Spatializing Stereo: A Framework For Creative Live Upmixing

Liam Martley
Northeastern University
Boston, USA
martley.l@northeastern.edu

ABSTRACT

In this paper I describe the process of the design, development, and implementation of a multichannel spatial audio system for use in live performance. Based on a method for stereo upmix established by Carlos Avendano and Jean-Marc Jot, a software named the Creative Upmixer was developed. The Creative Upmixer further expands on the capabilities of Avendano and Jot's method through the incorporation of a comb filtering trick dubbed the "Super Separator." A multi speaker arrangement will be created and tuned in order to reproduce the up-mixed signal. At the end, further developments and implications for this project are discussed

1. INTRODUCTION

1.1 What is Spatial Audio?

The term "immersive" has become somewhat of an overused buzzword, used to describe any sort of audio experience that captivates the listener. When it comes to spatial audio contexts, however, immersion is often used as a subjective qualifier for how much or how well the sound seems to surround or encapsulate during the experience. While there are many physical and psychological components to our perception of sound and its "immersion," its foundation in spatial audio is based on simply increasing the number of source points in the listening environment.

Today, the majority of people experience sound primarily in the two-channel stereo format. Headphones, laptops, and phones all take advantage of merely a left and right channel in order to capture the spatial dimension of sound. Stereo is easily accessible, and, through its singular dimension of left to right, generally provides enough spatial flexibility for most audio applications. However, advancements in media formats such as movies and music have created a demand for more "immersive" experiences. People looking for this immersion typically need to go to specific locations specifically set up for multichannel audio. Movie theaters and cars are perhaps the two most common places where one might encounter spatial audio. However, in the past two decades the rise of digital media has coincided with the desire for greater accessibility of surround sound.

The rise of spatial audio formats such as Dolby Atmos and Auro-3D have effectively standardized the equipment, software, and other requirements for setting up surround sound, whether in a movie theater or in a living room. Spatial audio consists of two parts: 1. physical speakers and 2. upmixing software. The term "upmixing" refers to the generation of additional channels of audio, in this case from a two-channel stereo signal. Mixing engineers often use proprietary software such as the Dolby Surround Upmixer (DSU) to convert stereo audio into formats such as Dolby Atmos and Auro-3D. The majority of movie theaters today deploy one of the major formats depending on their size and budget. However, the specificity of these setups may make implementing them a challenge for those looking to start working with spatial audio at home. Atmos and Auro-3D require specific numbers of speakers with a variety of positions and heights throughout a room. At minimum, the cost of upscaling a two-speaker set up to five, eleven, or even thirteen speakers adds up very quickly. Furthermore, the structural challenges posed by the positioning of these speakers disqualifies many rooms that would otherwise be very suitable listening environments. The relative lack of flexibility in speaker setups also restricts the use of tools such as DSU in upmixing stereo signals.

1.2 Focus, Scope, and Inspirations

This project addresses the aforementioned inaccessibility of spatial audio formats. Specifically, it provides a complete spatial audio system for use in the 2024 Northeastern University Music Technology Capstone Concert. This event is a formal showing for graduating students to both demonstrate and punctuate their experience in the Music Technology program. While the foundations of Music Technology at Northeastern are based in composition, the program's focus has shifted more towards technological applications of music in recent years. This year's graduates consist of a diverse set of interests, experiences, and careers ranging from recording artists, producers, developers, and engineers. Thus, the culminating performance consists of a number

of multimedia works, including recorded and live music, prerecorded video, and live demonstrations.

This year's performance takes place at Northeastern's Fenway Center located on their Boston campus. The Fenway Center, previously serving as St. Ann's Church, was converted to "a versatile space for choral music, recitals, banquets, lectures, and more" [1].The venue is equipped with a number of features useful for orchestrating the performance. For example, this project utilizes the in-house mixing desk, a Yamaha QL5, in order to route the student's audio out to eight speakers positioned around the audience. However, the venue also presents a number of constraints and challenges that have influenced the scope and implementation of this project. Firstly, the physical structure of the space prevents the use of spatial audio formats such as Dolby Atmos and Auro-3D. The major spatial audio formats available to consumers require the use of speakers placed vertically above the listening position, which is not a possibility for the Capstone Performance.

Because this limitation disqualifies the use of established upmixing software, one of the key components of this project is the creation of a custom upmixing software. The up-mixer was developed using MaxMSP, a visual patching environment used for the creation of audiovisual software. The finalized product implements the upmixing algorithm developed by Carlos Avendano and Jean-Marc Jot in their paper "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix." This method uses Short-Time Fourier Transform (STFT) in order to compute the correlation between the left and right channels, separating the "direct" and "ambient" components for each source channel. However, this method produces only four of the eight audio channels desired for this project. Additional channels are synthesized under a principle inspired by a mixing technique coined the "Super Separator" by Dan Worrall, a sound engineer and video producer for music software companies such as FabFilter and Camel Audio. The "Super Separator" involves strategically filtering two different sources of audio in order to make their spectrum more complementary to each other. In this case, the choice to use this method serves as a creatively-inspired choice rather than one invested in a realistic recreation of a listening environment. Nevertheless, the resulting upmixing software converts stereo input to eighth channels of output: two "direct" and six "ambient."

