Parametric and Adaptive Reproduction of Ambisonic Signals

Vaibhav Talwadker
Aalto University
Master's Programme CCIS / AAT

vaibhav.talwadker@aalto.fi

Abstract

The importance of spatial sound reproduction methods has increased with the widespread development of VR and AR technologies. Parametric reproduction methods exhibit an advantage over non-parametric methods in terms of flexibility of their input and scalability, due to the adoption of Ambisonic signals as their input. This review article studies the analysis and synthesis pipelines of four such parametric methods and their evaluation results. Since parametric methods operate in the time-frequency domain, the artifacts arising from their processing can be quite severe due to the time-variant modification of the reproduced spectrum. In addition to the studied methods, the article briefly explores one technique called covariance domain rendering that has been suggested to mitigate processing artifacts.

1 Introduction

Spatial sound reproduction algorithms are required for conveying spatial cues that mimic the original recorded scene, which include synthesizing a desired perception of the sound. Existing physical approaches such as Wave Field Synthesis recreate the entire sound scene but require a large number of loudspeakers to do so. Parametric methods bypass this issue by extracting time-dependent parameters from the signal based on some assumptions about the structure of the sound field and its properties. For the purpose of spatial sound reproduction, it is important to have some knowledge about the recording setup such as the geometry of the microphone array, the microphone directivity patterns and array orientation. Alternatively, the microphone signals could be processed in the spherical harmonic domain, thus, distinguishing the spatial properties of the sound scene from the array [4].

Many parametric methods estimate a dominant directional cue and another cue describing the dominant directional/reverberant ratio. Directional Audio Coding(DirAC) is one such method that performs energetic analysis of the sound field recorded using a B-format microphone,also described as a First Order Ambisonics(FOA) signal. The analysis in case of the FOA signals is relatively straightforward but due to their limited spatial resolution, spectral and spatial issues are observed in the reproduced sound. This led to the development of methods such as Higher-Order DirAC and COMPASS that utilize Higher Order Ambisonics(HOA) signals, providing increased spatial resolution. The remainder of this review article studies each of these four methods and their evaluation results.

2 First Order DirAC

2.1 Analysis

While DirAC works with a variety of microphone arrays,this section focuses on DirAC using B-format signals, also termed as First Order DirAC [8]. The B-format signal consists of an omnidirectional pressure pattern(b_w) and three dipole patterns, b_x, b_y, b_z, that approximate the pressure-gradient which is proportional to the particle velocity vector. Using this information, the intensity vector is formulated as the product of the channels corresponding to the omnidirectional and dipole patterns. This quantity proves to be essential for estimating the two main parameters in DirAC - Direction of Arrival(DoA) and Diffuseness(ψ). The omnidirectional signal is reproduced using Vector Base Amplitude Panning (VBAP) in the direction of the most prominent DoA. For reproduction, accurately analysing the DoA is mainly necessary in scenarios with a dominant plane wave so that its direction can be reproduced effectively. Meanwhile, the diffuseness coefficient is bound between 0 and 1, describing plane waves($\psi = 0$) and diffuse reverberation($\psi = 1$).

2.2 Synthesis

The analysed diffuseness coefficient is used to divide each frequency band into two streams - a direct stream which should ideally consist of only plane waves arriving at the microphone and a diffuse stream containing sound caused by surrounding reverberation. The diffuse and non-diffuse streams are weighted with coefficients $\sqrt{\psi}$ and $\sqrt{(1-\psi}$ respectively. DirAC addresses the high coherence problem in FOA by executing distinct operations for each stream. In the direct stream the number of loudspeakers producing a coherent signal is limited to a minimum of three and two for 3D and 2D cases respectively. Meanwhile, in the diffuse stream decorrelation methods are used which ideally produce an incoherent signal compared to the input to the diffuse stream. However, decorrelation changes the temporal structure of the signal and produces noticeable artifacts in impulsive signals such

as applause and speech. In certain cases, the non-diffuse sound may leak into the diffuse stream, producing considerable artifacts after decorrelation such as incorrect distance perception of sources and spectral filtering of sources with a room-like response. Minimizing the amount of decorrelated energy is achieved by a method known as covariance domain rendering [12].

