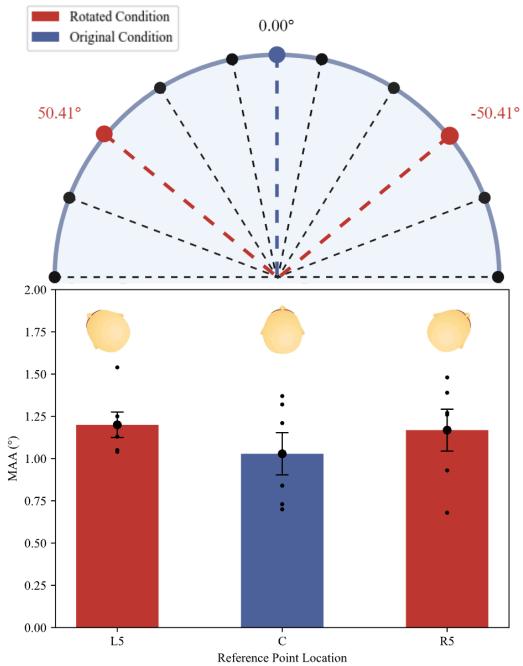

Figure 2.11 Average MAAs (large dots) when participants were faced at different reference points with variable densities of LSPs around them. Individual data are shown with small dots. Error bars show standard errors.

2.3.3 Spectral Effects of 9th-order Ambisonic Simulation

Frequency responses measured from microphones in the "ear canals" of the HATS for the chirp presentations with the single-channel and ambisonic methods are illustrated in *Figure 2.12*. For both methods, responses estimated for the left ear (channel) to chirp sweeps located at every position at the left are similar to the responses of the right ear (channel) to chirp sweeps located at the symmetrical position and vice versa.

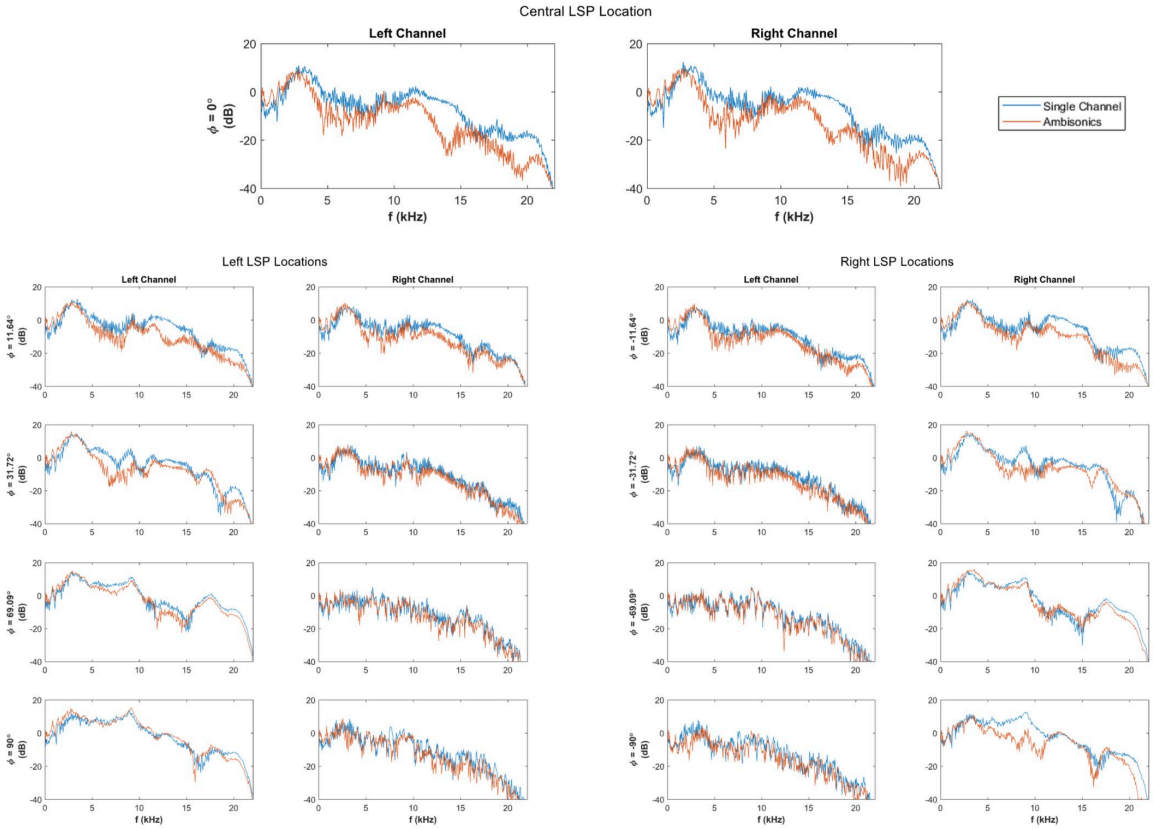

Figure 2.12 Estimated frequency responses for sound sources presented with single-channel and ambisonic methods at the location of LSPs in the frontal half of the horizontal plane.

2.4 Discussion

In this experiment, horizontal MAAs for sound sources rendered with ambisonics were estimated at 17 locations at the front. The estimated MAAs were consistently low around the midline and increased as the sources were moved to the sides. MAA values observed

here are consistent with previous reports of free-field MAAs for human listeners with single-channel sound presentation: MAAs for reference points within the ±25° azimuth range match with values reported by Mills (1958) measured by 500 Hz and 1000 Hz pure tones. For the other reference points, the MAAs are generally estimated to be higher than Mills reports, but the difference is not drastic. Also, the frontal MAAs in all conditions are smaller than the frontal MAAs reported by Strybel and Fujimoto (2000) in free-field listening to high-pass pink noise in a similar paradigm. These results were reassuring since they indicate that sound sources rendered with 9^{th}-order ambisonic panning are spatially resolved and afford human spatial acuity similar to that of natural free-field environments, at least for broadband sounds presented on the horizontal plane.

Additionally, the average MAAs estimated for left and right symmetrical lateralized reference points seem to differ. This asymmetry might be explained by the small sample size of the experiment and the uncertainty that increased as more listeners were excluded for estimating these MAAs. Also, participants anecdotally reported that they heard some extra noise at random times during the experiment. It was clarified later that the computer that controlled the head-tracker system and recorded its data, located outside of the Audio Dome near R7 location, had its fan turned on to cool down the system. Another explanation for the asymmetrical MAAs is this extra noise intervening with sounds presented during blocks or trials testing nearby locations; these noises could turn into an extra localization cue for the listener, which enabled them to compare each trial's noise burst locations with the location they perceived from the computer fan and then judge the relative location of those bursts based on their distance from this extra sound. Although the head-tracking data was collected for all six listeners of this experiment (to match their experience), I stopped head-tracking for the second experiment to avoid this issue. In addition to the elevation cue reports, this was another reason that data collection for this experiment stopped after six participants so both problems could be solved before further testing.

In this experiment, the ability of psychometric models to explain responses was well maintained for some listeners and drastically dropped for others at the lateralized reference points. This reflects the individual variability in localizing sound sources in

natural free-field experiences, which is another piece of evidence showing the importance of “listening with one’s own ears,” even in virtual acoustic spaces.

Next, listeners' frontal MAAs at virtually rendered locations with a sparse LSP distribution were compared with the frontal MAA rendered at the location of an LSP where there was the highest density of LSPs. As these values were not statistically different, it is concluded that there are no perceptually detectable differences in the spatial precision of 9^{th}-order ambisonic rendering at locations with different LSP distributions.

Finally, using a head and torso simulator (HATS), the amplitude of the frequency responses of the human average ear, head, and body to a chirp sweep at the locations of horizontal LSPs were estimated for the single-channel and ambisonic panning methods. Frequency responses for the single-channel method mimic the characteristics of actual human HRTFs (Blauert 1997) with peaks at around 2-5 kHz and attenuation for frequencies above 15 kHz. The frequency response at each position to the chirp sources rendered with ambisonics was compared with responses to sources rendered with single-channel. Visually comparing the two sets of responses, differences between the rendering methods are observable. However, these differences are more highlighted in higher frequencies, while they are not as drastic for lower frequencies (below ~5 kHz). Additionally, the difference between the responses tends to be modulated by the location of the sound source in the higher frequency range, most dominantly appearing as attenuations in ambisonic compared to single-channel rendering. These spatially dependent extra attenuations can affect the spectral balance of the heard sound and level-dependent spatial cue, both monaural and binaural, leading to incorrect elevation and azimuth localization. Because the frequency response patterns for below ~5 kHz are consistent across locations, a (universal) inverse filter can be designed and applied to the audio files before being presented with ambisonic rendering to compensate for the differences. However, designing such an inverse filter for higher frequencies is not possible because the distortions vary with the location of the sound and response to the base response (response to single-channel presentation) is not available for locations other than those at locations of LSPs.