The physical component of this project is the deployment of speakers in the venue. As the author of this project is a graduating member of the 2024 Music Technology program, it is heavily inspired by their experience and career as a live sound technician. A major focus in the framing of this project was to incorporate elements from concert audio. Concerts are essentially spatial audio systems, with various arrays of speakers positioned to project the audio across the crowd while simultaneously maintaining a clear and cohesive image. Time alignment, the delaying of certain speakers in order to better align the phases of the composite signal they produce, is an essential component of tuning concert audio systems and is utilized in this project as well. Implementing spatial audio in the Fenway Center presents a few acoustic challenges that will be highlighted in the sections below. The rest of this paper will provide more technical insight into the digital and physical components of this project. It will conclude by discussing the future implications for this project, including potential areas for improvement and development.

2. THE CREATIVE UPMIXER

2.1 Frequency Domain Ambience Extraction á la Avendano and Jot

The algorithm utilized in this project was developed by researchers Carlos Avendano and Jean-Marc Jot. Their paper, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mixing," notes that their method was inspired by the "direct/ambient approach for mixing multichannel audio as described in T. Holman's "Mixing the Sound" article for Surround Magazine. This approach defines a process for the creation of multichannel recordings, in which microphones are strategically placed either close to sound sources, such as instruments and amplifiers, or further away from the sound sources. The microphones placed near sound sources capture the "direct" sound, while the distant microphones pick up room character and reverberation, applause, or other background noises all constituting "ambience." During the mixing phase, the "direct" signal is generally panned to the front channels, while the rear channels receive the "ambient" signals. Avendano and Jot note that this particular routing of signals can create the impression of a "virtual listening environment" capable of simulating the effect of being "in the audience of a concert hall, in front of the stage (best seat in the house)" [2].

Citing analysis from sources such as "Spatial Hearing" by J. Blauert, Avendano and Jot explain that their method of ambience extraction was inspired by our own process of binaural hearing. Our brain takes in data from our left and right ears and performs a number of computations and analyses in order to derive spatial information. Specifically, the brain computes the cross-correlation of the data from each ear. Cross-correlation measures the similarity of two signals as a function of one's displacement relative to the other in terms of time [2]. In other words, it involves measuring one signal at discrete time intervals in comparison to the entirety of the second signal in order to determine how similar they are to each other. However, as Avendano and Jot note, this cross-correlation is analyzed as a function of both time and frequency. Our ears contain a structure called the basilar membrane, whose surface is activated in different regions based on frequency content. These

regions are referred to as critical bands, and our ears perform cross-correlation for each of these frequency regions in order to provide us with sound localization capabilities.

The upmixing algorithm developed by Avendano and Jot attempts to simulate this process of inter-ear cross-correlation in order to separate the ambient component of a stereo input. The two signals to be correlated are the left and right channels, and are fed into the Ambience Extraction module as shown in Figure 1 [2].

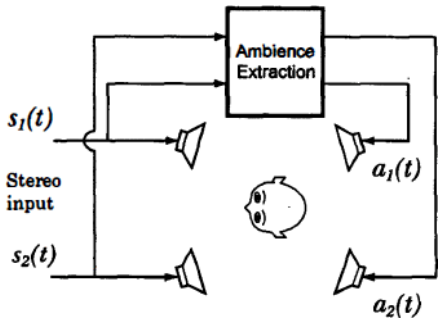

Figure 1. Upmixing algorithm courtesy of C. Avendano and J-M Jot.

However, audio signals are generally transmitted in the time domain. This means that the usable data from an audio signal conveys the amplitude of the signal over time. Conversion from the time domain to the frequency domain can be achieved using a Fourier Transform, such as the Direct Fourier Transform (DFT) or. most commonly, the Fast Fourier Transform (FFT). These processes convert the time-domain signal into its spectral components, producing a multitude of "spectral bins" each containing their relative "strength" or "energy" in the signal. However, in order to work with both time and frequency, the stereo input to be up-mixed must be processed using the Short-Time Fourier Transform (STFT). The STFT computes the Fourier Transform of a signal at regular intervals. Because Fourier Transforms apply a window function to the input in order to minimize spectral artifacts, each individual frame of the transform contains some amount of data loss at the edges of the window. In order to compensate for this loss, Consecutive frames are "overlapped." Conversion to the time-frequency plane allows for the cross-correlation of the left and right channels of a stereo signal for each bin created by the Fourier Transform, thus simulating the process of the basilar membrane.

