Upmixing to Multichannel Audio

Mitchell Leibowitz
Aalto University
Master's Programme CCIS / AAT

mitchell.h.leibowitz@aalto.fi

1 Introduction: What is Upmixing?

Most music recordings are made in stereo, which would usually be designed to be listened to on a setup where there are 2 speakers in front of the listener or through headphones. The signal is recorded to a left and a right channel. If you wanted to play a stereo recording over a loudspeaker setup with more than 2 channels it presents the problem of how to send a 2-channel signal to an output with more than 2 channels [1]. If the music was recorded in only 2 channels, without adding additional channels to the audio there would be no actual benefit to playing the audio over multiple speakers [2]. Multi-channel audio is commonly used in movie theatres or in surround sound home entertainment systems. However, with the widespread usage of stereo recordings for many applications it is still desirable to have a stereo version of the recording even if a multi-channel version exists [3].

One of the main limitations of stereo audio is that the back sources are missing so it makes it challenging to fully immerse the listener in the sound. The rear channels are important to reproduce properties such as ambient reflections that would help to give a more realistic impression of the space [4].

Stereo sources can create a “phantom source” between the loudspeakers, but this will only work if the listener is positioned correctly in the “sweet spot” between the loudspeakers. Outside of this area the stereo image will move to the closer loudspeaker. The optimum listening area will depend on how much the “phantom source” can be shifted. [5]. It has been known since the 1930s that sending two channels to a 3 loudspeaker configuration can make spatial image more stable over a larger listening area. This is valid for a setup with the third loudspeaker placed in the center, where the middle channel is an average of the left and right channels with a gain factor included. However, the improvement compared to two loudspeaker arrangements hasn’t been significant enough to make it widely used [6].

There is an increasing interest in broadcasting multichannel formats and new audio formats that would allow the user to modify the mix of music to match their own preferences. One popular technology for multi-channel streaming is Dolby Atmos, which sends out the audio signal with each source represented as a different object in the file [7]. Many DVDs and Blu-Ray offer an option for the audio in at least 5.1 sound [4].

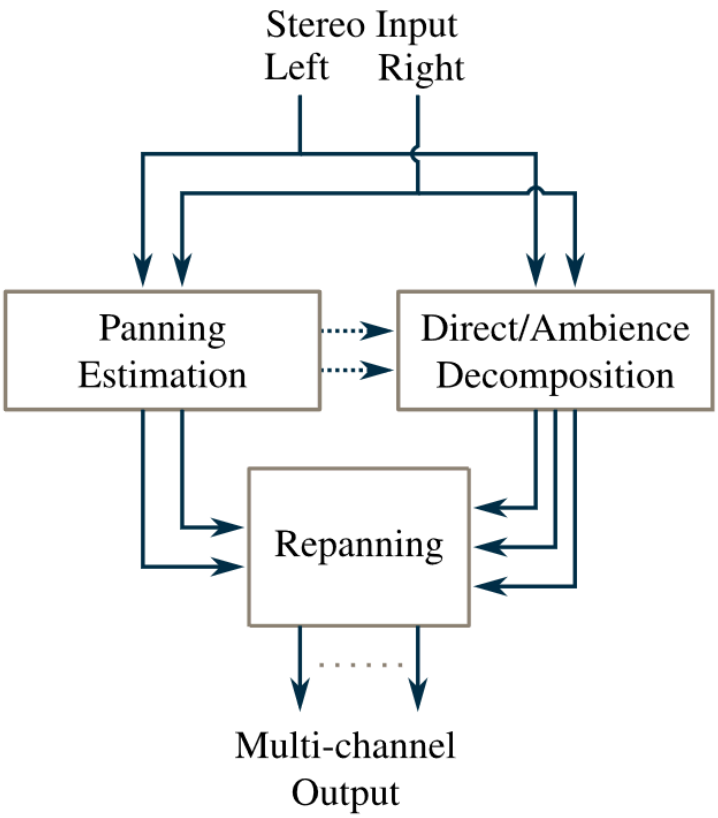

Figure 1: Exemplary processing steps in a typical stereo to surround upmix algorithm [4]

2 Challanges with Upmixing

The upmixing algorithm will depend on how the audio source was recorded and what the audio file consists of (music, speech, ambient sounds). Stereo recordings are generally either live recordings or studio recordings. In a studio recording different sources are individually recorded and then mixed to a single stereo channel. Reverberation will be added artificially.