The spectral modifications introduced through ambisonic rendering may have resulted in the percepts that all listeners reported for laterally presented sounds – listeners reported perceiving elevation cues that were not intentionally implemented in the experiment. This would be worth investigating in a follow-up experiment. An elevation discrimination task (similar to the azimuth discrimination task explained in this experiment) using two sets of high-frequency and low-frequency noise bursts at the lateral locations could both characterize the subjective elevation reports and clarify whether the illusory elevation of the sources is a result of the observed differences in frequency responses. Such effects might explain some portion of the variability in MAA estimates at lateralized locations. The MAA experiment should be replicated with a set of low-frequency noise bursts for more reliable estimates.

Another key observation in the spectral responses is the fluctuations of the response curves. These patterns that make the responses look noisy could be explained by reverberation effects while recording the data. Although the room was designed to reduce as much reverberation as possible, minor reverberations can have such effects. To avoid these reverberations, a robust impulse response estimation protocol such as Minimum-Length Sequence (MLS) should be utilized (Tominaga et al. 1975). This technique requires perfect synchrony between the recording device (HATS microphones) and the audio presentation device (Audio Dome LSPs) to succeed (Farina 2000). As synchronizing the digital clocks of the two systems faced a few challenges, this method was not implemented as I initially intended. Once this requirement is satisfied, this method should be applied for more robust frequency response estimations (that could directly be calculated using the estimated impulse responses). In addition, more reliable ITD and ILD cues could be estimated from recordings when the HATS and the Audio Dome are synchronized. This estimation will be helpful in a more detailed characterization of the spatial cues, even if they are not directly contributing to the elevation illusion that listeners reported.

In conclusion, this experiment showed that the spatial resolution of virtual sound sources, rendered with 9^{th}-order ambisonic technology in the Audio Dome, affords human spatial acuity consistent with that observed in the free field, regardless of the system's LSP

layout. Additionally, the spectral effects of ambisonic rendering seem to be consistent for low-frequency (below ~5 kHz) signals. In combination, these results promise high fidelity of the system's 9^{th}-order ambisonic panning to reproduce accurate virtual lower-frequency acoustic sources.

Chapter 3

3 The Interactions of Spatial and Pitch Cues in Auditory Scene Analysis

In this chapter, ABA sequences with variable pitch differences and presentation locations are presented to listeners to test their perception as both spatial and pitch cues are provided to them in this basic paradigm. After the reliability of the Audio Dome for testing humans was evaluated in Chapter 2, it was used to simulate auditory scenes in this study.

3.1 Introduction

Spatial and Pitch cues are among sound component attributes used by the auditory system to analyze auditory scenes; for example, sound components that originate from the same location tend to come from the same source, while sound components perceived from different locations are assumed to be initiated by different sources. Therefore, components coming from the same location are grouped together and segregated from those coming from other places. The same principle applies to the sound pitch in the formation of the perceptual organization (Bregman 1990).

Van Noorden's ABA triplets (1975) is a popular model for studying the effect of pitch cues in sequential streaming. In this paradigm, two interleaved sequences of pure tones with different pitches named A and B (B has a higher pitch than A) are presented to the listeners. When the pitch difference between the sequences (Δf) is small, they are perceptually grouped or "integrated" as a coherent rhythmic auditory stream; alternatively, at higher pitch differences or faster presentation tempi, the sequences are "segregated" and perceived as two simultaneous streams of pulses (Carlyon 2004; van Noorden 1975; Bregman 1990). It has been shown that for a certain range of Δf less than the Δf that yields immediate segregation, attention promotes segregation. At first, the listener perceives the ABA triplets as integrated; then, the segregated percept gradually forms within a few seconds. Also, in the intermediate values of Δf, both percepts are

accessible to listeners who can alternate between them willingly (Moore and Gockel 2012; Cusack et al. 2004).

A and B sequences are often presented through a loudspeaker or a pair of headphones in mono mode. One exception is the Boehnke and Phillips' study (2005), which showed that, when A and B sequences are delivered to different ears on a pair of stereo headphones (dichotic presentation), they are segregated. In the present study, A and B sequences were also displaced, but instead of dichotic presentation, they were presented on the horizontal plane at a range of different azimuth distances ($\Delta\varphi$), symmetrically around the median plane, in a free-field setting (*Figure 3.1*). This approach provided richer spatial information and aimed to investigate how space as a continuum, in combination with pitch cues, affects perceptual organization.

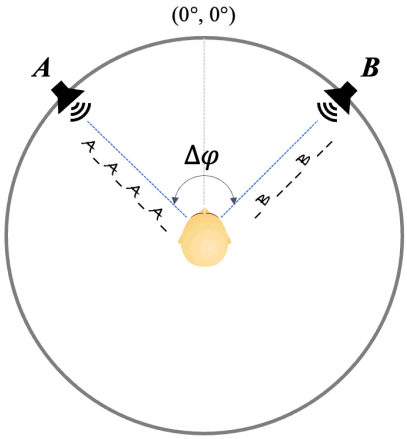

Figure 3.1 Top view schematic of displaced A and B sources on the horizontal plane

To account for individual differences in sensitivity to spatial and pitch cues, I first determined 84.1% thresholds for 1) pitch differences (for collocated sounds, Δf^*); and 2) spatial differences (with A and B at the same frequency, ($\Delta\varphi^*$). When pitch differences and spatial differences are simultaneously present, segregation was expected to be observed at pitch differences less than observed in 1) and at spatial differences less than observed in 2). I used a transformed adaptive up-down paradigm (Levitt 1971) for each parameter. Pitch and spatial differences between 0 and these 84.1% threshold values formed the coordinates of an individual-specific 2-dimensional psychophysical space.

Then, listeners were tested using the method of constant stimuli (Gescheider 1997) for their perception (segregated or integrated) of the ABA sequence at 49 coordinates within this 2-dimensional space.

Finally, inspired by the one-dimensional sigmoid function used in classic psychometric experiments (Gescheider 1997), a two-dimensional psychometric function was used to model the probability of segregation as a function of $\Delta\varphi$ and Δf. This model was defined as an extension of the one-parameter logistic regression model by introducing regressors for $\Delta\varphi$, Δf, and their numerical product (interaction). In the end, to compare the perceptual weights of spatial and pitch cues, the coefficients estimated for $\Delta\varphi$ and Δf parameters were compared at both individual and group levels.

3.2 Materials and Methods

3.2.1 Experiment Setup

To present sound sources flexibly from a range of locations, the experiment was conducted using 9th-order ambisonic panning in the Audio Dome, which I demonstrated in Chapter 2 to be reliable virtual acoustic technology for sound reproduction for the frequency range used in this experiment (below ~5 kHz). Similar to the experiment explained in Chapter 2, listeners sat in the middle of the Audio Dome, with their heads and ears adjusted at the center. The experiment was held in darkness, and the only source of light was the LED at (0°, 0°), which listeners fixated on throughout the trials.

3.2.2 The Experiment Paradigm

3.2.2.1 Task and Stimuli

On each trial, 32 repetitions of van Noorden's ABA triplets (1975) were presented to the listeners. A and B tones were each 125 ms long, and each triplet was followed by a 125 ms silent gap before the next triplet started (500 ms for each triplet; 16 s in total). The tone lengths were adopted from Cusack et al. study (2004) to set an intermediate value for the sequences' pace to make both percepts accessible. The frequency of A tones was always set to 400 Hz. The frequency of B tones was determined based on the Δf value for each trial. Δf ranged from 0 to 12 semitones (400-800 Hz) with 0.25 semitone

increment resolution. All A and B tones were enveloped with attack and decay cosine^2 ramps (5% of the beginning and the end of each tone's duration). To avoid perceptual carryover effects between trials, they were separated with 2 s inter-trial intervals during which no sound was presented to the listener. All the audio files were generated in MATLAB 2022b (The MathWorks Inc., 2022) software with an amplitude of -20 dB and a sampling rate of 44100 Hz; the same sampling rate was used to render sounds with the Audio Dome. Two repetitions of the triplets are illustrated in *Figure 3.2*.