For each time-frequency tile, the covariance rendering method provides an optimal mixing solution as a function of three matrices

- Input Signal Covariance Matrix (C_x) : This is the covariance matrix of the B-format input signal.
- Target Covariance Matrix (C_y) : This matrix is obtained by adding the non-diffuse(C_y^{ND}) and diffuse target covariance matrices(C_y^D). C_y^{ND} corresponds to the covariance matrix of the panning gains multiplied by $(1-\psi)$ and the total sound energy,R, of the current time-frequency tile. C_y^{ND} is a diagonal energy distributor matrix,D, multiplied by ψ and R.
- Prototype Matrix (Q) : This matrix depends on the microphone and loudspeaker configuration. For a B-format input signal, Q could be a virtual directional microphone matrix with the look directions pointed towards the loudspeakers. Meanwhile, a beamforming matrix could be suitable prototype matrix for a spaced omnidirectional microphone array input.

The listening test compared covariance rendering to the legacy DirAC rendering for a B-format input signal, a spaced array with 4 channels and a single omnidirectional microphone. The results shown in Figure 1 indicated that covariance rendering was specially useful in the case of multiple microphone signals for stimuli containing applause or speech.

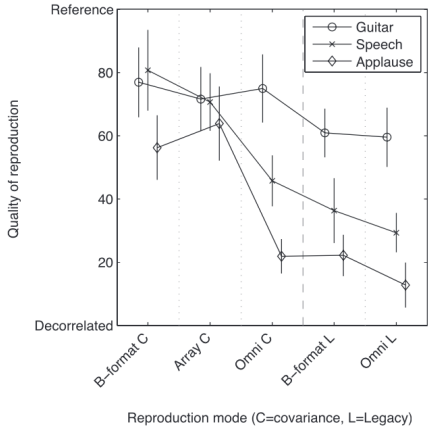


Figure 1: Mean and 95% intervals of the score of reproduction modes normalized with respect to the reference and decorrelated anchor [12]

2.3 Virtual Microphone DirAC

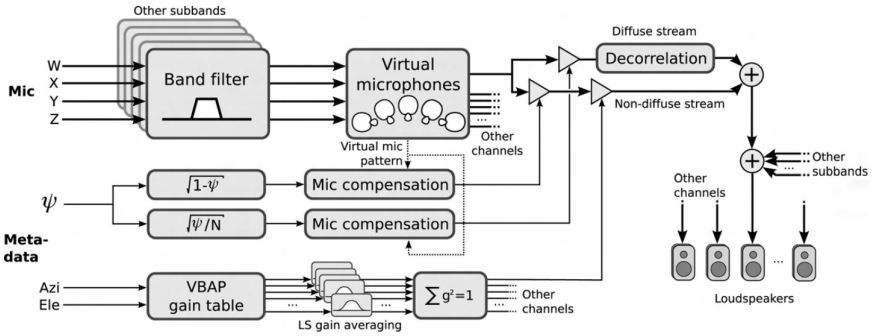


Fig. 4. Synthesis of loudspeaker signals in DirAC with B-format microphone signal and virtual directional microphones.

Figure 2: VMDirAC flow[11]

In applications favoring efficiency over quality such as teleconferencing, synthesis is usually achieved by simply distributing the microphone pressure signals to the loudspeakers according to the analysis parameters. It was found that using a linear combination of the B-format channels, termed as a virtual microphone [11], resulted in better reproduction of spaciousness, sound color and source localization. Fig 2 describes the analysis in Virtual Microphone DirAC (VM-DirAC). The VM signals are computed from bandpassed B-format signals and multiplied with the following factors for obtaining the non-diffuse and diffuse stream signals respectively.

$$\frac{1}{\sqrt{1+\psi(1/Q-1)}}\sqrt{1-\psi} \tag{1}$$

$$\sqrt{Q}\sqrt{\psi/N} \tag{2}$$

where N is the number of loudspeakers and Q is the microphone directivity factor The left hand terms in both equations (1) and (2) are gain coefficients that compensate the energy loss due to attenuation differences between plane waves and diffuse sound in microphones. The non-diffuse gain weighting limits the number of loudspeakers that the virtual microphone is applied to, thus reducing the high coherence between loudspeaker signals observed in FOA. But with multiple sources, DoA estimates can be incorrect and the VM may not be applied to the true sources,thus reducing the overall reproduced sound level.

