Audio Engineering Society

Conference Paper

Presented at the International Conference on
Spatial and Immersive Audio
2023 August 23–25, Huddersfield, UK

Hearing from Within a Sound: A Series of Techniques for Deconstructing and Spatialising Timbre

Lewis Wolstanholme[1], Cyrus Vahidi[1], and Andrew McPherson[2]

[1]*Centre for Digital Music, Queen Mary University of London*
[2]*Dyson School of Design Engineering, Imperial College London*

Correspondence should be addressed to Lewis Wolstanholme (`l.wolstanholme@qmul.ac.uk`)

ABSTRACT

We present a series of compositional techniques for deconstructing and spatialsing timbre in an immersive audio environment. These techniques aim to engulf a spectator within a given abstract timbre, by highlighting said timbre's distinct spectral and gestural characteristics through our approach to sound spatialisation. We have designed these techniques using both additive synthesis, and time-frequency analysis and resynthesis, building upon analytical methods such as the discrete Fourier transform and the joint time-frequency scattering transform. These spatialisation techniques can be used to deconstruct a sound into subsets of spectral and gestural information, which can then be independently positioned in unique locations within an immersive audio environment. We here survey and evaluate how perceptibly cohesive and aesthetically nuanced a timbre remains after deconstruction and spatialisation, when applied in both live performance and studio production contexts. In accordance with their varying design, each spatialisation technique engenders a unique aesthetic experience, affording a listener various means through which to hear from within a sound.

1 Introduction

Over recent years, spatial and immersive audio has become an increasingly prevalent feature in both musical and sound art practices. With many music venues and gallery spaces now beginning to rapidly adopt and install complex speaker arrays, these institutions present a unique opportunity for artists and sound engineers to both explore and expand upon their curatorial approaches to the dispersal, projection and positioning of sounds within a space. Stemming from our own creative experiences in these environments, we here present, analyse and compare three techniques for deconstructing and spatialising timbre. For a given audio source or abstract timbre, each of these techniques enables their sonic distribution throughout a space, creating an auditory landscape that both surrounds and engulfs a listener, with each technique serving to perceptually highlight particular spectral and gestural characteristics present in the original sound.

We base the first of our three spatialisation techniques on additive synthesis, with the subsequent two techniques developing upon the discrete Fourier transform (DFT) and the joint time-frequency scattering transform (JTFS). Through our explorations of these techniques, we here analyse their perceptive effectiveness and aesthetic nuance when applied in both live performance and studio production contexts. And, in doing so, it is our aim with this work to exhibit our aesthetic knowledge surrounding these techniques, influencing their incorporation in the future works of artists and sound engineers alike.

2 Background

When approaching spatial audio from a curatorial perspective, it is often the desire of the artist or sound engineer to engender an immersive listening experience for their spectators. "[I]mmersion describes omnidirectional, enveloping qualities ascribed to a specifically sonorous experience or sensibility. To be immersed in sound implies embodied presence…" [1, p.2] In our interpretation of Schrimshaw [1], we have here sought to engender such sensory experiences through the deconstruction of an abstract timbre, which, once spatialised, can be used to engulf a spectator, as if they were hearing from within the original sound source. This perspective on immersion can also be ascribed to a number of conceptual artworks, such as Charles Nichol's *Liberosis*, which places the listener inside the body of a violin, as well as Collective Act & Jon Hopkin's *Dreamachine*, which uses spatialised audio and lights to conjure introspective experiences.

The techniques towards spatial audio described here were made possible due to the concept of *object based audio* [2]. In comparison to channel based audio, which conceives of spatial audio in terms multiple channels, often pairing each channel with its own speaker, object based audio conceptualises spatialisation in response to the sonic materials or *sound objects* themselves. As a result, it is possible to design spatial audio techniques purely from a sonic perspective, without the need to prepare such a technique for a specific speaker configuration or technological approach towards sound spatialisation [2].