In a live recording microphones are spatially distributed to capture sound sources. The ambience will be naturally included in the recording. The ambience can also be introduced from other noise sources in the environment. Separating elements from the left and right channels of audio, as well as separating ambient components from the channels and determining background noise has to be done in order to upmix a recording. Primary components express spatially localizable sounds while

ambient components are usually distributed more diffusely [8]. There is usually very little information about how the stereo mix was created so models usually rely on more general information about how stereo audio is recorded [1]. Another challenge is how to store and playback a multi-channel signal. Options such as Binaural Cue Coding (BCC) can store multi-channel side information in a mono or stereo signal [9].

Binaural Cue Coding (BCC) [10] is a method to separate information for "the spatial perception of multi-channel audio signals and the basic audio content." It takes multichannel audio signals and considers them as a mono signal with BCC parameters. The mono source is the sum of the sound sources which are part of the spatial image. BCC is a way to reduce the bit rate for encoding stereo and multi-channel encoders nearly at the rate of a mono audio encoder. It can also be used to enhance mono broadcast systems.

Many algorithms are based on separating primary and ambient components from the mix [11]. There are various ways of doing this including Principal Component Analysis (PCA) based methods, least square methods, and spectral-based approaches, adaptive filter based methods. These methods will be discussed further in Section 3.

Unless a stereo and multichannel version of the same recording were made it would require an up-mixing algorithm to add the additional channels to the audio. Even if a multichannel version of a recording exists there could be a vast number of different multichannel speaker setups on which the audio could be played back on [3]. Many upmixing algorithms intend to preserve the integrity of the original mix without adding coloration to the signal. It should be considered if upmixing is enhancing or distracting from the mood of the original piece [2].

3 Overview of Methods for Upmixing

3.1 Frequency Domain Techniques For Stereo To Multichannel Upmix

The paper used a spectral-approach to separate primary and ambient sources. They used the inter-channel short-time coherence. By using the cross-correlation and autocorrelation between the stereo channels they were able to estimate the panning and ambience index [11]. They compared the STFT (Short-Time Fourier Transform) of the left and right stereo signals to help identify different components of the mix. They measured the inter-channel coherence to figure out the ambient contribution to the recording. The time correlation between channels will be higher in areas where the primary components of the recording are dominant and lower in regions dominated by reverberation or ambience. A non-linear mapping function was used to "extract the time-frequency regions of interest" and an inverse STFT was used to create the new signals. The nonlinear function separates regions that have a high ambience index and therefore do not need to be modified

and regions with a small ambience index, meaning they need to be reduced to remove the primary signal component. They used a measurement of similarity between the channels to determine a panning coefficient for various instruments. At each time frame they integrated the energy for frequency regions with a similar panning index to get a distribution of energy as a function of the panning index and time. The information could be used to identify the main sources in the mix and their panning coefficients.

The human hearing system determines the direction of sound in part due to the difference in signals reaching the left and right ear, known as the interaural time difference (ITD) and interaural level difference (ILD). In a mix, ambient sources will have low correlation between the two channels meaning that it will be difficult to determine exactly where the signal is coming from [11].

If a source is panned to the center then the difference of the left and right signals will "eliminate the direct-path component" and the resulting signal will only have the reverberation information. However, this only works for the signals panned to the center and will only give a monaural ambient signal.

They applied their method to create a 2 to 5 channel upmix. Their method worked mostly for studio mixes. The method was limited in situations when the relationship between the STFTs was not as straightforward and did not work as well for live recordings [1]. This might be due to their being more background noise in a live recording and it being more difficult to correctly set the resolution of the STFT to isolate sources. There might be too much of other sources blending and coloring the signal in live recordings.

3.2 Two-to-Five Channel Sound Processing

Five channel surround setups are commonly used for multichannel audio. The speaker configuration used in this paper is shown in Figure 1. The front channels are used to have accurate directional audio over a large listening area for sounds such as dialogue. The rear surround channels are used for more diffuse sounds such as environmental effects. The algorithm described can be extended to a 5.1 surround system [3] .

