Improved Method for Binaural Rendering Based on Parametric Processing

Chuhan Qiu[1]

[1] *Communication University of China, Beijing, China*

Parametric processing is a kind of method based on sound field modelling and perceptually motivated, which has wide applicability in spatial audio. Commonly, parametric processing utilizes adaptive filter to extract specific parameters and signals from the assumed sound field model, which is typically represented by the signals captured by microphone array. In this paper, we propose a rendering method based on parametric information processing and optimized HRTF filter, which enhances the articulation and spatial perception effect.

1 INTRODUCTION

In sound recording, we commonly use multiple microphones to reproduce the sound sources and their position in real sound field. The coverage angle, placement position, pointing angle and quantity of the microphones determine the different microphone techniques. Correspondingly, the captured signals' processing methods and the effect of them are completely different as well. Spatial audio recordings have specific microphone setups and focus on the reconstruction of the acoustic fields. In processing stage, as so called spatial audio rendering, we want to better reproduce the performance of the sound sources and acoustic field, mainly manifested as whether the position of the virtual sound source is accurate and the spatial perception is close to reality, etc. Based on these premises, many methods have been proposed to enhance the performance of spatial information, and parametric spatial processing is usually a flexible and efficient solution for most sound scene.

The stepping stone in development of parametric spatial audio techniques is based on decomposing the spatial impulse response rendering (SIRR) into one direct sound and a diffuse residual for each time-frequency bin [5]. The underlying theoretical basis of this method actually assumes that source signals tend to be sparse in the time-frequency domain, especially for signals with obvious harmonic structures, e.g. music and speech.

2 METHOD

[6] proposed directional audio coding (DirAC), whose assumed sound field model is obtained from a zeroth-order (omnidirectional) signal and two parameters: the DOA (Direction-of-Arrival) and the diffuse. DirAC is a processing solution for B-format (1st-order Ambisonics), which totally has four channels.

$$W = S \cdot \frac{1}{\sqrt{2}} \tag{1}$$

$$X = S \cdot cos\theta cos\phi \tag{2}$$

$$Y = S \cdot sin\theta cos\phi \tag{3}$$

$$Z = S \cdot sin\phi \tag{4}$$

Obviously, W signal corresponds to the omnidirectional microphone, whereas XYZ are the components along three spatial axes. In this way, DirAC utilizes XYZ to estimate the intensity vector i, and estimates the sound field energy density e in conjunction with

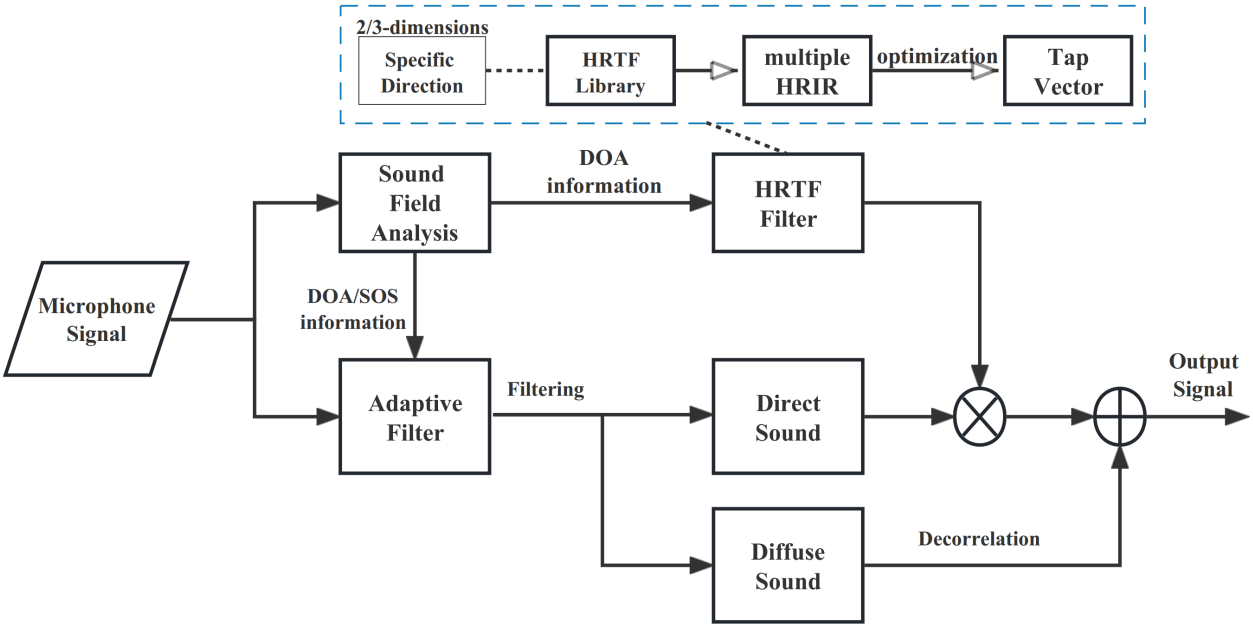

Figure 1: *Flow chart of the improved method for binaural rendering*

W. Then, the DOA and the diffuse φ can be further derived from i and e.

$$DOA = -\angle E[i] \tag{5}$$

$$\varphi = 1 - \frac{||E[i]||}{cE[e]} \tag{6}$$

where the operator $\angle$ gives the 3D angle of a vector, and $E[\,]$ is the expectation operator.