When designing immersive audio experiences, it is also important to bare in mind some of the psychoacoustic phenomena that shape how each spatialisation technique will be perceived. In doing so, it is useful to conceive of one's listening capability as having two distinct modalities - *holistic listening* and *analytic listening* [3, pp.25-27]. Holistic listening describes the perception of a complex sound as a single sonic entity, whilst analytic listening describes the ability to hear out the individual sonic components of a sound, be that a single instrumental colour amongst a complex orchestration, or a partial frequency present in a complex waveform. In general, moving between these two modalities requires the active attention of the listener themselves, however, within the context of both sound design and immersive audio practices, there are a number of means through which an artist can persuasively invoke a listener to prioritise one of these two modalities. *Tonal fusion*, which describes the holistic perception of a sound, can be influenced by numerous sonic qualities such as "attack, envelope, vibrato, harmonicity" [3, p.27], as well as similar approaches to sound design such as the Gestalt principles of auditory grouping [4]. Conversely, the process of deconstructing a timbre and spatialising its components naturally encourages a listener to perceive the original sound analytically, through each of these sonic components being presented to the listener from distinct spatial locations.

3 Spatialisation Techniques

For this work, we designed three distinct spatialisation techniques for engendering immersive audio experiences. The first technique builds upon the process of additive synthesis, whilst the latter two techniques build upon the DFT and JTFS transformations. Although each of the techniques presented here produces a collection of sonic objects which can then be individually spatialised, these techniques also carry with them their own unique aesthetic and technological details.

3.1 Additive Synthesis

Additive synthesis is typically understood as the creation of a complex waveform from numerous sinusoidal components. The history of this technique can be traced throughout the field of sound synthesis [5], and has been used to create sonic palettes ranging from simple artistic gestures to realistic sonifications of physical objects. In its simplest form, additive synthesis can be used to create an arbitrary waveform $W(t)$, such that:

$$W(t) = \sum_{n=1}^{N} \sin(2\pi f_n t + \phi)\alpha_n \tag{1}$$

where f_n is the frequency in Hertz of an individual sinusoidal component, α_n is the amplitude of that sinusoid, ϕ is the sinusoid's phase, and N is the total amount of sinusoids used to construct the complex waveform. Here, as in previous works [6, 7], we treat each sinusoidal component as its own sonic object, which can then be independently positioned within a space.

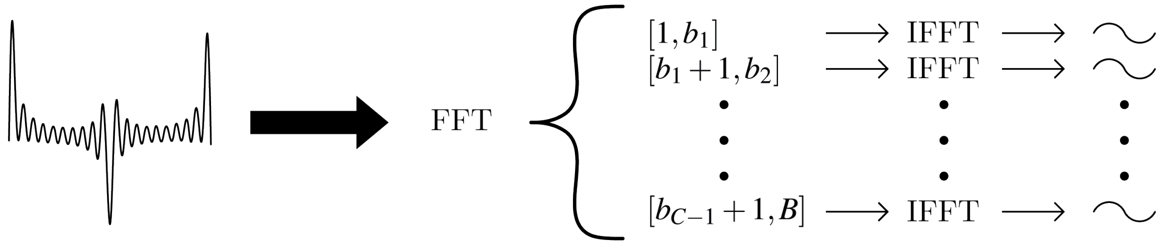
FFT
$[1, b_1] \longrightarrow$ IFFT $\longrightarrow$
$[b_1+1, b_2] \longrightarrow$ IFFT $\longrightarrow$
$[b_{C-1}+1, B] \longrightarrow$ IFFT $\longrightarrow$

Fig. 1: Diagrammatic representation of our DFT approach to spatialisation, portraying how a single sound is separated into multiple sonic objects. Here, B is used to represent the total number of FFT bins, b_n is the maximum bin index in the n^{th} subset, and C is the total number of subsets.

3.2 Discrete Fourier Transform

By building upon the fast Fourier transform (FFT) and the inverse fast Fourier transform (IFFT) [8], we designed a system which separates an incoming audio stream into multiple sonic objects, each of which can then be independently spatialised. Essentially, we begin by performing an FFT, and then separate its resulting FFT bins into subsets, preserving both their magnitude and phase, and resynthesise each subset using a dedicated IFFT. This process, for each IFFT, works the same as zeroing out all of the FFT bins that are not contained within that particular subset. A graphic representation of this design can be seen in Fig. 1. For this work, when performing our FFT, we used a sample rate of 48kHz, a window size of 2048 audio samples and a hop size of 512 audio samples to produce 1024 bins.