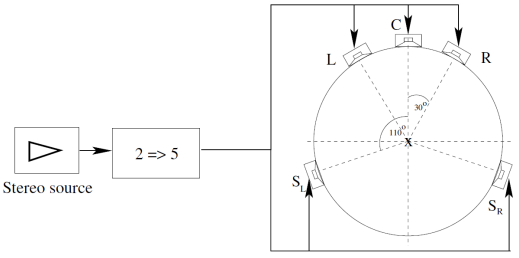

Figure 2: ITU reference configuration [9]. —reference listening position (sweet spot). Left and right channels are placed at angles 30° from C; two surround channels are placed at angles 110° from C [3]

The paper first considered a three-channel setup, where the central loudspeaker was centrally located between the left and right speaker. In this setup an additive blend of the left and right channels is sent to the center channel. However, blending the signals in this way will narrow the stereo image due to crosstalk from the left and right channels. If crosstalk is in phase it can reduce the stereo image. The left and right channels will lose some of the uniqueness of the signal so it will be perceived as less clearly directional [12]. Typically some sort of crosstalk cancellation algorithm would be necessary to address these issues and to help perceive the unique directionality of the signals [13]. This paper used Principal Component analysis (PCA) to derive the center channel and avoid the issues with crosstalk. That is because this technique produces two vectors with the direction of the main signal and the remaining signal. The two vectors can be used to create a new coordinate system. The two signals are used as the basis signals in the matrix decoding.

"Principal Component Analysis uses correlation between the two channels to extract the correlated signals from the mixture as the primary source while the ambient sources are assumed to be the residuals, which show low correlation." PCA is only suitable for extracting primary sources when their is an intensity difference between the two channels. It doesn't use time-difference information. It has low accuracy if there is no prominent primary source, for example if the recording is mostly from ambient sources. However, some variations on the method take into account additional weighting methods. [11]

In order to get the center channel's gain based on the direction of the stereo image, they processed each sample of an audio sample as a linear combination of the left and right channels. To find the correct weighting vector they tried to maximize the energy of the signal using a steepest-descent optimization method.

To compute the ambient components they took the difference between the left and right channels of the original signal and then look at how correlated the remaining signal is and how distributed it is between the two channels [3] .

3.3 Optimum Reproduction Matrices For Multispeaker Stereo

One of the simplest ways to upmix audio is by using a time domain mixing matrice with phase shifting to create additional channels. [4]

An example of this method is described by [6]. The matrix decoder for upmixing channels should preserve the total energy of all stereo inputs. This is necessary to preserve the level-balance amongst different sounds in the mix and also has psychoacoustic relevance as well." The level balance between direct sounds and early reflections in a recording environment conveys important cues about sound-source distance." So the recordings will not sound like they are the correct distance away. Additionally there are many different ways to record stereo, different stereo panning techniques use amplitude and time delay. If these recordings are sent through a network that doesn't preserve the total energy then the time-delay will

have a comb-filtering effect as the signals will add and cancel out, which will color the signal. Preserving the energy with the decoder also helps to create a wide stereo image. If the energy isn't preserved then likely the integrity of the original recording will not be maintained and the level balance between sources may be distorted.

The energy-preservation criteria led to setting up orthogonal matrices which makes finding ideal values a geometric optimization problem. It is also important for the localization to be as similar as possible to the original mix. To evaluate the localization performance they used a method based on acoustic velocity and another method based on energy and sound-intensity.

Their algorithm was intended for cases where all loudspeakers are the same distance from an ideally positioned listener.

They described their matrix decoder in terms of the sum and difference of the signals to simplify the equations, which could then be reverted back to directly left or right channels later on.