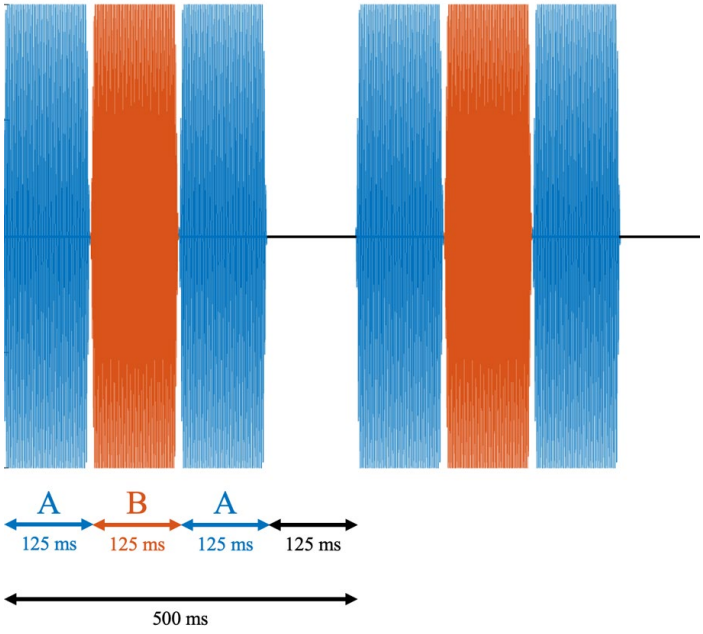

Figure 3.2 Two (of 32) repetitions of the ABA triplet were presented to listeners on each trial. The A and B tones are shown in blue and red, respectively. Each tone was 125 ms long, and the triplets were separated by a 125 ms silent gap.

The A and B sounds were presented on the horizontal plane, at symmetrical locations with respect to the median plane and azimuth angle difference of $\Delta\varphi$ (i.e., they were located at ($\pm\frac{\Delta\varphi}{2}$, 0°) position) with $\Delta\varphi$ ranging from 0° to 360° (*Figure 3.1*). In all trials with non-zero $\Delta\varphi$ values, the side (left/right) on which the A tones were located was randomly chosen with an equal probability of 0.5; B tones were presented from the other side.

At the beginning of the experiment, each listener completed two adaptive procedures that estimated two thresholds (84.1%) that captured their dynamic range of responses as a function of $\Delta\varphi$ and Δf. Then the range between 0 and each threshold was divided into seven equally spaced levels (including 0 and the threshold value) that pairing them as $[\Delta f, \Delta\varphi]$ made a seven-by-seven two-dimensional coordinate space (*Figure 3.3*). Triplet sequences reflecting each coordinate were each presented to the listeners ten times.

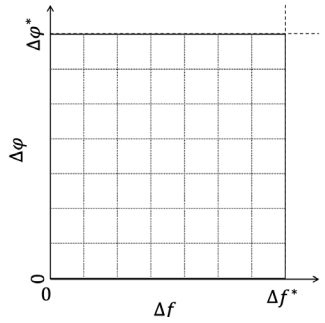

Figure 3.3 The seven-by-seven two-dimensional psychometric space

Throughout each trial, listeners reported their subjective percept (integrated/segregated) via key press (Cusack et al. 2004) on a small wireless keypad. To capture perception build-up and any effect of attention (Cusack et al. 2004) instead of a single response for each trial, listeners were told to start responding from the beginning of the trial, and they were allowed to change their response as many times as they wished until the trial ended. For each trial, the proportion of the time the listener reported segregation was used as an estimation of the probability of segregation at the tested coordinate (*Figure 3.4*).

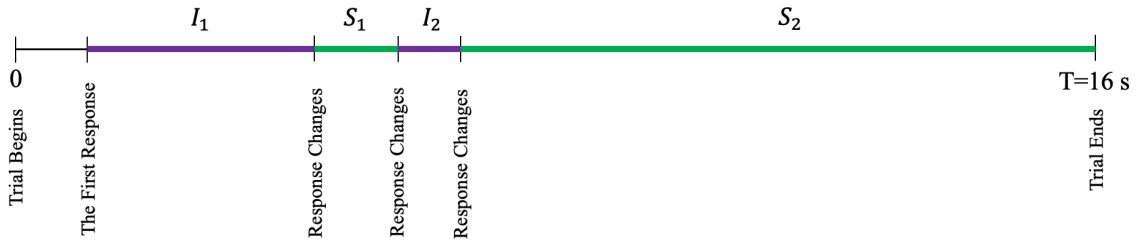

Figure 3.4 The progression in a sample trial: "I" and "S" indicate "integrated" and "segregated" percept, respectively. The segregation probability estimated in this trial is (S1+S2)/16.

3.2.2.2 Adaptive Up-Down Procedures

Assuming segregation probability increases with $\Delta\varphi$ or Δf, a monotonic ascending function (sigmoid) was selected to model this probability. To capture the highest sensitivity and avoid ceiling effects in the measurements, at the beginning of the experiment, the parameter representing each dimension was adjusted by holding the other dimension at $\Delta = 0$ to estimate the dynamic range of that dimension. To do so, a transformed Up-Down adaptive procedure was followed to estimate the value of each parameter, leading to an 84.1% chance of segregation when the other parameter is set to zero (Levitt 1971). These parameters were referred to as $\Delta\varphi^*$ and Δf^*. To estimate $\Delta\varphi^*$, sequences were presented at $\Delta f = 0$ semitone (frequency of both sequences was 400 Hz), then starting with $\Delta\varphi = 0°$, the value of $\Delta\varphi$ on each new trial was determined based on the responses on the previous trials: if the listener reported an integrated percept for less than 90% of the duration of previous trial, $\Delta\varphi$ increased; if listeners reported segregated for more than 90% of the duration all four previous trials, $\Delta\varphi$ decreased; otherwise $\Delta\varphi$ did not change on the next trial. This process was completed for six runs (changes of parameter value direction) with 8° step size for the first two runs and 4° for the last four runs. Then, $\Delta\varphi^*$ was estimated by the average of $\Delta\varphi$ values at the end of the six runs. Δf^* was estimated similarly by setting $\Delta\varphi = 0°$ (collocating A and B sources at $\varphi = 0°$) and adjusting the values of Δf systematically by the same rules, starting from $\Delta f = 0$ semitone. For the first two runs of estimating Δf^*, the step size was set to two semitones, and for the final four runs, one semitone was the change increment. The two thresholds were tested independently, but to prevent the listener from recognizing patterns, the two adaptive procedures were interleaved. If the participant reached the higher bounds of $\Delta\varphi$ (360°) or Δf (12 semitones), the procedure was stopped, and the boundary value was assumed as the estimated 84.1% parameter. The Δf^* value was rounded up to the closest value in the 0 to 12 semitone range with a 0.25 semitone resolution. Finally, the range between 0° and $\Delta\varphi^*$ with increments of $\frac{\Delta\varphi^*}{6}$, and 0 and Δf^*semitone with increments of $\frac{\Delta f^*}{6}$ shaped the two-dimensional psychophysical space.

3.2.2.3 Sessions and Blocks

Each of the 49 coordinates of the psychophysical space was tested ten times, distributed among ten experimental blocks. Each block had 49 trials, one trial from each of the coordinate parameters. The order of trials was randomized such that no similar $\Delta\varphi$ or Δf values were presented in the two consecutive trials (to avoid carryover effects between trials). Because of the lengthiness of the experiment, blocks were split into two sessions.

Similar to the experiment explained in Chapter 2, listeners consented, were tested for hearing thresholds, had their head measurements recorded, and filled out the pre-experiment questionnaire at the beginning of the first session. Then, they were familiarized with the task and carefully instructed. To ensure participants had learned the task and were comfortable with the experiment setup, they completed a practice block of 15 trials. Finally, the transformed Up-Down procedures were presented in a separate block before the experimental blocks. Because the instructions and convergence of the adaptive procedure for some listeners were time-consuming, the number of experimental blocks completed in the first sessions differed between participants (between 3 and 5). In the second session, which was one or two days after the first one, listeners completed a practice block to ensure they remembered the task, then they completed the remaining experimental blocks and, finally, filled out the post-experiment questionnaire.

3.2.2.4 Instructions and Feedback

To familiarize the participants with the task, they were first presented with two (somewhat extreme) sample trials with Δf = 0.75 and Δf = 11.25 semitones with sources collocated at φ = 0°. After they described what they heard and how the sample trials were different, they were told that all the trials would be similar to these two samples, with two "sounds" present: the "galloping rhythm" and "morse code pulses" metaphors were also introduced to them to provide some examples of the differences between the two patterns intended to be discriminated in the experiment and how the two sample trials are more similar to one or the other. The two samples were presented to them again, and they were asked if they could understand the differences between them; they were also told to try to segregate the sequences when presented with both samples. (More examples were

presented to them on rare occasions when they asked or were confused. The number of examples was kept minimal to avoid biasing or imposing effects.) Once they understood the alternative percepts, they were told that they should respond “integrated” or “segregated” (with their own words and labels for these percepts) as soon as they had a percept and to listen carefully throughout the trial to report any potential changes in their response. Most importantly, to unify the definition of segregation between the participants, although they knew that there were always two sequences presented to them, they were told that they should respond “segregated” if they could focus on one sequence (no matter which one or if they could switch between the streams) and “integrated” otherwise. Listeners were told that they were allowed to toggle between the two options without any limits and that they should change their response as soon as they realized their perception had changed. Finally, before each experimental block, they were reminded that they should try not to move as much as possible, that they should fixate on the red LED while it is on, and that there are no correct answers in this experiment.