For perceptual evaluation, reference scenarios were created in an anechoic room by convolving dry recordings with impulse responses of four different types of rooms modelled using DIVA [9]. These scenarios were measured using a simulated,ideal B-format microphone and a real B-format microphone as well,which was reproduced by eight reproduction methods. The results of a listening test with 14 participants can be found in Figure 3. It can be seen that DirAC outperforms conventional Ambisonics reproduction for both, simulated and real B-format inputs. DirAC's best performance is observed in reverberant spaces and free field scenarios. The quality

of the reproduced sound was dependent on the prominence of early reflections and the sound quality of the source signal. For example, transient signals such as drum sounds produced noticeable defects in reproduction.

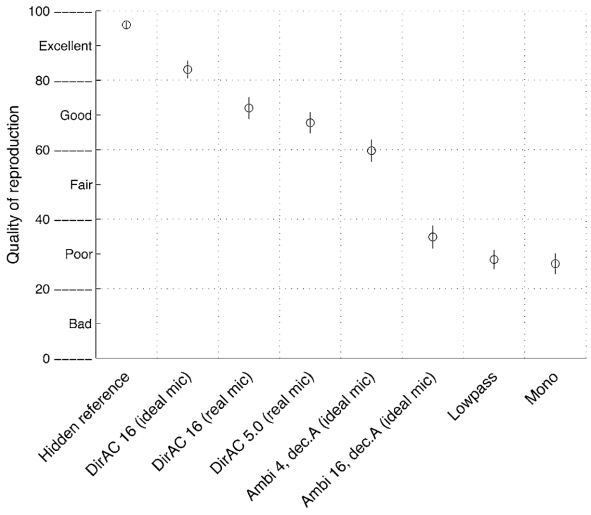


Figure 3: Mean opinion scores and 95% confidence intervals testing different reproduction methods in an anechoic chamber (VM-DirAC) [11]

3 Higher-Order DirAC

3.1 Analysis

Higher-order DirAC spatially weights the HOA signal into sectors which provides better results for complex sound scenes compared to First-Order DirAC [5]. In sector processing, parameter analysis is conducted independently which means that each sector has its own local active intensity vector, energy density and local diffuseness estimate. However, this is only possible if the order of the microphone array is greater than 1. Favorable sector designs include the minimal number of sectors that fulfill the criteria of axial symmetry and energy preservation. For a sector of analysis order N, uniform arrangements of points on a sphere known as t-designs can be used with a degree of $2N - 2$ to meet the sector design conditions [7]. For the synthesis, HO-DirAC utilizes the optimal mixing solution defined in Section 2.2. Based on the assumption that the directional component is uncorrelated to the diffuse component, the target spatial covariance matrix is calculated for each of the components for each sector. The directional component is assumed to be maximally concentrated at the analyzed DoAs, using panning function for loudspeaker rendering, and HRTFs for binaural rendering. The HO-DirAC processing block diagram is presented in Figure 4.