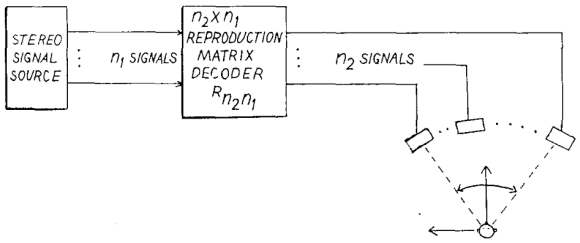

Figure 3: Schematic of n2-1oudspeaker stereo reproduction via decoding matrix from hi-loudspeaker stereo source. [6]

3.4 Automatic Audio Upmixing Based On Source Separation And Ambient Extraction Algorithms

This paper presents another method for upmixing from 2 to 5.1 format. The problem they were trying to solve with their algorithm is that extracting only the direct and primary sound sources was not enough to create good quality multichannel surround sound. They used a deep neural network that can separate 4 sources (drums, bass, vocals, other) in the mix. It can also extract ambiance information and primary information. However, they neglected the primary extracted information as they considered that much of that information would already be contained in the information from extracting the 4 sources.

In a stereo recording the primary sources will be highly correlated sounds and directional sounds. Ambient or diffuse sounds will have signals that have low correlation with no certain direction. Fig 1. describes the 5.1 reference speaker format also used for this algorithm. Many upmixing methods based on adding or subtracting the difference between the two channels, or based on adaptive filtering had issues that the results had artifacts in the audio, or didn't sound natural.

They could also only work for a certain type of source like music. However, after reviewing a variety of sources they found that the adaptive panning algorithm proposed by [3], which used Least Mean Squares to maximize energy and determine the weighting vectors corresponding to the left and right channels and the center channel gain. This algorithm performed best from several compared methods so they based their work off of this.

They first considered the right channel as the desired signal and then sent the left channel to the adaptive filter. After going through a series of iterations for the training stage they repeated the process for the other channel, using the other channel in the adaptive filter. They used a weighted combination of the bass and drum channels through a low-pass finite impulse response (FIR) filter for the low-end channel to send to the sub-woofer. For the front channels they used a high-pass FIR filter as well as a combination of the original audio and the vocal source in order to get a wider distribution for speech, which could be particularly helpful for dialogue from film sources. They also had only the ambient components sent to the back channels.

Least mean squares methods can also be used to extract primary and ambient sources to help minimize the error between the extracted signal and the original signal. [11]

Adaptive filtering "equalizes two input signals with respect to both spectral magnitude and phase before differencing to remove correlated components." Many least-mean square based algorithms determine the gain of a principle sound for relevant frequency bands. An issue with these approaches occurs when the input signals have maximum correlation. When this occurs the algorithm cannot cancel the correlated sound components and direct sound will color the ambient channel. Adaptive filtering by using a normalized-least-mean squares (NLMS) adaptive filter to "align the input signal both spectrally and temporarily before differencing" which allows the correlated components to be removed [14].

3.5 Directional Audio Coding In Spatial Sound Reproduction And Stereo Upmixing

Directional audio coding (DirAC) is a way to represent spatial sounds. It can be used for arbitrary audio reproductions methods. DirAC can also send spatial information as side information to monophonic audio channels. Methods such as Ambisonics can use any loudspeaker setup to map spatial audio, however current technical limitations in ambisonic microphones cause the spatial image to be blurred. Using higher order ambisonics could provide better quality, but will increase the cost due to the large number of microphones required.

Figure 2 demonstrates how DirAC works. It is based on the assumption that temporal and spectral resolution of the signal processing should model the spatial hearing of the auditory system. In this case the microphones signals are split into

frequency bands which are based on how the inner ear decomposes frequencies. The direction of arrival and diffuseness have comparable accuracy to the auditory system at each frequency band. The number of transmitted channels can vary depending on the application. The directional analysis was based on analyzing the energy of the sound field. Two different methods are used to produce the diffuse sound. The first method is to convolve the signal with exponentially decaying white noise to decorrelate it, each loudspeaker channel is decorrelated separately. This can produce artefacts. The other option would be to send all of the channels of the diffuse sound and get a virtual cardioid microphone which points towards each loudspeaker direction. The difference between this sort of cardioid diffusion is similar to first order ambisonics. However, in DirAC only the loudspeaker signals are diffused, which helps to minimize some of the coloration problems that can occur in ambisonics [15].