DirAC is a classical example of parametric spatial audio, which obtain a perceptive meaningful description of the recording sound field by extracting the direct and diffuse sound components and some parametric information [2]. We may use single-channel or multi-channel adaptive filter to extract the direct and diffuse signal components and other parametric information from the mono W signal or other signals captured by different microphone arrays, then updating the tap weights of the adaptive filters which enable the filters to adapt to fast changing acoustics and provide a good trade-off between robustness and attenuation of undesired signals [1].

Figure 1 shows the whole flow chart of the method proposed in this paper, and each module of it is explained separately in the following sections.

2.1 Adaptive Spatial Filter

Both Single-channel filter and multi-channel filter can be used for the extraction of direct and diffuse sound signal. To compute the tap weights of the filters, we may exploit knowledge about the DOA estimate of the direct sound or the second-order-statistics (SOS) information of the sound field [2]. Commonly, in single-channel filter, the signal-to-diffuse ratio $SDR(k,n)$ or the diffuseness $\psi(k,n)$ need to be estimated. And in multi-channel filter, more parameters are required include the DOA $\theta(k,n)$ of the direct sound, the diffuse sound power $\phi_d(k,n)$ and the PSD matrix $\Phi_n(k)$ of slowly time-varying noise.

2.1.1 Single-channel filter

Based on single-channel filter, we can estimate the direct and the diffuse components effectively, which is actually by applying a spectral gain factor to a single microphone signal.The direct signal estimation can be expressed as

$$\hat{X}_s(k,n,d_1) = W_s(k,n)X(k,n,d_1) \tag{7}$$

where $W_s(k,n)$ is the single-channel filter, which is multiplied with the reference microphone signal to obtain the direct sound at d_1.

We may obtain the optimal filter by minimizing the mean-square error between the true and estimated direct sound, which yields the well-known Wiener filter (WF).

2.1.2 Multi-channel filter

Different from single-channel filter, multi-channel filter recomputes the tap weights each time and frequency with updated information on the DOA and second-order-statistics. Multi-filter can adapt fast to changing acoustics and overcome many limitations of the single-channel filter. The direct sound estimated by multi-channel filter is

$$\hat{X}_s(k,n,d_1) = w_s^H(k,n)x(k,n) \tag{8}$$

To obtain the optimal filter w_s, there are two filter types: the linearly-constrained minimum variance (LCMV) filter and the parametric multi-channel Wiener filter. For LCMV filter, the optimal filter can be found by minimizing the noise-plus-diffuse power while extracting the distortionless response for the direct sound

$$w_{sLCMV}(k,n) = arg_{w_s} min\, w_s^H[\Phi_d(k,n) + \Phi_n(k)]w_s$$
$$s.t.\ w_s^H(k,n)v(k,n) = 1 \tag{9}$$

where the propagation vector$v(k,n)$ depends on the array geometry and DOA $\theta(k,n)$ of the direct sound.

For multi-channel Wiener filter, the loss function is the same as single-channel Wiener filter but with a linear constraint

$$w_{sPMW}(k,n) = arg_{w_s} min\, w_s^H[\Phi_d(k,n) + \Phi_n(k)]w_s$$
$$s.t.\ E\{|w_s^H(k,n)x_s(k,n) - P_s(k,n,d_1)|^2\} \leq \sigma^2(k,n) \tag{10}$$

Here, both in (9) and (10), $\Phi_d(k,n)$ is the power spectral density (PSD) matrix of the diffuse sound, which can be written as

$$\begin{aligned}\Phi_d(k,n) &= E\{x_d(k,n)x_d^H(k,n)\} \\ &= \phi_d(k,n)\Gamma_d(k)\end{aligned} \tag{11}$$

where $\phi_d(k,n)$ is the power of the diffuse sound and $\Gamma_d(k)$ is the diffuse sound coherence matrix. When assuming a specific diffuse field characteristic, each elements of $Gamma_d(k)$ is typically known a priori. In this way, to compute the loss function of the diffuse sound, we mainly need to obtain the power of the diffuse sound $\phi_d(k,n)$ and the PSD matrix of the noise $\Phi_n(k)$.