$$\Phi_{ij}(k)=E\{S_i(m, k)S_j*(m, k)\} \quad (1)$$

Equation (1) shows the correlation function used by Avendano and Jot as only a function of the current frequency bin, k [2]. The * operator refers to complex conjugation, in which the output of, in this case, a Fourier Transform is multiplied by its complex conjugate in order to calculate correlation. Complex conjugation refers to the fact that Fourier Transforms output a value containing both real and imaginary parts; a complex number in the form a + bi. The complex conjugate takes a complex number and flips the sign of the imaginary part, in this case becoming a - bi.

$$\Phi_{ij}(m, k)=\lambda\Phi_{ij}(m-1, k) + (1-\lambda)S_i(m, k)S_j*(m, k) \quad (2)$$

Equation (2) shows the expansion of this correlation function to the time domain, denoted by variable m [2]. There are two important additions to note here. Firstly, Avendano and Jot introduce the forgetting factor λ to account for the dynamism of an audio signal over time. Secondly, they add a recursive element into the correlation function denoted by the expression $\Phi_{ij}(m-1,k)$. These two additions serve to account for the fact that the correlation of an audio signal will constantly change as a function of time.

$$\Phi(m, k) = \frac{\Phi_{12}(m,k)}{[\Phi_{11}(m,k)\Phi_{22}(m,k)]^{\frac{1}{2}}} \quad (3)$$

Thus, (3) showcases the fully realized "inter-channel short-time coherence function" as defined by Avendano and Jot [2]. It should be noted that there is a distinction between the use of "correlation" and "coherence." Correlation refers to similarity over time, whereas coherence refers to the similarity of two signals with respect to frequency content. In this case, the correlation of the left and right signals is used to calculate their coherence as defined by (3).

In order to extract the ambient signal, $A_i(m, k)$, Avendano and Jot "weigh the channel short-time transforms with a non-linear function of the short-time coherence" [2].

$$A_i(m, k) = S_i(m, k)\Gamma[\Phi(m, k)] \quad (4)$$

Equation (4) shows the final ambient signal composition as defined by Avendano and Jot [2]. Note here that the "non-linear short-time coherence function" is represented by Γ. In their paper, Avendano and Jot explain that the purpose of this function is to leave bins of low coherence unmodified, while heavily attenuating bins of high coherence. The researchers identified the hyperbolic tangent function as a suitable candidate [2].

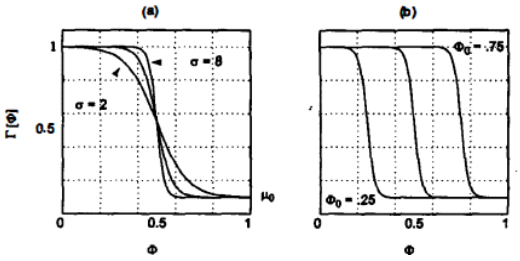

Figure 2. Mapping function courtesy of C. Avendano and J-M Jot

Figure (2) shows a graphical representation of Avendano and Jot's proposed mapping function as a function of the short-time coherence function $\Phi(m, k)$ [2]. Indeed, the hyperbolic tangent preserves at low values of $\Phi(m, k)$ while attenuating as $\Phi(m, k)$ approaches 1. The two graphs demonstrate the behavior of the mapping function due to changes in input parameters for slope and threshold.

$$\Gamma[\Phi] = \left(\frac{\mu_1 - \mu_0}{2}\right) tanh\{\sigma\pi(\Phi_0 - \Phi)\} + \left(\frac{\mu_1 - \mu_0}{2}\right) \quad (5)$$

Equation (5) defines the full equation for the "non-linear short-time coherence function" as proposed by Avendano and Jot [2]. The various parameters of this equation constitute the changes in behavior demonstrated in Figure 2. For example, u_1 and u_0 control the floor and ceiling of the function output for each frequency bin. Avendano and Jot note that u_1 should generally maintain a value of one in order to simply preserve the low-coherence bins, while u_0 should hold a small value greater than one to achieve proper attenuation in highly-correlated bins. The variable σ controls the slope of the function, which is demonstrated in graph (a) of Figure 2 [2]. Greater values of σ create a steeper slope of attenuation at the threshold. This threshold is determined by the value of Φ_0 as denoted by graph (b) of Figure 2. Higher values of Φ_0 leads to a less strict attenuation threshold.