During our initial tests for this design, we noticed that, when distributing our resulting FFT bins linearly throughout a space, our spatialisation scheme suffered from a lack sonic variety, with many of the low to mid range frequencies being grouped together during spatialisation. This is due to the DFT acting linearly across the audible frequency range, whereas a listener conversely perceives sound logarithmically [9]. To account for this discrepancy, we created a flexible, logarithmic distribution of our FFT bins, such that, for a given audio source, we could modify its spectral and spatial distribution in response to the original audio source's sonic content. For a given FFT bin, we calculate its corresponding sonic object according to:

$$c = \left\lceil \left(\frac{b_i}{B}\right)^x C + 1 \right\rceil \qquad (2)$$

where B is the total number of FFT bins, b_i is the index of a single FFT bin according to the closed interval $[1, B] \in \mathbb{Z}$, and C is the total number of sonic objects such that $C \leq B$, with x being used to parametrise the distribution according to the open interval $(0, 1] \in \mathbb{R}$. For this work, we used a value of $C = 32$. By employing this method of distribution, we were able to create a more detailed spatialisation of the low to mid frequency range, whilst at the same time limiting the number of sonic objects which corresponded to FFT bins in the higher frequency range.

3.3 Joint Time-Frequency Scattering Transform

The JTFS transformation is used to analyse joint spectrotemporal modulations present within a sound, i.e amplitude modulations and frequency modulations [10]. As an instance of deep scattering networks [11], JTFS analyses an audio signal using multiple layers consisting of a wavelet filter convolution, multirate subsampling, pointwise modulus nonlinearity and a low-pass filter. Using this transformation, it is possible to partially reconstruct a complex audio signal, by utilising the analysed spectrotemporal modulations to iteratively reform the original signal from noise. Through this reconstructive process, the original gestural characteristics of a sound gradually emerge after each iteration.

One of the earliest creative employments of JTFS can be heard in Florian Hecker's work *Modulator (Scattering Transform)*[1], which showcased JTFS as a practical method towards the deconstruction of timbre [12]. For this piece, Hecker emphasised the iterative nature of the JTFS resynthesis process, utilising a multichannel speaker array to sonify its various iterative steps.

[1] Documentation of Florian Hecker's *Modulator (Scattering Transform)*:
`https://www.dreamideamachine.com/?p=21042`

Our work employs a similar usage of the JTFS transform, however, instead of spatialising different iterations of the JTFS resynthesis process, we run the reconstruction algorithm using subsets of the JTFS transform's coefficients to create our individual sonic objects. Through this approach, we were able to spatialise individual components of spectrotemporal modulation present within the original sound source, highlighting the gestural characteristics of said sound. As part of our systematic design, we also maintained the ability to performatively interpolate between each iteration of the resynthesis process, so as to bring these gestural characteristics in and out of perceptual focus.

Although a full explanation of the JTFS transform is beyond the scope of this work, it is important to still give an overview of the transformation and resynthesis process, so as to convey our precise method of spatialisation. However, we also refer the reader to the works of Andén et al. [10], Muradeli et al. [13] and Vahidi et al. [14] for more in depth mathematical details. Given an audio source x, we define the *first-order joint-time scattering coefficients* $\mathbf{S}_1 x(t,\lambda)$, indexed by time and log-frequency. This corresponds to a constant-Q transform [15] that is locally averaged in time, using Q bins per octaves and J octaves to create a total of JQ bins. We then define the *second-order joint time-frequency scattering coefficients*, $\mathbf{S}_2 x(t,\lambda,\alpha,\beta)$, indexed by time, log-frequency, temporal modulation rate and frequential modulation scale. This corresponds to a constant-Q wavelet filterbank [14] that is locally averaged across frequency. Concatenating the first-order and second-order JTFS coefficients forms the full JTFS transform:

$$\mathbf{S}x(t,p) = [\mathbf{S}_1 x,\ \mathbf{S}_2 x] \tag{3}$$

where p is here used to represent the scattering path, such that for the first-order coefficients $p = (\lambda)$, and for the second-order coefficients $p = (\lambda,\alpha,\beta)$.