The only feedback provided to the participants was after they did the practice block. The participants were not expected to segregate collocated sources at Δf = 0 semitones, whereas they should have shown evidence of segregation of those sources at Δf above nine semitones (Micheyl et al. 2005). If the general trend of the responses to the practice trials did not follow an ascending pattern with integration at Δf = 0 semitone, they were asked if they were confused, then any potential confusion about the percepts and responding was clarified, and the practice block was repeated to ensure the listener fully understood the task.

3.2.3 Participants

Twelve young (aged 18-30 years), normally hearing (tested with audiometry) listeners (ten female) with no reported hearing or neurological abnormalities completed this experiment. Five participants had no musical training, and the other seven had some musical experience but were out of practice at the time of their participation.

Six other individuals were recruited for the experiment, but their data is not used for analysis here; one participant was excluded because their responses for only 60 trials (out

of 490) were available. One participant always responded with segregation from the beginning to the end of the trials; this participant also reported frequent ear infections. The four remaining participants were excluded because they reported segregation at the control condition ($\Delta f = 0, \Delta\varphi = 0$) with an average probability of at least 48% (*Appendix A*). In this condition, no information is provided that leads to segregation at any time; reporting segregation at this coordinate suggests that the listener was perhaps responding in the way they thought the researcher desired.

3.2.4 Analysis

3.2.4.1 Preprocessing

After responses were collected, the trials with no responses (maximum of three per listener) were removed from the analysis. Then, for each trial, the intervals in which the listener reported segregation were identified, and the segregation proportion was calculated.

$\Delta\varphi^*$ value for four listeners was greater than 180°; therefore, some trials were tested with sources located at the back of the listener's head ($\Delta\varphi$ > 180°). In the preliminary analyses, it was observed that not only these trials violated the assumption of $\Delta\varphi$ modulating segregation probability by an ascending function, but also such trials effectively mimicked the segregation probability pattern of the same trials with $\Delta\varphi$ values mirrored with respect to the φ = -90° to φ = 90° line. Additionally, the stimulus frequency ranged from 400 Hz to 800 Hz. In this region, ITD cues are the dominant cue for localization, and because they are identical for the front and the back, they leave some ambiguity if the listener does not rotate their head (which, in the case of this experiment, they did not). Therefore, $\Delta\varphi$ values greater than 180° were transformed to their effective mirrored values to keep the model reasonably simple before the rest of the analysis steps. (Another piece of evidence that justifies this decision was two of such listeners' verbal reports mentioning that they thought all sounds were coming from the front.)

3.2.4.2 Psychophysical Modelling

The one-dimensional sigmoid function used to model psychometric functions with one parameter is shown in *Equation 3.1* in which ψ refers to the psychometric function, x is the physical variable level, β_0 is the constant bias term, and β_x is the variable coefficient which shows the dependency of sensation or perception on the physical stimulus level. In this model, the exponent of constant e (Euler's number) has a linear relationship with the variable, and it could be rewritten such that this linear relationship is magnified (right side of *Equation 3.1*).

$$\psi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_x x)}} = \left(1 + \exp\left(-(\beta_0 + \beta_x x)\right)\right)^{-1}$$

Equation 3.1 The one-dimensional sigmoid function model (the right side of the equation is a form of illustration that visually magnifies the exponential regression term and the β coefficients)

This model could be extended to explain functions with more variables by adding appropriate terms in the exponent. The extended model for this experiment is shown in *Equation 3.2*.

$$\psi(\Delta f, \Delta\varphi) = \left(1 + \exp\left(-\left(\beta_0 + \beta_f(\Delta f) + \beta_\varphi(\Delta\varphi) + \beta_{f\varphi}(\Delta f)(\Delta\varphi)\right)\right)\right)^{-1}$$

Equation 3.2 The sigmoid model extended to two dimensions

In the proposed model ψ is the psychometric function model that depends on Δf and $\Delta\varphi$. β_0 is the bias term, β_f and β_φ are the model coefficients for Δf and $\Delta\varphi$ respectively, and $\beta_{f\varphi}$ is the coefficient for the numerical product of the two variables. Coefficients β_f and β_φ reflect the perceptual weights of pitch and spatial cues in the task, and $\beta_{f\varphi}$ reflects the potential effect of the interactions of the two variables. Also, by substituting $\Delta\varphi$ with zero, this model turns into a one-dimensional sigmoid function with Δf as the only variable which aligns with the previous models that explain the segregation probability for collocated sources.

Because the scale and physical nature of the two variables are different, before further analysis, the Δf and $\Delta \varphi$ values were divided by their maximum value presented to the listener to normalize them and make them dimensionless. Instead of the numerical comparison of beta coefficients, their share in explaining the variance of the observed data was compared.

To estimate each coefficient's share in explaining the variance, four different forms of the model were fitted to each listener's data as labelled and described below:

1) The full model: All beta coefficients were estimated.
2) ~fϕ model: $\beta_{f\varphi}$ was constrained to be 0 while β_0, β_f, and β_φ were estimated. (segregation probability explained by separate terms for Δf and $\Delta \varphi$, but not the interactive term)
3) ~ϕ model: β_φ and $\beta_{f\varphi}$ were constrained to be 0 while β_0 and β_f were estimated. (segregation probability explained just by Δf)
4) ~f model: β_f and $\beta_{f\varphi}$ were constrained to be 0 while β_0 and β_φ were estimated. (segregation probability explained just by $\Delta \varphi$)

Then, the Explained Variance (EV) of each model was estimated by dividing the model's estimation Mean Standard Error (MSE) by the data variance (similar to the explained variance measure in Chapter 2). Finally, the proportion of EV each coefficient took from the full model with their absence estimated the coefficient's contribution to explaining the variance (*Equation 3.3*).

$$EV(\beta) = \frac{EV(full) - EV(\sim\beta)}{EV(full)}$$

Equation 3.3 The contribution of coefficient β is estimated by the proportion of the unexplained variance after it is removed from the full model.

3.2.4.3 Statistical Analysis

Explained variance shares of the coefficients were compared using a one-way ANOVA followed by a set of one-tailed t-tests for post-hoc comparisons.

3.2.4.4 Group Modelling

In addition to the individual-specific models, a group model was trained on the data from all listeners. Similar to the individual-specific models, the group model was fitted to estimate the segregation probability from the normalized Δf and $\Delta\varphi$ values (divided by their maximum value for each listener as explained earlier). Because the Up-Down procedure would ideally capture the 0-84.1% dynamic range for both dimensions, the strategy of concatenating different participants' responses seemed reasonable. In the end, the coefficients' shares in explaining the variance of the observations were estimated with a similar method used for individual-specific models.

3.3 Results

3.3.1 Individual Variability in Perceptual Weights of Spatial and Pitch Cues for Auditory Segregation

The average segregation probability, the estimated two-dimensional psychometric model, and the models used to estimate coefficients' contributions to explaining the data variance are illustrated for a sample participant in *Figure 3.5*.

The average segregation probability heatmap for all participants had the same pattern of low probabilities (dark red) at the bottom left corner, with a gradual change to higher probabilities (blue) moving to the sides, center, and top right corner. However, individual differences were observed in the rate of change along each axis and the diagonal line (*Figure 3.6*). These differences were reflected in the estimated coefficients for each listener (Color-coded dots in *Figure 3.7*).

The segregation probability at the collocated conditions ($\Delta\varphi = 0°$) was separately analyzed to compare with previous studies that presented collocated sources. Δf values for subjective equal probability of segregation and integration in these conditions ranged from 0.86 semitone to 6.42 semitones with an average of 3.85 semitones (*Appendix B*).

The total explained variance with the full model on average was 0.62±0.09. Of this total explained variance, the average contribution portion of β_f (frequency difference

coefficient) was equal to 0.66±0.18, β_{φ} (spatial separation coefficient) equal to 0.46±0.22, and $\beta_{f\varphi}$ (interaction coefficient) equal to 0.03±0.03 (*Figure 3.6*). These proportions are statistically different (one-way ANOVA; $F = 46.55$, $p < 10^{-9}$). Based on the post-hoc analysis, β_f trended towards being greater than β_{φ}, and the difference was marginally significant (one-tail paired sample T-test; $t(11) = 1.79$, $p = 0.0501$). Also, β_f proportion is higher than $\beta_{f\varphi}$ proportion (one-tail paired sample T-test; $t(11) = 12.04$, $p < 10^{-7}$), and β_{φ} proportion is also higher than $\beta_{f\varphi}$ proportion (one-tail paired sample T-test; $t(11) = 6.87$, $p < 10^{-4}$). Additionally, β_f proportions were larger than zero (one-tail one sample T-test; $t(11) = 12.79$, $p < 10^{-7}$) as well as β_{φ} proportions (one-tail one sample T-test; $t(11) = 7.26$, $p < 10^{-4}$), and $\beta_{f\varphi}$ proportions (two-tail one sample T-test; $t(11) = 2.83$, $p = 0.0163$); showing all coefficients' significant contribution to explained variance (although $\beta_{f\varphi}$ proportions seem to be numerically small).