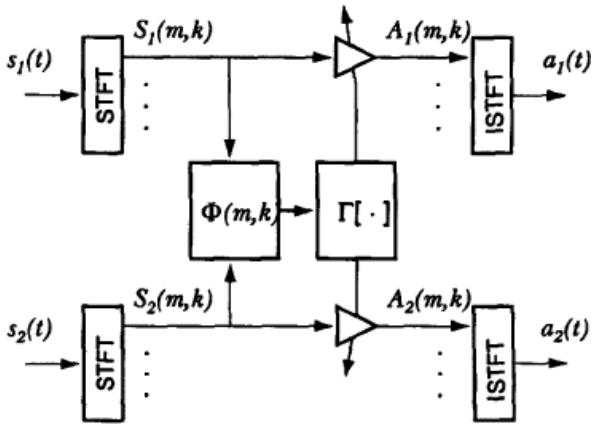

Figure 3. Ambience Extraction module (courtesy of C. Avendano and J-M Jot)

Figure 3 diagrams the Ambience Extraction module as defined in Figure 1. Note that the composite ambient signal must then be reversed back into the time domain via the Inverse Short-Time Fourier Transform (STFT).

2.2 Implementation in MaxMSP

The application for Avendano and Jot's upmixing method was created using MaxMSP, a visual patching environment used for the creation of audiovisual software. MaxMSP provides a number of objects with predefined functions which makes it an ideal tool for easily implementing ideas. For example, the pfft~ object contains all of the functionality to perform an STFT by simply providing a signal to transform. The object's parameters include frame size and overlap factor, which control the number of frequency bins generated and the number of total frames overlapped to form the composite STFT respectively. The pfft~ object also handles the ISTFT within itself, making it an invaluable tool for this project. Furthermore, MaxMSP is efficiently optimized for interaction with external hardware. This patch needs to take multiple stereo inputs and output eight channels of audio for each source, which is mostly handled via the adc~ and dac~ objects. The adc~ object handles all of the input sources, while the dac~ object routes audio out to the channels of whatever output device is set.

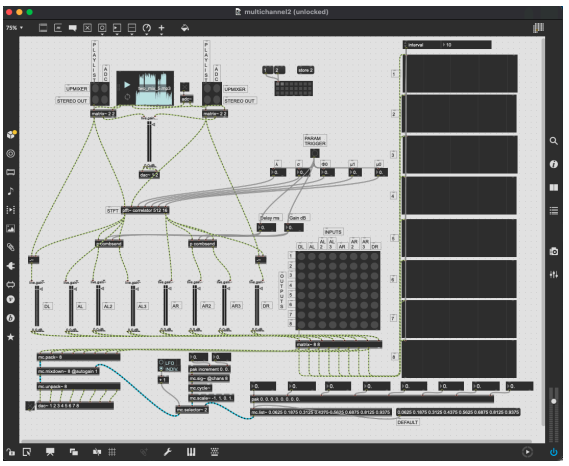

Figure 4. User-level view of the Creative Upmixer

Figure 4 shows the top level of the Creative Upmixer for this project using the method developed by Avendano and Jot. Due to the complex math involved in computing the ambient signals, the various functions employed in Avendano and Jot's method were encapsulated into subpatches within this top layer. The top layer serves as the main control interface for the Upmixer; users should not have to go into the subpatches unless there is a bug. Thus, this top layer is divided into four sections: input, processing, analysis, and output.

The input section sits in the top left quadrant of the patch. As of right now the Upmixer is optimized for two types of sources: inputs from the selected hardware device and preloaded audio. The preloaded audio sits in a playlist~ object, which serves as a buffer which can hold both lossless and compressed audio formats. The external inputs are handled via the adc~ object as mentioned above. Both of these objects output left and right channels separately, which are routed to individual matrix~ objects. Matrix~ objects are useful for sending copies of a signal to multiple places. In the input section, the matrix~ objects are used to allow the user to route each channel to either the Ambience Extraction module or simply to a normal stereo output. The matrix~ objects are controlled by matrixctrl objects, which are represented by the arrays of buttons. When connected to a matrix~ object, these matrixctrl objects allow the user to route each input to the desired output at the click of a button. The adc~ object is currently set up to only handle

one stereo input at a time, due to the constraints of the hardware used to develop and test the Upmixer. However, accommodating more inputs is as simple as providing the adc~ object with specified input numbers in its arguments.

The input section of the patch routes audio to the processing unit via the pfft~ object. As stated previously, the pfft~ object handles both the STFT and the ISTFT for each input. However, the object does this through the creation of a new subpatch; the Ambience Extraction module is contained within the top-layer of the Upmixer.

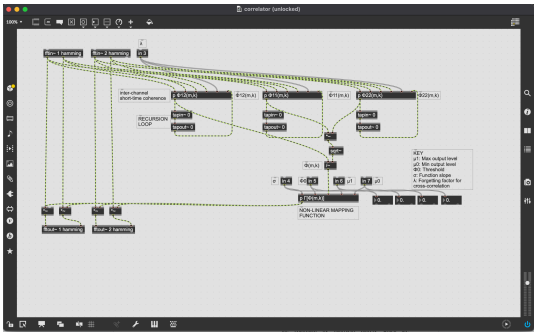

Figure 5. Ambience extraction subpatch

Figure 5 shows the top-level view of the Ambience Extraction module. In order to help with organization, the various mathematical functions employed by Avendano and Jot are contained within their own subpatches. However, the objects that should first be noted are fftin~ and fftout~. These objects serve as inlets and outlets for the subpatch, but they are also responsible for the calculations of the STFT and ISTFT respectively. Audio enters via fftin~, being sent to the various $\Phi_{ij}(m, k)$ function subpatches, as well as to the *~ objects positioned before the ISTFT to calculate the value for $S_i(m, k)\Gamma[\Phi(m, k)]$. Each copy of the $\Phi_{ij}(m, k)$ is assigned either the left and right channels, both left, or both right channels, in order to calculate their respective correlations.