For an arbitrary audio source, we compute its JTFS transform $\mathbf{S}x$ using a differentiable approach first introduced by Muradeli et al. [13] and Vahidi et al. [14]. To then reconstruct said audio source, we initialize the process with Gaussian noise $y_0(t)$ and perform gradient descent under the following normalized L1-error:

$$E(y_n) = \frac{||\mathbf{S}x - \mathbf{S}y_n||}{||\mathbf{S}x||} \tag{4}$$

where $\mathbf{S}y_n$ is the JTFS transform of the approximated waveform at iteration n. We use backpropagation to iteratively update y using a learning rate μ over 150 iterations, such that:

$$y_{n+1}(t) = y_n(t) + \mu \nabla E(y_n) \tag{5}$$

Since JTFS locally averages across time and frequency, sonic information from the original audio source is lost. However, this method of resynthesis using scattering coefficients serves to emphasise the spectrotemporal modulations captured by the JTFS transform.

When performing this resynthesis process, we extract our individual sonic objects using an octave subset of the first-order coefficients, j_n, relative to a given scattering path p. For each first-order octave j_i, where i corresponds to the index of each octave, encompassing the closed interval $[1,J] \in \mathbb{Z}$, we set $\mathbf{S}_2 x(p) = 0$ where $j_n \neq j_i$. In other words, we zero out all second-order coefficients whose first-order parent is not within the octave j_n. This scale-rate audio effect equates to the removal of modulation information in the constant-Q spectrogram located outside of octave j_n. After repeatedly performing this resynthesis process for all possible values of n, which for this work was limited to $J = 10$, we obtain a set of sonic objects, each of which emphasise particular subsets of the spectrotemporal modulations described by the original JTFS transform.

4 Experimental Design

To explore our spatialisation techniques, we used Cycling 74's Max to create a patch that enabled us to employ additive synthesis, sample playback, DFT transformation and various spatialisation techniques all in real-time. This choice of technological design allowed us to flexibly deploy these techniques for both performative exhibition and studio production. Due to the time complexity of performing the JTFS transformation, we performed the JTFS resynthesis algorithm offline using the Python library *Kymatio* [16]. However, for exploring the results of this methodology, we developed our Max patch to allow for real-time interaction with both the resynthesised JTFS transformations and their respective iterations.[2]

When spatialising our sonic material, we experimented both with a static dispersal of sounds, and by rotating our spatialisation about its polar and azimuthal axes.

[2] The Python code and Max patch used to develop this work has been made publicly available online: `https://github.com/lewiswolf/hearing-from-within-a-sound`

We kept the relative position of our sonic objects constant, positioning each sonic object n according to the spherical coordinate system:

$$\begin{aligned} r &= 1 \\ \theta &= \frac{2\pi n}{15} \\ \varphi &= \frac{\pi n}{15} + \frac{3\pi}{4} \end{aligned} \tag{6}$$

We chose this system primarily for its symmetrical shape when spatialising 32 sonic objects, however this symmetry was not present when working with the limited number of sonic objects produced using the JTFS spatialisation technique. The complete spatialisation pattern for 32 sonic objects can be seen in Fig. 2.

4.1 Synthesis

To explore the additive synthesis portion of this work, we implemented two synthesis models - a physically informed percussion model, derived from a rectangular membrane instrument,[3] and a custom implementation of the ripple spectra first described by Kowalski et al. [17]. We chose to explore these two models for their inharmonicity and gestural variation, with the rectangular percussion model having also been employed during our previous spatialisation works [7].