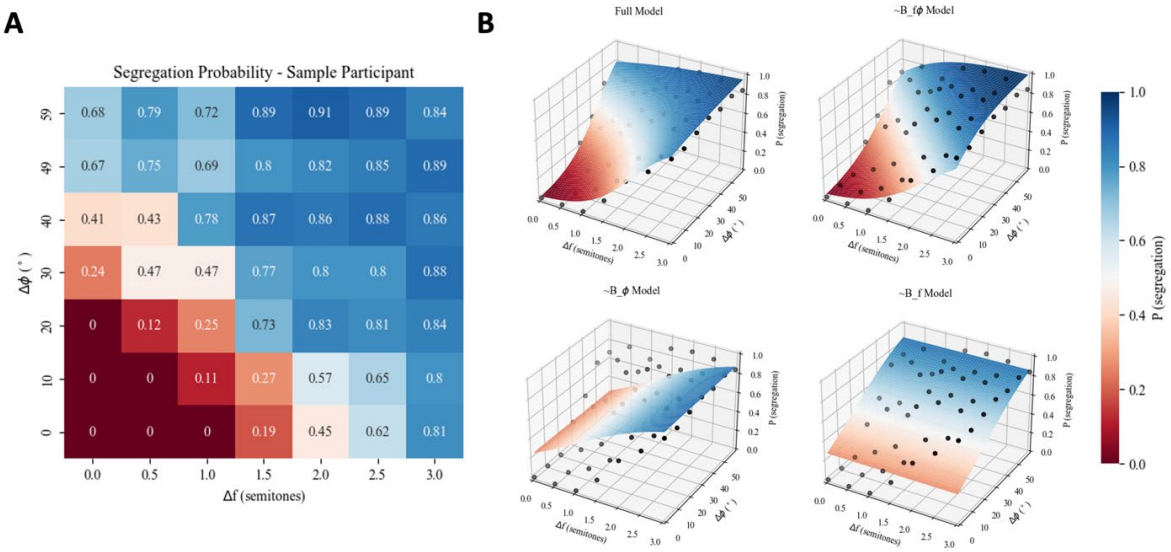

Figure 3.5 A. The average probability of segregation at each coordinate for a sample participant. B. The four models trained on fitted to the data to estimate each coefficient's contribution in explaining the data variance. Black dots represent the values in panel A. (For model fitting, all ten observations at each coordinate were fed into the model.)

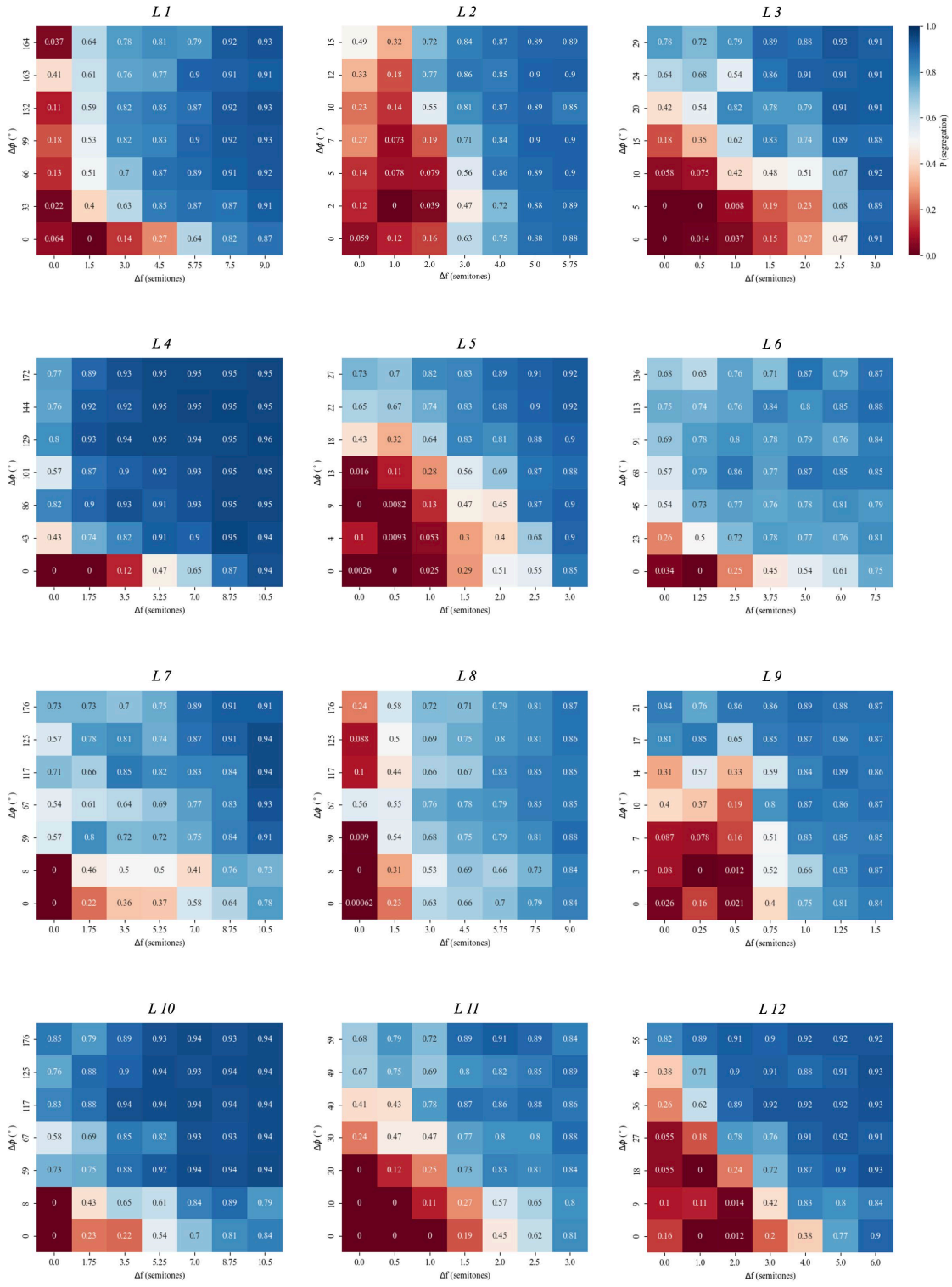

Figure 3.6 Segregation probability in the two-dimensional psychophysical spaces for all twelve listeners.

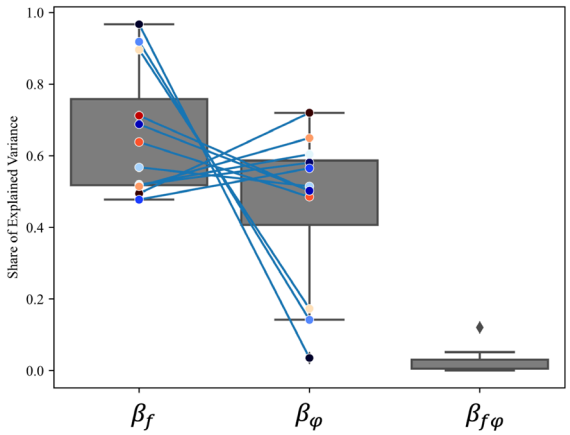

Figure 3.7 Beta coefficients contributions to the total explained variance. Individual listeners' β_f and β_φ contributions are illustrated with color-coded dots that are connected with a blue line for each listener.

3.3.2 The Group Model's Relative Weights of Spatial and Pitch Cues for Auditory Segregation

The coefficients of the group model trained on the data from all listeners are summarized in *Table 3.1*, and the estimated surface is shown in *Figure 3.8*. This model explained the total of 0.50 of data variance.

Coefficient	**Value ± *STD***	**Contribution to the explained variance proportion**
β_0	-2.32 ± *0.07*	-
β_f	4.01 ± *0.12*	0.67
β_φ	3.15 ± *0.11*	0.40
$\beta_{f\varphi}$	-1.57 ± *0.23*	0.01

Table 3.1 Estimated model parameters and their share in the total 0.50 explained variance. STD in the second column shows the parameter estimation standard deviation.