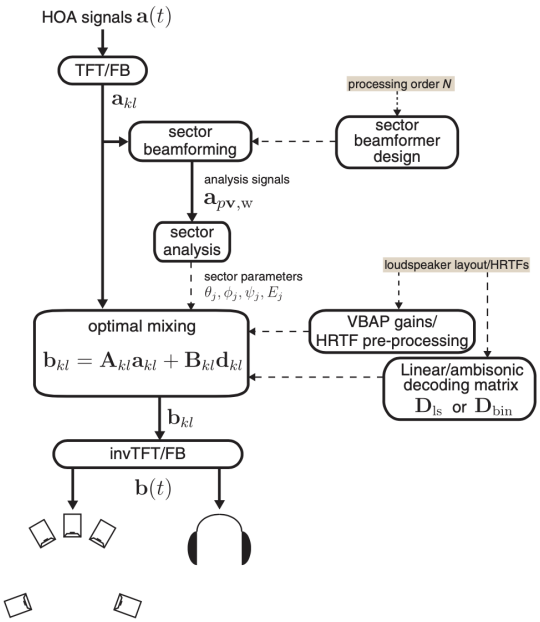
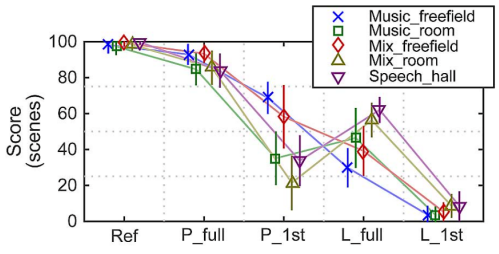
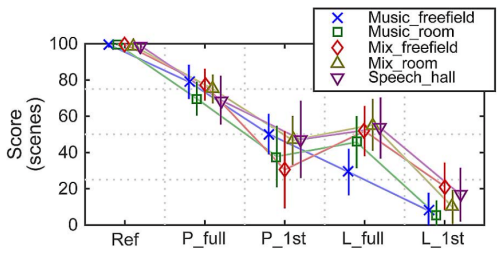


Figure 4: Block Diagram for Higher Order DirAC processing [5]

A listening test was conducted to compare higher-order,first order and Ambisonics reproduction to a reference 28-channel recording [7]. There were five reference scenes including two cases in free-field, two in a room and one in a hall. The reference signals were encoded as ideal fourth order SH signals in one test and as simulated microphone signals with added gaussian noise at 45dB SNR. The results in Figure 5 show that for the ideal SH signals,fourth-order parametric reproduction sounded almost the same as the reference and for noisy microphone signals,fourth order reproduction was closer to the reference than first-order parametric reproduction and Ambisonics reproduction.



(a) Mean and 95% confidence interval for the listening test with ideal signals



(b) Mean and 95% confidence interval for the listening test with noisy signals

Figure 5: HO-DirAC Listening Test Results. P and L denote parametric and linear reproduction respectively [7]

4 HARPEX

HARPEX is proposed as a method that combines parametric decoding with linear decoding, in order to reduce artifacts observed in parametric decoding of FOA signals [1]. HARPEX processes first-order B-format signals by utilizing overlapping windows, zero padding, and Fast Fourier Transform (FFT) to derive eight numerical values representing each time/frequency segment of a FOA signal. The decomposition is defined as the product of two matrices,V containing real-valued unit vectors that point to the DoA of the plane waves, and matrix A containing the complex amplitude of each plane wave. In some scenarios such as isotropic noise fields, the decomposition may not exist which warrants other decomposition methods. Reproducing the parametrically decomposed DoAs could be done using a panning function that generates a loudspeaker signal by multiplying the panning weight with the complex amplitude of the plane wave. But DoA vectors can rapidly change between time frames which would require smoothing to avoid time-domain artifacts. Since this alters the DoA of the plane waves, two more plane waves would need to be added for the decomposition to be valid again.

The listening test setup consisted of 12 loudspeakers forming an octagon and standard ITU5.0 layout. Six scenes were decoded with 1st and 3rd order max-RE to the octagon(1-8, 3-8 in Figure 6) and 1st order HARPEX followed by 3dB pairwise panning to both, the octagon and the ITU5.0 layout (H-8,H-5). To account for differences between pairwise panning and ambisonics decoding, one system(H-3-8) upmixed a first order signal to 3rd order using HARPEX followed by max-RE decoding to the octagon.