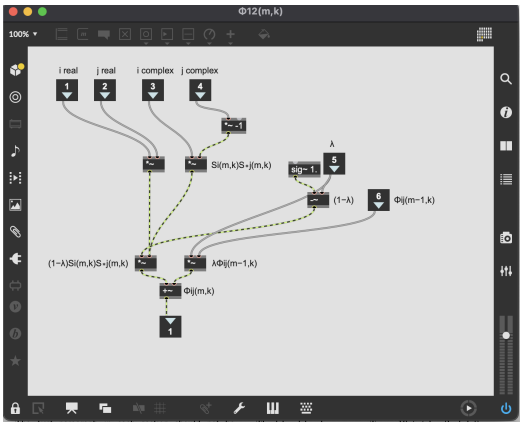

Figure 6. Correlation and coherence subpatch

Figure 6 shows the layout of one of the $\Phi_{ij}(m, k)$ subpatches. Because MaxMSP cannot perform equations directly on signals, each step in the $\Phi_{ij}(m, k)$ equation is separated into steps using the various *~ objects and +~ objects. It is important to note here that the numbered inlets are receiving input from the top-layer control patch of the Creative Upmixer. Referring back to Figure 4, the various objects containing floating point values are sending these values to the $\Phi_{ij}(m, k)$ equation based on their label. The equation subpatch receives these values at instantiation and uses them as constants to calculate the correlation values. It is also important to note that each $\Phi_{ij}(m, k)$ module feeds its output back into itself. As noted in the analysis of Avendano and Jot's equation, the $\Phi_{ij}(m, k)$ contains a recursive element. Thus, the output from each $\Phi_{ij}(m, k)$ module is looped back with a delay created by the tapin~ and tapout~ objects. These objects work in tandem to delay signals entering their sequence. In this case the delay value is set to zero, since its only purpose is to prevent an infinite recursion loop.

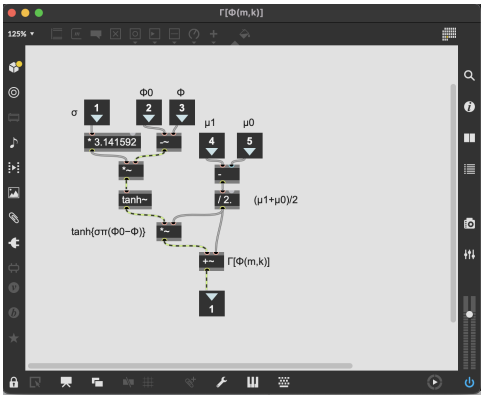

Figure 7. Mapping function subpatch

Figure 7 shows the layout of the non-linear mapping function $\Gamma[\Phi(m,k)]$. Once again referring back to Figure 4, the majority of the parameters offered at the top level are used in this mapping function. Once again, the equation is broken down into steps via serial sequences of *~ and +~ objects. The output of the $\Gamma[\Phi(m,k)]$ subpatch is finally routed to the *~ objects that perform the calculations for $S_i(m, k)\Gamma[\Phi(m, k)]$. Because fftout~ handles the real and imaginary components of the STFT separately, applying the mapping function is as simple as multiplying each component by the output value.

Once the ambient signal is converted back into the time domain and routed out of the Ambience Extractor, it is then copied to form the remaining four channels of audio. Both the left and right ambient signals are copied two more times, forming six total channels of ambient audio. In order to provide variation, the copied signals are further processed using a method inspired by the "Super Separator" coined by Dan Worrall. Dan Worrall is a sound engineer who has worked with companies such as FabFilter and Camel Audio to create tutorials on mixing plugins and synthesizers. On his eponymous YouTube channel, Worrall documents a number of interesting mixing techniques that take advantage of some more nuanced concepts in music technology. The "Super Separator" trick takes advantage of comb filtering in order to provide better clarity to parts

with competing frequency spectrums. Comb filtering can be created by taking a signal and running it in parallel to a delay in the sub-audible range. This process creates "notches" of attenuation and non-attenuation evenly dispersed along the frequency spectrum of a sound. Worrall's "Super Separator" takes advantage of this behavior by comb filtering the two parts with competing frequency spectrums. The trick is to essentially invert the frequency response of the comb filter in one of the parts so that one signal's attenuation bands align with the other signal's passbands. The behavior of comb filters is relatively predictable with a known delay time, with a fundamental frequency calculated by (6)