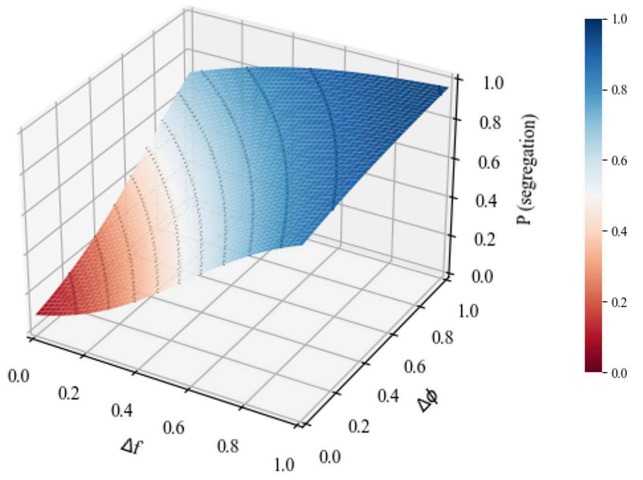

Figure 3.8 The model fitted to the group data. The dashed lines represent contours of equal segregation probabilities on the surface.

3.4 Discussion

In this experiment, the perception of ABA triplet sequences (van Noorden 1975) was tested with A and B sequences presented at seven different pairs of locations on the horizontal plane at seven frequency difference levels, calibrated for each individual to be between 0 and the 84.1% threshold (Levitt 1971) for each dimension. The combination of these two sets of parameters formed a two-dimensional psychophysical space with 49 coordinates. The angle difference between the location pairs and the frequency difference between the two sequences served as spatial and pitch cues contributing to perceptual organization that was reflected in the integration and segregation of the sequences. The proportion of the duration for which participants reported segregation during each trial estimated the probability of segregation at each coordinate. Despite some individual variability, a gradual transition from integration to segregation was observed when frequency difference or spatial separation increased (separately and when they covaried together) in all participants.

Listeners' frequency difference value to equally segregate and integrate the collocated sequences was, on average, 3.85 semitones, which is comparable with the expected value of 3 semitones reported previously for this paradigm (Micheyl et al. 2005).

A two-dimensional sigmoid function with coefficients for spatial cues, pitch cues, and their product value was defined to model the observed data. These coefficients would indicate the perceptual weights of spatial and pitch cues and how much benefit their interaction provides as they are simultaneously provided to the listener. This model was first fitted to the individual listeners' data. Then, the proportion of each coefficient's contribution to the explained variance was estimated by the unexplained variance for a retrained model constrained to set that coefficient to zero. High variability was observed between the listeners' psychometric functions, which was also reflected in the variability of model coefficients and relative perceptual weights; some weighted pitch cues more than spatial cues to organize their perception, and some did the other way around. However, on average, these models could explain an acceptable portion of the variance (0.62 ± 0.09), with pitch cues' coefficients contributing to this explained variance being marginally higher than the contribution of spatial cues. Although the contribution of the interaction coefficient numerically is small, it significantly contributes to the explained variance as well as the other coefficients.

Finally, a group model was trained on the data recorded from all listeners. The variance explained by the group model was lower than for the individually trained models. The contributions of this model's coefficients to explain the variance were in the range of variance estimated by the individual models, with the highest share for the pitch cues' coefficient and the least share for the interaction coefficient. The difference between the spatial and pitch cues share is considerable in the group model (0.27). These results, in combination, suggest that spatial and pitch cues are both essential to organizing perception, and perceptual organization might rely on pitch cues more. Finally, the small effect of the interaction coefficient suggests that combined evidence for segregation does not drastically affect perception in this paradigm and that perception gradually changes as spatial separation and pitch difference increase.

This design, in which listeners reported their subjective perception, had a few caveats that should be considered in future implementations of this experiment: In the subjective task, the listeners always knew that there were two sequences of tones and might have been biased to always report segregation at some point throughout the trial; because they

concluded that it is a “correct” answer to this task and did not want to fail the experiment or have their auditory abilities questioned. Indeed, I excluded four participants based on their segregation responses in a condition in which A and B sequences had zero pitch difference and were collocated, so there were no cues for segregation at all. On the other hand, more conservative listeners who were well-familiarized with the task might have considered reporting segregation only when they were certain. Hesitation when reporting segregation, leads to slowed reaction times and decreases the segregation intervals and, therefore, segregation probability estimated for a trial. Such behavior would contribute to individual differences and might explain some of the observed large values for parameters determined in the adaptive procedure, particularly for the spatial dimension.

Given the observed variability is mostly highlighted in the spatial domain and is less in the pitch domain, and the pitch cues were found to be perceptually weighted more, one might be concerned that this was because the listeners were familiarized with the task and practiced it only with pitch manipulations. This might have led to listeners’ full understanding of how to report perception based on pitch differences and confusion when exposed to spatial manipulations. While this concern might be valid, the initial reasoning behind not introducing the spatial domain in the familiarization block was that the purpose of this experiment was to explore the added spatial domain to a paradigm that is well-studied for pitch differences. Therefore, I familiarized participants with what we expected them to do based on previous experiments and left the spatial differences to be explored in the experiment without exposure or training biases.

An objective task similar to those used by Micheyl et al. (2005) and Thompson, Carlyon, and Cusack (2011) could be employed to account for the described biases and differences in responding strategies across listeners. In these objective designs, some elements of one or two of the sequences are manipulated (modulated or delayed), and participants are asked to report such manipulations. In some designs, targets are accessible only when the listener segregates the sequences, and in some, they are easier to catch when A and B sequences are integrated. Therefore, integration/segregation is implied from the listener’s behavior, and they cannot develop different strategies for direct percept reports. Such designs would also mitigate the concern about listeners being familiarized with the task

and practicing it by pitch differences. Finally, I recommend limiting the spatial separation of the sources to 180° or less at the maximum in future experiments.

Chapter 4

4 Discussion

Auditory scene analysis (ASA) enables humans to organize sound components into meaningful sound streams perceptually. The principles of perceptual organization that govern auditory scene analysis rely on several attributes of sound, such as location and pitch, to segregate auditory streams. More explicitly, the sound components originating from different locations or those with different pitches are less likely to belong to the same source and are, therefore, perceptually segregated (Bregman 1990). As spatial and pitch cues in ASA have been mostly studied with focus on one or the other, it was not clear how perception would be organized when they both systematically are manipulated. In other words, it was unclear whether perceptual segregation based on one domain's cues would dominate integration evidence from the other domain and also how strongly auditory perception would promote segregation when there is some evidence for perceptual segregation in both domain's cues. This project aimed to answer these questions by characterizing perceptual weights of spatial and pitch differences in segregating sounds in a model auditory scene.

To study the interaction of spatial and pitch cues in ASA, an auditory scene model with two sequences of sounds (van Noorden 1975) that could differ in pitch and presentation location was simulated. A virtual acoustic space with 9th-order ambisonic rendering (the Audio Dome) was utilized to present sound sources on a continuum of locations. Because this virtual space had not been used for psychophysical experiments, its reliability for sound reproduction was assessed in the first experimental chapter before it was used to present auditory scenes in the second experimental chapter.

In the first experimental chapter, the minimum audible angles for human listeners were estimated at several locations on the frontal half of the horizontal plane for sounds that were rendered by 9^{th} order ambisonic panning method. The estimated minimum audible angles align with previously reported values with loudspeaker single-channel design (Mills 1958; Strybel and Fujimoto 2000), showing the system's ability to render detailed

enough source locations for human listeners. Additionally, the same level of detail was perceivable for the sources rendered at locations with different nearby loudspeaker densities, indicating the uniformity of the system's precision across the space and its independence from the loudspeakers' layout. In this experiment, I observed an asymmetry between the minimum audible angle values on the lateralized parts of the left and right hemifields. Although human spatial acuity in these locations is very low and listeners performed very poorly in the task, the asymmetry might be explained with interventions of the noises made by one of the experiment setup computers during the experiment.

The minimum audible angles were estimated for broad-band stimuli in which participants verbally reported perceiving unexpected elevation variations at lateral locations. As elevation cues mostly rely on spectral characteristics of the sound, I hypothesized that ambisonic rendering might modify the frequency content of the audio files (e.g., filter some frequencies) that makes them spectrally different from when they are presented (more naturally) from one loudspeaker. To test this hypothesis, I characterized the differences between the frequency content of sound sources that were rendered with ambisonic and single-channel panning methods. To do so, I presented broad-band (linear chirp sweep) sound sources at various locations on the horizontal plane and recorded the signal that was received in the artificial ears of a head-and-torso simulator (HATS). Comparing the frequency response, I showed that ambisonic rendering reliably reproduces frequency content of the lower portion of the spectrum (below ~5 kHz) across the space while it manipulates the frequency content at the higher frequencies that depend on the source location. I concluded that the unexpected elevation cues might have been caused by the high-frequency distortions identified for ambisonic rendering in the Audio Dome. This possibility could be empirically explored in future experiments. Additionally, a more comprehensive assessment can characterize the differences between sound localization cues (ITDs and ILD) perceived for single-channel and ambisonic panning methods. Once the source of fake elevation cues is identified, the minimum audible angles could be estimated again with adjusted stimuli (for example, if the distortions in high-frequency content are identified as a source of fake spatial cues, low-frequency

stimuli should be utilized). Although, it is unlikely that the values would change much, given that they are already at a level consistent with the literature.