We derived our synthesis model of a rectangular membrane instrument from the two-dimensional wave equation, and formulated a simplified linear approximation according to:

$$W(t) = \sum_{m=1}^{M} \sum_{n=1}^{N} \sin(2\pi f_{mn} t) \alpha_{mn} \tag{7}$$

where m and n are the modal indices, and M and N are the maximum modal indices [18, p.72], here chosen to be 8 and 4 respectively. For a given rectangular domain with aspect ratio ε, the modal frequencies f_{mn} can be calculated according to:

$$f_{mn} = f_0 \sqrt{\frac{m^2}{\varepsilon} + \varepsilon n^2} \tag{8}$$

where f_0 is an arbitrary fundamental frequency in Hertz. Under a Dirichlet boundary condition, the modal amplitudes α_{mn} can be calculated, relative to a cartesian strike location, [19, p.309] according to:

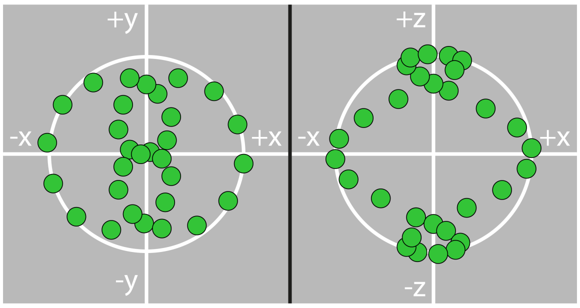

Fig. 2: Spatialisation pattern used throughout the experimentation process.

$$\alpha_{mn}(x,y) = \sin\left(\frac{m\pi x}{\sqrt{\varepsilon}}\right) \sin\left(n\pi y\sqrt{\varepsilon}\right) \tag{9}$$

For the ripple spectra, we devised our system based on the original implementation by Kowalski et al. [17], according to:

$$W(t) = \sum_{n=0}^{N-1} \sin(2\pi f_n t) \alpha_n \tag{10}$$

where N is total number of frequency components, here set to be 32. For this synthesis model, each frequency component is logarithmically spaced between an arbitrary harmonic interval, according to:

$$f_n = f_{\min} \cdot \left(\frac{f_{\max}}{f_{\min}}\right)^{n/(N-1)} \tag{11}$$

where $f_{\min}$ and $f_{\max}$ represent the minimum and maximum frequencies in Hertz of said interval. The amplitude of each frequency component is modulated, according to:

$$\alpha_n = 1 + \sin\left(2\pi\omega t + \frac{\log_2(f_n/f_{\min})}{\Omega}\right) \tag{12}$$

where ω and Ω parametrise the frequency and phase characteristics of the amplitude modulation.

4.2 Sampling & Mono Audio

For testing the DFT and JTFS spatialisation techniques, we rendered each of our additive synthesis models in

[3] The physically informed percussion instrument was implemented in Max using the external library *kac_maxmsp*: `https://github.com/lewiswolf/kac_maxmsp`

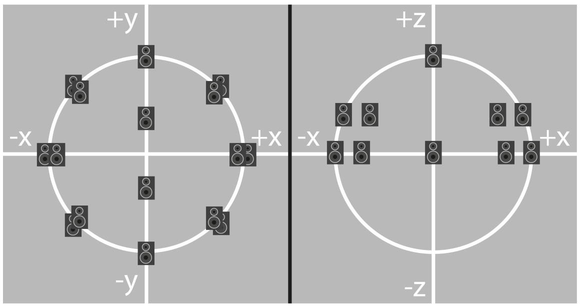

Fig. 3: Idealised representation of the speaker configuration at Iklectik Art Lab.

mono. Although these synthesis models present a useful baseline for testing the DFT and JTFS spatialisation techniques, we also chose to create an additional number of short musical segments involving sampled percussion instruments. Percussive music, in general, is well suited to testing these transformation based techniques, given that percussive music largely contains a range of inharmonic and transient spectral contents, coupled with a rich set of gestural and spectrotemporal variations. Most of this percussive material was created using the *kac_drumset* sample library [20, 21], with additional material being composed out of percussive samples taken from recordings of Iannis Xenakis' *Rebonds A* and *Rebonds B*.