$$F_1 = \frac{1}{delay\ time\ (ms)} \tag{6}$$

For example, a comb filter with a delay time of 2 ms will create passbands at the frequency $\frac{1}{0.002} = 500\ Hz$ and its multiples: 1000 Hz, 1500 Hz, etc [3]. This also means that the signal's attenuation bands will occur directly in between; in the case of the example above this would be at 750 Hz, 1250 Hz, etc. Thus, a version of Worrall's "Super Separator" can be implemented simply by running a signal through two comb filters and reversing the polarity of the output of one. Reversing the polarity of a signal involves rotating its phase 180 degrees. In practice in MaxMSP, this simply means multiplying the signal by -1 such that its positive amplitude values become negative and vice versa. By reversing the polarity of one of the comb filters, the pass bands become regions of attenuation and the regions of attenuation become pass bands. When recombined with the signal from the first comb filter, their respective passbands and attenuation regions are aligned opposite each other. In a mixing context, these regions fit together almost like teeth to create space for competing parts while providing the illusion of their spectrums being whole. Within the context of this project, however, the "Super Separator" technique is employed simply as a creative choice to provide additional variation between the synthesized ambience channels. MaxMSP provides a
comb~ object for easy implementation and greater control of comb filtering, and is used as shown in Figure 8.

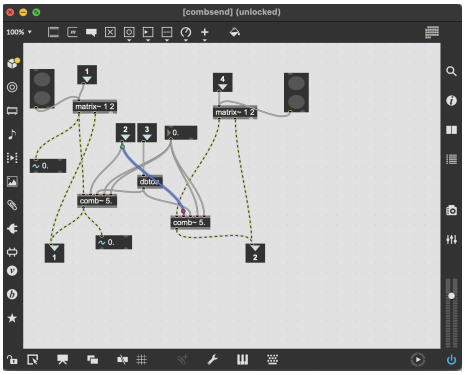

Figure 8. "Super Separator" subpatch

Once again, matrix~ and matrixctrl objects are used to provide greater flexibility; the user can choose to bypass the comb filters and simply utilize unaltered copies of the ambient signal if they desire. Additionally, it should be noted that parameter values for delay time and filter output gain are both provided via floating point number objects in the main patch.

From the processing section, the synthesized channels enter the routing and analysis portion of the patch. Each channel enters this section via labeled live.gain~ objects, which provide visual feedback on the levels of the signals as well as control via a decibel slider. The signals then enter the final routing matrix~, which provides control over signal assignment prior to the output stage of the patch. Each column of the corresponding matrixctrl object represents a signal source, while each row represents the output channel. The signals are then sent to individual spectrometers which provide additional visual feedback on the frequency spectrum for each channel.

The eight signals are also sent to an mc.mixdown~ object which provides opportunity for final panning adjustments prior to the output stage. The eight individual channels are combined into a single multichannel signal via the mc.pack~ object, which is then sent to the mc.mixdown~ object. The right input of the mc.mixdown~ object controls the pan of each individual channel, and accepts another multichannel signal with the same number of channels as the input. Thus, each individual channel can be panned independently and simultaneously through a single cable.

This project has currently implemented two different forms of panning that users can utilize for greater creative control. The default panning mode is individual assignment, which utilizes eight individual floating point number objects to control the individual channels. These values are combined into a multichannel signal via the pak~ and mc.list~ sequence. The pak~ object first combines the individual signals into a generic list, which is then converted into a multichannel signal through mc.pack~. The second panning method utilizes a multichannel signal generated by a low frequency oscillator (LFO). The mc.sig~ object generates an eight-channel multichannel signal which is used to drive the mc.cycle~ oscillator. The mc.scale~ object then converts the bipolar (ranging from -1 to 1) amplitude values of the LFO to unipolar (ranging from 0 to 1) values which can be used to drive mc.mixdown~. Because each output channel operates at its own rate, creating interesting, if disorienting, soundscapes is easily achievable. Finally, the multichannel signal output from mc.mixdown~ is separated once again into individual channels which are routed out to the hardware via dac~. The resulting patch effectively up-mixes stereo input into eight channels of output, separated into direct and ambient components.

3. THE CONCERT

The purpose of the Creative Upmixer was for use for students in the Northeastern University 2024 Music Technology Capstone Performance. Effective implementation of this new tool involves the curation of a multi-speaker listening environment within the Fenway Center (Figure 9).

Figure 9. View from the stage of the Fenway Center looking out at the audience.

While the space is handily equipped with a number of features that will ease the production of the performance, the space is not normally equipped for spatial audio. The creation of a spatial audio setup will thus require equipment contributions from a number of different sources.