In the second experimental chapter, after the fidelity of the Audio Dome to sound reproduction was established, it was used to simulate ABA triplet sequences (van Noorden 1975) with a spatial dimension. In this experiment, in addition to the variable frequency of A and B tones, their presentation location could be different, unlike usual ABA streaming experiments. Seven levels of frequency difference and seven levels of horizontal spatial separation in individually estimated dynamic ranges for each listener formed a two-dimensional pitch-space psychometric space. Then, a two-dimensional sigmoid function was trained to explain the probability of perceptual segregation at the coordinates of this psychometric space for each listener and for the entire group of participants.

The most highlighted observation was the high individual variability in segregation probability patterns, which was also evident in model coefficients and individualized parameter ranges. Indeed, individual variability was expected to a certain degree, but such a high variability could reflect different biases and listeners' different strategies in responding. These differences probably exist because the task required subjective reports; despite my efforts to carefully instruct the listeners to develop the same understating of the task, they seem to have taken different approaches. Some listeners, regardless of the condition, tended to always report segregation as they knew there were always two sequences to the point that they segregated the sequences even when there were no differences between the two (these listeners were excluded from further analysis). On the other hand, some more conservative listeners hesitated to report the segregated percept unless they were absolutely sure. These differences could be solved by replacing the subjective task with an objective task in which perception is inferred from the listeners' behavior. In paradigms in which listeners should detect some targets when streaming sequences, they cannot develop different strategies because those tasks are designed to be performed accurately only when the sequences are perceptually integrated/segregated (Micheyl et al. 2005; Thompson, Carlyon, and Cusack 2011).

However, at a group level, the contribution of model parameters to explain data variance revealed that spatial and pitch cues are both essential to organize perception in this paradigm, and on average, pitch cues have a higher perceptual weight. Also, the transition from integration to segregation in this paradigm was observed to be gradual, and segregation was not drastically promoted when there was some evidence for segregation in both domains.

In future work, after I have validated an objective task as a robust and sensitive measure to quantify the contributions of spatial and pitch cues to promoting segregation, I plan to recruit music ensemble conductors to investigate the effect ensemble conducting practices on perceptual weighting of these cues. As conductors are generally known for detecting errors in a group of performers that are spatially distributed, I predict that they will show higher perceptual weights for spatial cues compared to non-musicians.

In summary, this project first verified the reliability and precision of the Audio Dome's 9^{th}-order ambisonic rendering for psychoacoustic experiments with lower-frequency (below ~5 kHz) audio stimuli. Then, I showed that for a non-musician listener cohort, spatial and pitch cues are both essential for organizing perception in auditory scenes and despite pitch cues seeming to be more influential at the group level, there is high individual variability.

Bibliography

American Standards Association. 1960. “Acoustical Terminology SI.” *1-1960, American Standards Association*. https://cir.nii.ac.jp/crid/1570572700315960960.

Bertet, St\’ephanie, J\’erôme Daniel, Etienne Parizet, and Olivier Warusfel. 2013. “Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources.” *Acta Acustica United with Acustica* 99 (4): 642–57.

Blauert, Jens. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press.

Boehnke, Susan E., and Dennis P. Phillips. 2005. “The Relation between Auditory Temporal Interval Processing and Sequential Stream Segregation Examined with Stimulus Laterality Differences.” *Perception & Psychophysics* 67 (6): 1088–1101.

Bregman, Albert S. 1990. “Auditory Scene Analysis: The Perceptual Organization of Sound” 773.

———. 2005. “Auditory Scene Analysis and the Role of Phenomenology in Experimental Psychology.” *Canadian Psychology/Psychologie Canadienne* 46 (1): 32.

Carlyon, Robert P. 2004. “How the Brain Separates Sounds.” *Trends in Cognitive Sciences* 8 (10): 465–71.

Cusack, Rhodri, John Deeks, Genevieve Aikman, and Robert P. Carlyon. 2004. “Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis.” *Journal of Experimental Psychology. Human Perception and Performance* 30 (4): 643–56.

Darwin, C. J., and R. P. Carlyon. 1995. “Auditory Grouping, in Handbook of Perception and Cognition: Hearing, Vol.6.” San Diego: Academic Press.

Darwin, C. J., and R. W. Hukin. 1999. “Auditory Objects of Attention: The Role of Interaural Time Differences.” *Journal of Experimental Psychology. Human Perception and Performance* 25 (3): 617–29.

Eramudugolla, Ranmalee, Dexter R. F. Irvine, Ken I. McAnally, Russell L. Martin, and Jason B. Mattingley. 2005. “Directed Attention Eliminates ‘change Deafness’ in Complex Auditory Scenes.” *Current Biology: CB* 15 (12): 1108–13.

Farina, Angelo. 2000. “Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique.” In *Audio Engineering Society Convention 108*. Audio Engineering Society. https://www.aes.org/e-lib/browse.cfm?elib=10211.

Gescheider, G. A. 1997. “Psychophysics: The Fundamentals.” *Lawrence Erlbaum Associates*.

Guzman, Adolfo. 1969. “Decomposition of a Visual Scene into Three-Dimensional Bodies,” January. https://dspace.mit.edu/handle/1721.1/6173?show=full?show=full.

Koffka, K. 1935. “Principles of Gestalt Psychology” 720. https://psycnet.apa.org/fulltext/1935-03991-000.pdf.

Lahav, Amir, and Erika Skoe. 2014. “An Acoustic Gap between the NICU and Womb: A Potential Risk for Compromised Neuroplasticity of the Auditory System in Preterm Infants.” *Frontiers in Neuroscience* 8 (December): 381.

Levitt, H. 1971. “Transformed Up-down Methods in Psychoacoustics.” *The Journal of the Acoustical Society of America* 49 (2B): 467–77.

Micheyl, Christophe, Robert P. Carlyon, Rhodri Cusack, and Brian C. J. Moore. 2005. "Performance Measures of Auditory Organization." *Auditory Signal Processing*. https://doi.org/10.1007/0-387-27045-0_25.

Mills, A. W. 1958. "On the Minimum Audible Angle." *The Journal of the Acoustical Society of America* 30 (4): 237–46.

Moore, Brian C. J., and Hedwig E. Gockel. 2012. "Properties of Auditory Stream Formation." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367 (1591): 919–31.

Neal, Matthew T., and Pavel Zahorik. 2022. "The Impact of Head-Related Impulse Response Delay Treatment Strategy on Psychoacoustic Cue Reconstruction Errors from Virtual Loudspeaker Arrays." *The Journal of the Acoustical Society of America* 151 (6): 3729.

Noorden, L. P. A. S. van. 1975. "Temporal Coherence in the Perception of Tone Sequences (Unpublished Doctoral Thesis)." *Eindhoven University of Technology*.

Plack, Christopher J. 2018. *The Sense of Hearing*. Third edition. | Abingdon, Oxon; New York, NY: Routledge, 2018.: Routledge.

Pulkki, Ville. 1997. "Virtual Sound Source Positioning Using Vector Base Amplitude Panning." *Journal of the Audio Engineering Society* 45 (6): 456–66.

Strybel, T. Z., and K. Fujimoto. 2000. "Minimum Audible Angles in the Horizontal and Vertical Planes: Effects of Stimulus Onset Asynchrony and Burst Duration." *The Journal of the Acoustical Society of America* 108 (6): 3092–95.

Tabry, Vanessa, Robert J. Zatorre, and Patrice Voss. 2013. "The Influence of Vision on Sound Localization Abilities in Both the Horizontal and Vertical Planes." *Frontiers in Psychology* 4 (December): 932.

Thompson, Sarah K., Robert P. Carlyon, and Rhodri Cusack. 2011. "An Objective Measurement of the Build-up of Auditory Streaming and of Its Modulation by Attention." *Journal of Experimental Psychology. Human Perception and Performance* 37 (4): 1253–62.

Tominaga, Shoji, Shinichi Tamura, Kokichi Tanaka, and Seihaku Higuchi. 1975. "Uncorrelated Minimum-Length Sequence and Its Application to Parameter Estimation." *Information Sciences* 9 (2): 151–68.

Zotter, Franz, and Matthias Frank. 2019. *Ambisonics*. Springer International Publishing.

Appendices

Appendix A

Four listeners in the experiment presented in Chapter 3 reported the segregated percept for at least 48% of the time in the collocated sources of sound with the same frequency (this is the (0,0) coordinate of the two-dimensional psychophysical space, *Figure A 1*). In this condition there are no physical differences between the A and B sounds. Therefore, it is not possible to perceptually segregate them and high segregation probability indicates listeners' biases. For this reason, these listeners were excluded from the analysis.