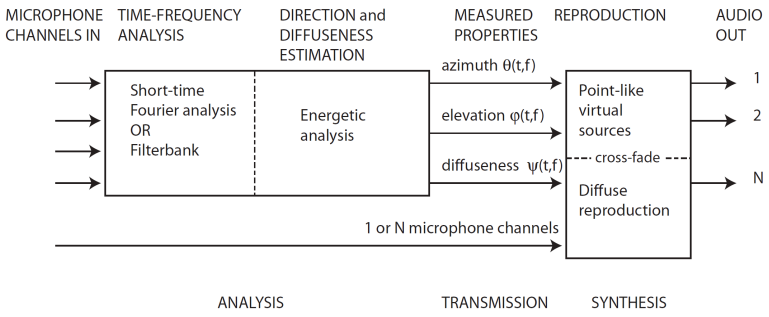

Figure 4: Overall flow diagram of Directional audio coding. [15]

Bformat audio is the main audio format used for higher order ambisonics. It is a multichannel audio format where each channel does not correspond to a specific speaker. Instead the channels contain information about the soundfield and spatial information of how each source is distributed. This information is used later in the decoding process to translate into specific audio streams that can be sent to a speaker output [16].

When using DirAC for upmixing, the left and right channels are applied to simulated loudspeakers and a B-format microphone is simulated in the optimal far-field listening position. The far-field is approximately a distance of 1 wavelength away from the the sound source. At this distance the source can be modelled as a point and the wave front can be approximated as a plane-wave. The sound wave may be simpler to measure and model in the far-field compared to the near-field [17].

4 Listening Tests & Audio Evaluation

There are few audio samples available that can be help validate the results of upmixing algorithms. Listening tests are a common way to evaluate audio quality. The MUSHRA test is a common listening test to determine audio quality. [2] used the MUSHRA listening test to determine how successful their up mixing algorithm was. The MUSHRA test compares a reference audio sample to a high

quality example to determine how noticeable impairments in audio quality are. The reference sample could be also an unaltered version of the recording such as the original stereo mix. It could also be something like a 5.1 channel mix from a film soundtrack, or some recording that was made and already mixed for a multi-channel setup. Typically the MUSHRA test will also include a low-quality version of the recording possibly with some artifacts included in order to tell if listeners are able to perceive the artifacts in the recording. If the algorithm produces some distortion to the original signal but listeners are not able to perceive it, it may be acceptable. Not every application needs the highest possible audio quality anyways. In some cases such as broadcasting certain compromises will have to be made due to technical limitations of how much data can be sent and retrieved over the network [18].

Measuring the signal to distortion ratio (SDR) is another metric to evaluate the quality of upmixing algorithms. SDR is generally used to evaluate the overall audio quality of an audio source. A limitation of this method is that it only tells how good an audio source is compared to a reference. additionally it can be difficult to compare upmixing algorithms as they may be designed to work with different types of audio sources. For example, [2] was trying to measure the SDR of their method compared to several other methods including a method that used Vector-Based Amplitude Panning (VBAP). A difficulty they encountered trying to compare the different methods was that their method produced a 6 channel signal while some of the other methods they were evaluating produced a 5 channel signal, which made it difficult to compare the audio fairly. Additionally the algorithms had different requirements for how the audio should be processed as an input [2].

To validate their approach from 2 to 5 channel upmixing [3] used a listening test. They had the subjects listen to music encoded to multichannel audio with different methods and then they indicated their preference for which encoding method they preferred. The subjects were listening to the audio both in the sweet spot and 1m away from it. They listened to the samples in an acoustically dry room. Which allowed subjects to judge localization and crosstalk between the channels. They compared the method they described in their paper as well as a variation with more low-frequency to the surround system and some other commercially available speakers. The listening tests used samples from movies and music recordings.

The audio produced by DirAC was analyzed through listening tests done in anechoic conditions. “The test played back 16-channel 3-D reproductions of virtual reality generated with 16 loudspeakers.” The test conformed to ITU BS.1116-1 for a listening room with a 5.0 system. A virtual acoustic environment was generated with different versions of DirAC. A small, medium, and large room were generated for the virtual environments. A snare drum or male speech exert were convolved with the impulse response from the virtual environment, which created natural sounding samples. The B-format recording was simulated with the samples being reproduced with DirAC.

DirAC reproduces the spatial properties of sounds. The timbre of the reflections was slightly changed. Their can be a timbral change between the reference and reproduction. Depending on the source this could be more or less noticeable.