2.2 Sound Field Analysis

As what has been told in last section, the parameters need to be extracted in sound field analysis stage depend on the type of adaptive spatial filter we used

Single-channel Wiener filter signal-to-diffuse ratio $SDR(k,n)$ or the diffuseness $\psi(k,n)$

LCMV filter DOA $\theta(k,n)$ of the direct sound, diffuse sound power $\phi_d(k,n)$ and PSD matrix $Phi_n(k)$

2.3 HRTF Filter

In the application of binaural rendering, blocking the input signal and convolution it with HRIR is a very common method. However, this still causes much performance overhead compared to directly use the tap weights for filtering. A simple test in Matlab has been shown in Figure 2, which reflects the time cost between filtering by convolution and by tap weights. Obviously, in most cases, the latter takes an order of magnitude less time that the former.

Table 1: *Running time cost of different filtering method (average value)*

Filter Order	Convolution	Tap Weights
2nd-order	0.006276	0.000451
5th-order	0.006341	0.000761
10th-order	0.006375	0.000762
15th-order	0.007016	0.000861

As shown in Figure 1, we obtain the tap vectors of HRIR by optimization algorithm [7]. In rendering stage, we select the HRIR tap vector closest to the DOA direction estimated by sound field analysis module, and filtering the direct sound extracted by adaptive spatial filter.

The iteration of multi-direction HRIR is also a part of the time cost. So we may reduce the HRIR quantity by choosing more representative directions appropriately. And for 2-D reproduction rendering, we only select the horizontal plane HRIR. Only when 3-D playback is required and introduce the HRIR with evaluation. All selected direction HRIRs will be pre calculated and stored, and only need to be called during real-time rendering.

To prevent the problem may caused by phase distortion and feedback loops, the tap vectors of HRIRs only include numerator coefficients, i.e. using FIR filters.

2.4 Direct Sound Synthesis

After obtaining the corresponding direction HRIR tap weights, we synthesise the final direct sound by filtering the direct sound using HRTF filter estimated. As what has been proved in Matlab, this method could reduce much time for computing in real-time.

2.5 Diffuse Sound Synthesis

Binaural reproduction includes only two channels, but decorrelation still plays an very import role. Decorrelation is important for auditory spatial impression [8] [3], and can extend the size of the listening area in loudspeaker reproduction which is less important in binaural reproduction because the listener is always at the sweet point [3].

The decorrelation can be implemented using frequency-dependent delays, and the result is a signal that has a random delay at each frequency band but the magnitude response has not been changed[4]. To prevent the decorrelation contributing to the perception of direction or spaciousness, the delays are within eligible boundaries.

3 IMPLEMENTATION

At present, the collection of theories has been relatively comprehensive. At the same time, various factors have been considered and a large number of review documents have been used for reference. However, in the actual implementation process, several problems have emerged. Due to time constraints, a lot of time and energy have been spent in the literature research stage, so this assignment does not give a complete implementation process. The implementation ideas have been fully introduced in the previous article. The main implementation difficulties will be introduced in the appendix of this submission.

References

[1] Emanuël A. P. Habets and Oliver Thiergart. "Parametric Spatial Audio Processing: An Overview and Recent Advances". 140th International Audio Engineer Society Convention. June 2016. URL: https://www.aes.org/events/140/tutorials/?ID=4927.

[2] M. Taseska G. D. Galdo V.Pulkki Emanuel A. P. Habets K. Kowalczyk O. Thiergart. "Parametric Spatial Sound Processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction". In: *IEEE Signal Processing Magazine* 32.2 (2015), pp. 31–42. DOI: 10.1109/MSP.2014.2369531.

[3] Koichiro Hiyama Kimio Hamasaki. "Reproducing Spatial Impression With Multichannel Audio". In: vol. 19. June 2003.

[4] Mikko-Ville Laitinen and Ville Pulkki. "Binaural reproduction for Directional Audio Coding". In: *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2009, pp. 337–340. DOI: 10.1109/ASPAA.2009.5346545.

[5] K. Kowalczyk M. Cobos J. Ahrens and A. Politis. "An Overview of Machine Learning and Other Data-based Methods for Spatial Audio Capture, Processing, and Reproduction". In: *EURASIP Journal on Audio, Speech, and Music Processing* 1 (May 2022), pp. 4477–4479. DOI: 10.1186/s13636-022-00242-x.

[6] Ville Pulkki. "Spatial Sound Reproduction with Directional Audio Coding". In: *Journal of the Audio Engineering Society* 55.6 (2007), pp. 503–516. ISSN: 1549-4950.

[7] Renjith V. Ravi et al. "Optimization algorithms, an effective tool for the design of digital filters; a review". In: *Journal of Ambient Intelligence and Humanized Computing* (2019). DOI: 10.1007/s12652-019-01431-x.

[8] T. Okano T. Hidaka and Leo Beranek. "Interaural cross correlation (IACC) as a measure of spaciousness and envelopment in concert halls". In: *Journal of The Acoustical Society of America - J ACOUST SOC AMER* 92 (Oct. 1992). DOI: 10.1121/1.404472.