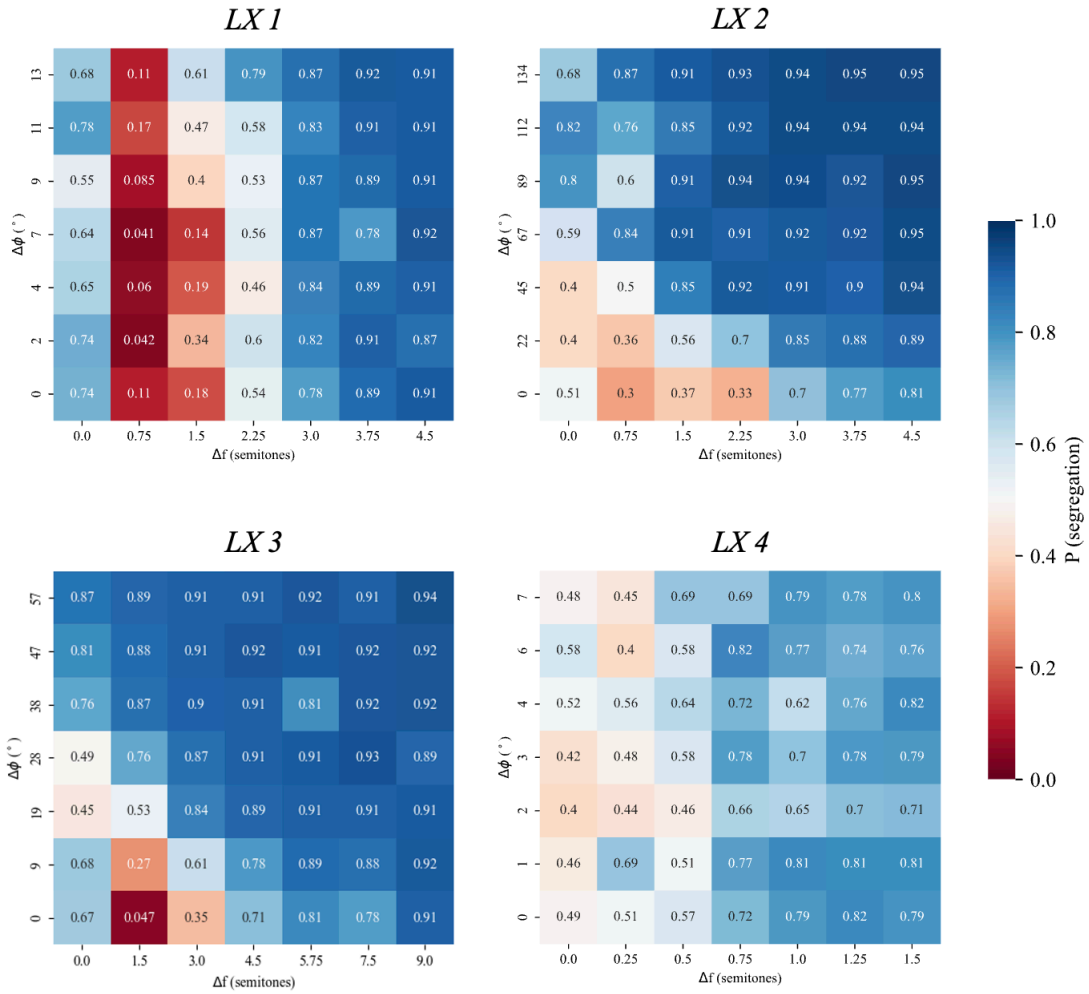

Figure A 1. Segregation probability in the two-dimensional psychophysical spaces for the four listeners excluded from the analysis in Chapter 3.

Appendix B

For the experiment presented in Chapter 3, listeners' behaviors in the conditions where the sources were collocated ($\Delta\varphi = 0°$) was analyzed and compared with the previous studies that presented such stimuli. For each listener a one-dimensional sigmoid function (*Equation 3.1*) was fitted to estimate segregation probability based on Δf only on the data from trials in which the sources were collocated (the bottom horizontal line of the two-dimensional psychophysical spaces). Then PSE (50% segregation probability) was estimated based on the psychometric function for each listener (*Figure A 2*). PSEs ranged from 0.86 semitone to 6.42 semitones with an average of 3.85 semitones.

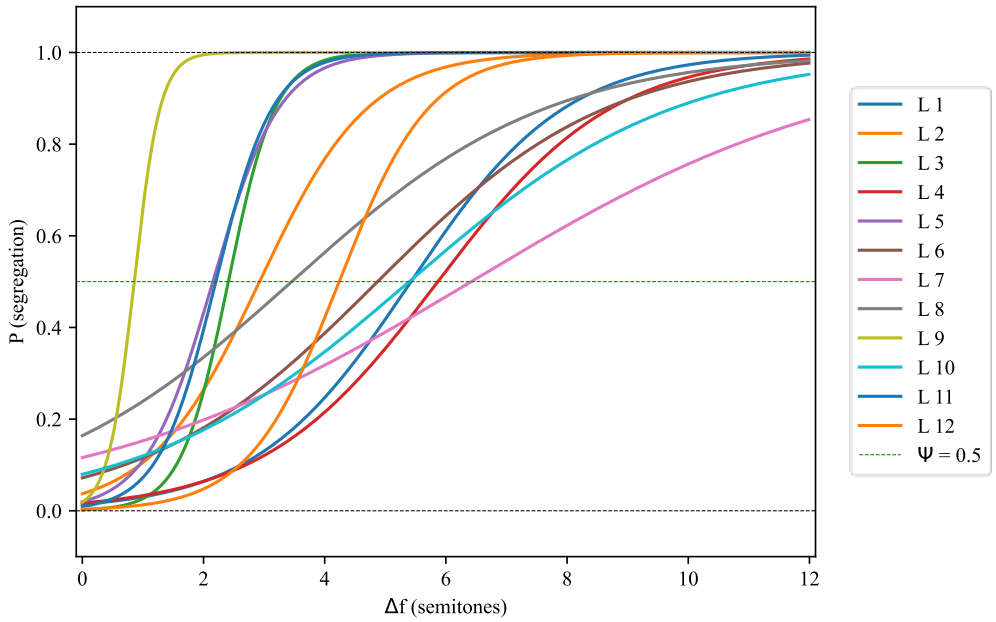

Figure A 2. Estimated segregation probability functions in the collocated sources conditions for the twelve listeners of the experiment presented in Chapter 3. The intersection of each listener's curve and the green dashed line approximates PSE's Δf value for that listener.

Appendix C

The experiments presented in this thesis were conducted under the approval of the Western University Non-Medical Research Board.

Western Research

Date: 17 April 2023

To: Dr. Ingrid Johnsrude

Project ID: 121478

Study Title: Behavioral studies of the effects of musical training on the perceptual organization of sound in auditory scenes

Short Title: Perceptual organization of music

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: 05/May/2023

Date Approval Issued: 17/Apr/2023 16:00

REB Approval Expiry Date: 17/Apr/2024

Dear Dr. Ingrid Johnsrude

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study.**

Documents Approved:

Document Name	Document Type	Document Date	Document Version
4.7_BrainSCAN_2023_02_16_Johnsrude_121478	Recruitment Materials	16/Feb/2023	1
2.5.DemographicQuestionnaire_V2	Paper Survey	28/Mar/2023	2
2.10.Debriefing_V1	Debriefing document	28/Mar/2023	1
2.5.PostExperimentQuestionnaire_V1	Paper Survey	05/Dec/2022	1
Poster_V2_Experiment1	Recruitment Materials	29/Mar/2023	1
Poster_V2_Experiment2	Recruitment Materials	29/Mar/2023	2
Poster_V2_Experiment3	Recruitment Materials	29/Mar/2023	1
4.1.6f_email script_V2	Recruitment Materials	29/Mar/2023	2
4.7.BrainSCAN_2023_03_29_Johnsrude_121478	Recruitment Materials	29/Mar/2023	2
2.5.pre-screening survey_V2	Online Survey	14/Apr/2023	2
2.5.eligible participant email collection survey_V1	Online Survey	14/Apr/2023	1
5.5.LOI_Consent_V3	Written Consent/Assent	14/Apr/2023	3

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Ms. Katelyn Harris , Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Curriculum Vitae

Name:	Nima Zargarnezhad
Post-secondary Education and Degrees:	Sharif University of Technology Tehran, Iran 2016-2021 B.Sc. The University of Western Ontario London, Ontario, Canada 2021-2024 M.Sc.
Related Work Experience	Teaching Assistant The University of Western Ontario 2022-2023