Another test involved listening to the diffuse and non-diffuse parts of B-format recordings separately. The diffuse reproductions were missing transients and the direction of the sound sources was not clear. The artifacts were not audible though. They found that with the non-diffuse sound if the constants were set incorrectly there could be some artifacts. When compared to a hypercardiod Ambisonic setup the reverb of the rooms sounded wider with DirAC than with Ambisonics [15].

5 Conclusion

This paper has presented an introduction to what upmixing is and why it is used. A summary was given of some of the common challenges for upmixing audio from stereo to multi-channel formats. The specific upmixing algorithm used will depend on the intended multichannel setup, the starting source of the audio and the desired effect to be created from converting the audio to a multichannel setup. An overview was given of various methods used to upmix audio. Many of these methods rely on separating directional information from the left and right channels of the recording, determining how correlated or separated information is between the channels and using this to determine primary or ambient effects in the audio. Then typically the algorithm will weight the signals to be properly sent out to the new loudspeaker setup. Some methods also rely first on mapping the audio to a 3D spatial sound field and using this information to convert to a signal that can be sent out. There is the difficulty of maintain the integrity of the original recording and not introducing artifacts or colouration in the upmixing process. Many upmixing algorithms will use listening tests to evaluate the success of their algorithm. Some of the most common methods for upmixing audio are time domain mixing matrice with phase shifting, least squares based methods, spectral-based approaches extracting panning data from the STFT of the channels, and adaptive filter based methods.

5.1 Notes on AI Usage

I used ChatGPT to find sources. For example, determining some keywords that might be relevant to the papers I was looking at. Or trying to find sources that would describe summaries of common ideas/topics that are mentioned in the paper, but not explained in detail. Sending queries like “Where could I find sources that would explain Primary and Ambient sources in audio recordings.” Using it to point me towards sources, rather than relying on anything it said as valid information.

References

[1] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, 2002.

[2] M. Negru, B. Moroşanu, A. Neacşu, D. Drăghicescu, and C. Negrescu, "Automatic Audio Upmixing Based on Source Separation and Ambient Extraction Algorithms," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 12–17.

[3] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, 2002.

[4] S. Kraft and U. Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *18th International Conference on Digital Audio Effects (DAFx)*, 2015.

[5] S. Merchel and S. Groth, "Analysis and implementation of a stereophonic play back system for adjusting the "sweet spot" to the listener's position," in *Audio Engineering Society Convention 126*, 2009.

[6] M. A. Gerzon, "Optimal reproduction matrices for multispeaker stereo," in *Audio Engineering Society Convention 91*, 1991.

[7] J. M. DeFilippis, "Mastering and Distributing Immersive Sound," *Immersive Sound Production*. Focal Press, pp. 123–140, 2022.

[8] L. McCormack and A. Politis, "Estimating and reproducing ambience in ambisonic recordings," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 314–318.

[9] H.-G. Moon, "A low-complexity design for an MP3 multi-channel audio decoding system," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 314–321, 2011.

[10] C. Faller and F. Baumgarte, "Binaural cue coding: a novel and efficient representation of spatial audio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, p. II–1841.

[11] K. M. Ibrahim and M. Allam, "Primary-ambient source separation for upmixing to surround sound systems," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 431–435.

[12] Hugh Robjohns, "Q. Is Crosstalk a good thing?." [Online]. Available: https://www.soundonsound.com/sound-advice/q-crosstalk-good-thing

[13] Y. Huang, J. Benesty, and J. Chen, "On crosstalk cancellation and equalization with multiple loudspeakers for 3-D sound reproduction," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 649–652, 2007.

[14] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2141–2150, 2007.

[15] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, 2006.

[16] Blue Ripple Sound Limited, "HOA Technical Notes - SN3D B-Format." [Online]. Available: http://www.blueripplesound.com/notes/bformat

[17] Scott MacDonald, "Sound Fields: Free versus Diffuse Field, Near versus Far Field." [Online]. Available: https://community.sw.siemens.com/s/article/sound-fields-free-versus-diffuse-field-near-versus-far-field

[18] I. BS, "1534-3,"Method for the subjective assessment of intermediate quality level of audio systems„" *International Telecommunication Union, Geneva, Switzerland*, 2015.