3.1 Gear

As stated previously, the Fenway Center comes equipped with a Yamaha QL5 that will be used to route the audio out to the speakers. The Music Technology department at Northeastern has provided a collection of Genelec 8050As with stands that will be used for all speakers in the setup. Additionally, Dr. Ronald Smith of Northeastern University has provided a MOTU 828 MKII to handle audio input and output from the computer hosting the Creative Upmixer. Other requirements, such as stage power and signaling, are outlined on the tech rider accompanying this project and will be provided by the Fenway Center and/or the Music Technology department.

3.2 Tuning

As all venues do, the Fenway Center presents a unique set of acoustic challenges for the creation of an "immersive" spatial audio setup. The two aspects crucial to the tuning of this system will be tonal balance and time alignment. Tonal balance refers to the way that the speakers color the frequency response of the listening position. In most applications, pink noise is used as a test signal because it contains all frequencies at equal loudness. Our ears are more sensitive to certain frequencies than others, which can be a problem when trying to assess the characteristics of a space objectively. This makes signals such as white noise, which contains all frequencies at equal amplitude, not suitable for tuning speakers. This project thus utilizes pink noise as a test frequency in order to assess the recreation of frequencies in the audience of the Fenway Center. Tonal balance can be adjusted on multiple levels, such as via the QL5 with additional adjustment via the tone controls of the 8050As.

Speaker Mounting Position	Treble tilt	Bass tilt	Bass roll-off	Desktop LF
Flat anechoic response	None	None	None	None
Free standing in a damped room	None	-2 dB	None	None
Free standing in a reverberant room	None	-4 dB	None	None
Near field on a reflective surface	None	-2 dB	None	-4 dB
In a corner	None	-4 dB	-4 dB	None

Table 1. Suggested tone control settings in some typical situations

Figure 11. A table of recommended settings for the tone control settings presented on various Genelec speakers, courtesy of Genelec.

Time alignment refers to the relationships in phase between the output of each speaker in a system. Almost every sound has a phase, which quantifies the position of perception of a periodic waveform. Most music consists of periodic waveforms, which means it also has a phase. Sound is a function of both space and time, meaning that the distance of the listener from the sound source affects the way signals are perceived. In particular, multiple signals traveling in the same direction will interact with each other based on their intra-phase relationships. This concept is called wave superposition, which refers to the addition or cancellation of amplitude in a composite signal made up of multiple source signals. Phase cancellation occurs when two sound sources with similar frequency content have phases that align opposite of each other. In these cases, the motion in the positive direction of one wave is counteracted by an equal and opposite motion in the negative direction, resulting in a net-zero change in amplitude. On the contrary, if the phases of these multiple sources happen to align such that the positive and negative regions align, the result is an amplification in amplitude via addition of the signals. In this way, the sources work together to create a composite signal with greater amplitude than the individual parts. This amplification is clearly desirable in successful surround sound systems; at the very least, phase cancellation is extremely undesirable. Thus, time alignment is used to delay certain sources in multi-speaker arrangements in order to better match their phases collectively. Poor time alignment results in comb filtering across the listening position, which is not desired in this context.

3.3 Speaker Placement

Spatial audio formats follow a set of principles based on acoustic properties such as sound diffusion and the aforementioned wave superposition. Diffusion generally refers to the way sound from a particular source will be spread throughout a room based on the size of the

room and the number and type of reflective surfaces. A large room with minimal reflective surfaces will naturally spread sound since the waves need to travel a significant distance to find any reflection points.

3.3.1 *Distance*

One of the first measurements to consider, then, is the distance of each speaker from the center of the listening position. Companies such as Dolby recommend that every speaker be no more than five meters away from this center point [4]. Other companies, such as Genelec, state that a distance of less than four meters is typically ideal [5]. The distance factor for this project, while taking these guidelines into consideration, will heavily be influenced by the acoustics of the Fenway Center. Experts familiar with the space, such as Dr. Ronald Smith, have commented that "The problem of reflectiveness is less of a problem than the sound being too diffuse due to the size of the space and where people are typically seated during the event." Miriam Rosenau, the Technical Operations and Events Manager of the Fenway Center, made a similar comment: "the important thing to remember is that it's a very live and reflective space … drum kits famously sound less than ideal." Thus, the acoustic properties of the Fenway Center encourage very close placement of the speakers in relation to the audience. Special care will be taken to observe the "less than four meters" guideline recommended by Genelec and, to a degree, Dolby.

3.3.1 *Angles*

Another important factor to consider in the creation of multi-speaker arrangements is the angle of rotation for each individual speaker. It is important to establish that each angle is again in reference to the middle of the expected listening position. Logically, it makes sense to point each speaker at the listening position. However, the introduction of these additional speakers creates additional choices to be made.