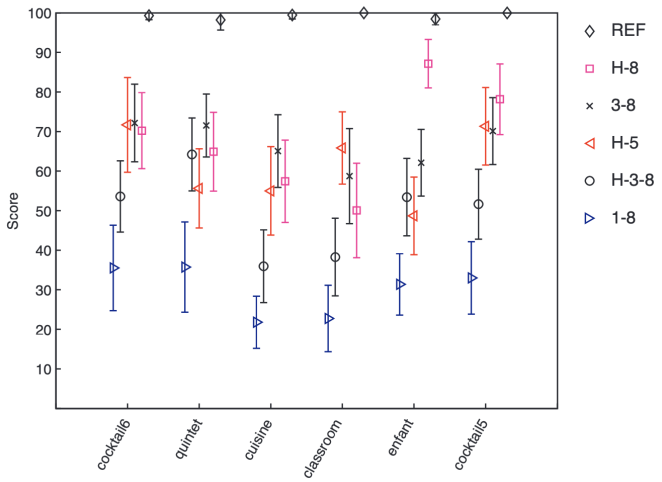


Figure 6: Mean and 95% confidence intervals for HARPEX listening test [1]

5 COMPASS

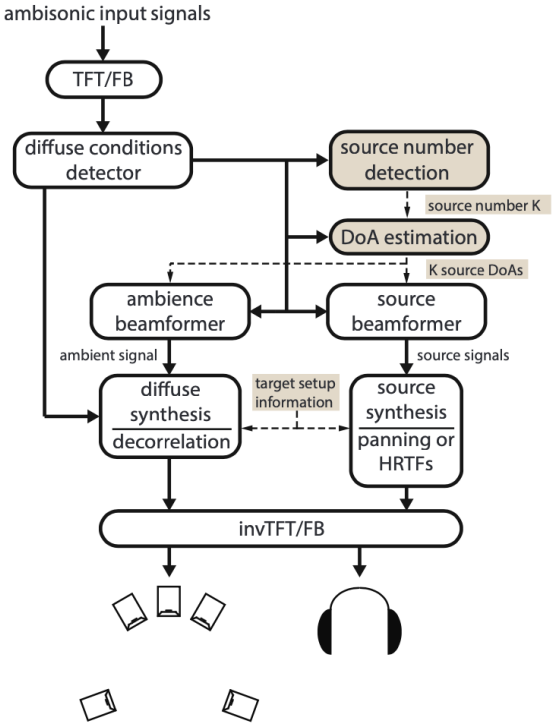


Figure 7: Block Diagram for COMPASS [6]

COMPASS is a signal-dependent reproduction method that stands for Coding and Multidirectional Parametrization of Ambisonic Sound Scenes. The acoustic model of COMPASS assumes multiple sound sources as foreground signals and ambient sound as background signals. Compared to other methods, this ambient sound does not have to be assumed as isotropically diffuse. COMPASS aims to improve reproduction using the flexibility offered by Ambisonics and is not limited to FOA, as seen in the case of HARPEX. The next paragraph explains the analysis and synthesis steps seen in Figure 7.

In COMPASS, parameters are analysed by an eigenvalue decomposition(EVD) of the PSD matrix. The PSD matrix is calculated, assuming that both the direct and ambient components are uncorrelated. The resulting eigenvalues from the EVD are sorted and used in calculating the number of sources by the SORTE method [2]. The diffuseness coefficient ψ, as seen earlier in DirAC, is calculated from the variance of the eigenvalues. Once the number of sources is known, their DoA is estimated using another statistical method called MUSIC [10]. During synthesis, the foreground source signals are first spatialized by a linear Ambisonics decoder followed by an adaptive matrix. While linear decoding effectively addresses diffuse signals, further decorrelation may be required for low-order signals, where decoded outputs may exhibit coherence. Lastly for rendering, COMPASS utilizes three control parameters

that temporally smooth the synthesis matrix, vary between fully linear decoding and adaptive decoding and modify the source-ambience ratio to be frequency dependent.

For the evaluation, three MUSHRA tests for headphone playback were conducted with 12 listeners for testing overall, spatial and timbral quality. As the stimuli, five sound scenes were simulated including three free field and two reverberant scenes. A binaural spatial room impulse response (SRIR) was obtained for each source in the scene. Convolving each of these with the appropriate HRTFs provided a reference binaural version of the scene. The test results in Figure 8 demonstrated that for the same order of input, COMPASS yielded better spatial, timbral and overall perceived quality compared to Ambisonics.