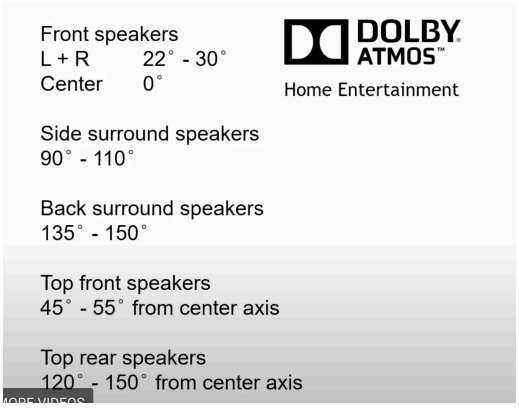

Figure 11. Recommended angle ranges for groupings of speakers in Dolby Atmos, courtesy of Genelec

Figure 11 lists the recommended angles for use in a Dolby Atmos system. It should be recognized that this list groups multiple speakers into sides of the room - front, sides, and back. Additionally, the "top" denotation refers to the fact that commercial spatial audio systems such as Atmos require a set of speakers placed vertically above the listening position. Although the limitations of the Fenway center did not accommodate this vertical layer, respect will be paid to the recommendations for the front, side, and back surround groupings.

4. FUTURE DEVELOPMENTS

Although it has yet to be utilized at the time of this paper, the Creative Upmixer has the potential to be a tool utilized in future iterations of the Northeastern University Music Technology Capstone Performance. The most pertinent development will be to create a more user-friendly user interface that will minimize the tool's learning curve. The main patch is organized in a relatively intuitive way and all pertinent parts and functions are labeled clearly, but the patch lacks the visual appeal of a tool that feels substantial. The patch has yet to be optimized for MaxMSP's presentation mode, which allows the developer to choose which objects and elements to display and hides the rest, while also allowing for reorganization of the visible objects in a way that doesn't need to follow the logic of their physical routing. Some of the parameter controls, such as with the values for the individual panning mode, are cumbersome to use and relatively unintuitive. Floating point number objects are great for storing data, but they lack both the ease and function and appeal that a control should have. The limitations in resolution of the live.dial~ object provided by MaxMSP prevented their use as a control knob for this project, which inspired the creation of a custom knob using MaxMSP's jsui objects. However, the timeline of this project did not allow for such developments. Another reasonable development for the Creative Upmixer would be more pan options. There are a number of creative panning algorithms that can be driven off of various data outputs of the audio at numerous points, which could provide even greater creative potential for interesting sonic soundscapes during the performance. A similar approach could likely also be taken towards the synthesization of additional ambient channels. The "Super Separator" trick was chosen mostly in respect to the work of Dan Worrall, but there are a number of other synthesization options that could provide more realistic or more creative ambient channels. Something considered during the development of this project was the creation of a convolution reverb module to be applied on a scale based on speaker position.

5. CONCLUSIONS

The products of Spatializing Stereo consist of the Creative Upmixer, and an upcoming demonstration of its capabilities through the Northeastern University 2024

Music Technology Capstone Performance. The Creative Upmixer utilized an upmixing method based on the principles of direct and ambient sound signals developed by Carlos Avendano and Jean-Marc Jot. The Creative Upmixer further expanded on Avendano and Jot's method through the incorporation of the "Super Separator" trick demonstrated by Dan Worrall. Finally, the product provides flexible signal routing at multiple points, as well as nuanced options for panning control. The potential success in the application of the Creative Upmixer will hopefully inspire and aid future iterations of the Music Technology Capstone Performance.

Acknowledgments

I would like to formally thank the Northeastern University Music Technology Department, including Dr. Anthony De Ritis and Dr. Ronald Smith, for their support, guidance, and contributions to this project. I would also like to thank the team at the Fenway Center, Marie Siopy, Miriam Rosenau, Jeremy Reger, Arthur Rishi, and Eric Dana, for allowing the Music Technology Department to use their space and for allowing me to scope out the venue prior to the performance. Finally, I would like to acknowledge the 2024 Northeastern University Music Technology graduating class for their hard work and contributions towards the overall capstone project.

6. REFERENCES

[1] Theatre Projects Consultants. "Northeastern University, Fenway Center." theatreprojects.com. https://theatreprojects.com/project/northeastern-university-fenway-center/ (accessed Apr. 20, 2024).

[2] C. Avendano and J. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. II-1957-II-1960. doi: 10.1109/ICASSP.2002.5745013.

[3] Cycling '74. MSP Delay Tutorial 6: Comb Filter. [Online] Available: https://docs.cycling74.com/max8/tutorials/15_delaychapter06

[4] DolbyAtmos® Specifications. Accessed: Apr. 24, 2024. [Online]. Available: https://professional.dolby.com/siteassets/cinema-products---documents/dolby-atmos-specifications.pdf

[5] Genelec. G LearningLab | Immersive speaker setups. Your room and the scalability of our systems. (Mar. 24, 2022). Accessed:Apr. 24, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=mhpbZAo2L_w