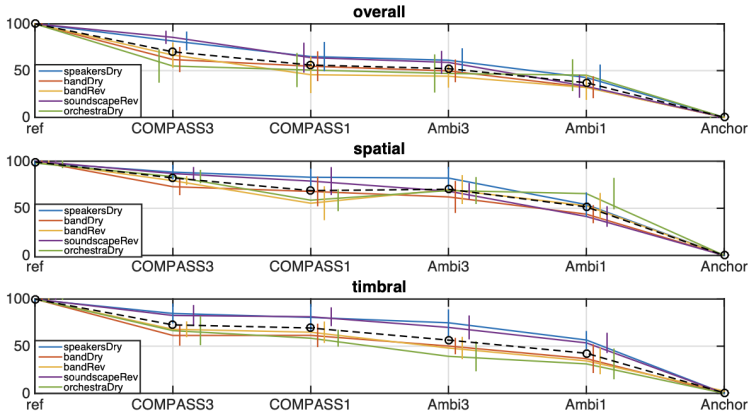


Figure 8: Mean and 95% confidence interval across subjects for the COMPASS MUSHRA test [6]

6 Conclusion

This article provided a summary of a few parametric reproduction methods for Ambisonic signals. One method which was not explored, is based on designing filters using a sparse plane-wave decomposition approach, for upscaling HOA sigals to higher order [13]. The upscaling was found to improve the listening sweet spot and fidelity of reproduction. Choosing between the parametric methods depends on the requirements of the reproduction and the type of signals available. For instance, if the reproduced sound quality has to be good for all scenarios and HOA signals are available, HO-DirAC would be a suitable option. But if computational resources are limited, first order DirAC would be more preferable. While this article focused on spatial sound reproduction, HO-DirAC has also been reformulated in recent work to be used in a spatial audio codec for compression of HOA signals [3].

References

[1] Berge, S., and Barrett, N. High angular resolution planewave expansion. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May* (2010), pp. 6–7.

[2] Han, K., and Nehorai, A. Improved source number detection and direction estimation with nested arrays and ulas using jackknifing. *IEEE Transactions on Signal Processing 61*, 23 (2013), 6118–6128.

[3] Hold, C., Pulkki, V., Politis, A., and McCormack, L. Compression of higher-order ambisonic signals using directional audio coding. *IEEE/ACM Transactions on Audio Speech and Language Processing 32* (2024), 651–665. Publisher Copyright: Authors.

[4] Politis, A., Delikaris-Manias, S., and Pulkki, V. Overview of time–frequency domain parametric spatial audio techniques. *Parametric Time-Frequency Domain Spatial Audio* (2017), 69–88.

[5] Politis, A., and Pulkki, V. *Higher-Order Directional Audio Coding*. John Wiley Sons, Ltd, 2017, ch. 6, pp. 141–159.

[6] Politis, A., Tervo, S., and Pulkki, V. Compass: Coding and multidirectional parameterization of ambisonic sound scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 6802–6806.

[7] Politis, A., Vilkamo, J., and Pulkki, V. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing 9*, 5 (2015), 852–866.

[8] Pulkki, V., Politis, A., Laitinen, M.-V., Vilkamo, J., and Ahonen, J. *First-Order Directional Audio Coding (DirAC)*. John Wiley Sons, Ltd, 2017, ch. 5, pp. 89–140.

[9] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society 47*, 9 (Sept. 1999), 675–705.

[10] Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation 34*, 3 (1986), 276–280.

[11] Vilkamo, J., Lokki, T., and Pulkki, V. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *J. Audio Eng. Soc 57*, 9 (2009), 709–724.

[12] Vilkamo, J., and Pulkki, V. Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering. *J. Audio Eng. Soc 61*, 9 (2013), 637–646.

[13] Wabnitz, A., Epain, N., and Jin, C. T. A frequency-domain algorithm to upscale ambisonic sound scenes. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 385–388.