5 Performative Evaluation

To evaluate these methodologies in a performative environment, we curated a private demo session at Iklectik Art Lab in London. Iklectik houses a 16 channel speaker system, originally installed by Amoenus[4], arranged in three tiers to form a semi-sphere. An idealised visualisation of this speaker configuration is shown in Fig. 3, however in both practice and implementation, the speakers at Iklectik Art Lab deviate from this ideal positioning, due to the architecture of the building itself. When performing these evaluations, we positioned all of our sonic objects in three-dimensional space using vector base amplitude panning [22], as implemented in the IRCAM Spat∼ library for Max [23].[5]

When evaluating our spatialisation techniques, we chose to respond to a set of three distinct questions. Each question served to assess the perceptual and aesthetic effectiveness of these techniques, and allowed us to structure our analytical remarks systematically.

5.1 Does each approach towards spectral decomposition perceptually distort the original sound?

When experiencing our additive synthesis spatialisation technique, the synthesised timbre remained clearly audible to us as a somewhat tangible, yet malleable complex sound. Once spatialised, the constructive design of each additive synthesis model became markedly more evident. When perceived in mono, each individual sinusoidal component typically fuses to form a perceptually unified entity. Through spatialisation, however, it became possible to aurally explore the content of these complex sounds, through physical movement or through focused listening in a particular direction. This method of spatialisation made evident our perceptual capacities towards analytic and holistic listening, emphasising our ability to hear out the partial frequencies of a sound, but without fully sacrificing our holistic impression of the original sonic entity.

The DFT approach is by far the least destructive of the three spatialisation techniques, leaving almost entirely no trace of its deconstructive approach to spatialisation. In accordance with the typical lack of sonic impact the FFT and IFFT processes have on a sound, the original audio remained perceptually intact once spatialised, with the additional feature of particular spectral characteristics becoming more clearly pronounced through their localisation in space. When listening with clear intent, it is possible to hear subtle artefacts of the deconstructed spectral content, however in general this spatialisation technique enables a truly immersive impression of the original sonic entity.

Conversely, the JTFS spatialisation technique is incredibly destructive towards its original sound source. Although this technique largely preserves the spectral content of a sound, the JTFS resynthesis process typically produces a disparate impression of its original sound source, due to the previously described loss of information that this resynthesis process incurs. Similarly, our approach to resynthesising subsets of the analysed spectrotemporal modulations do not combine to form the same sonic product produced by resynthesising the full JTFS transformation without deconstruction. This technique thus produces an entirely original sonic

[4] Amoenus: `https://amoenus.co.uk/`

[5] IRCAM Spat∼: `https://forum.ircam.fr/projects/detail/spat`

impression of a timbre, providing unique aesthetic nuances that rely upon this spatialised implementation to be at all perceptible.

5.2 How does the harmonicity/percussivity of a sound affect each spatialisation technique?

When using additive synthesis, we noticed that the previously outlined sonic qualities which encourage tonal fusion do indeed have a demonstrative effect on a sound's holistic perception. When a sound's spectral content is inharmonic, its perception as a holistic sonic entity becomes markedly more delicate, as each partial frequency has the increased potential to aurally distinguish itself. Similarly, amplitude modulations, such as percussive envelopes, have little effect on the holistic perception of a sound when used in a unified manner, such that each partial frequency follows the same dynamic contour. For similar reasons, the ripple spectra was far more difficult to perceive holistically, due to its incorporation of phase modulation.

For the DFT approach, the harmonicity or percussivity of a sound did not affect the perceptual outcome of this spatialisation technique. This approach works equally well for both harmonic and inharmonic sounds, as well as percussive and long form sonic envelopes. In practice, the successes of the DFT approach are only marginally reliant upon the spectral content of the sound source. For example, if one were to choose a sound source that contained a limited spectrum, the overall spatialisation would also be limited, as it would not be possible to spatialise content if it was not there to begin with.

As the JTFS transformation is designed to analyse modulation characteristics, the harmonicity and percussivity of a sound does also play a crucial role when employing this technique. If one were to apply this technique to a purely harmonic sound that contained no amplitude or frequency modulation, the outcome of this transformation would be largely uninteresting. If one were to instead choose an inharmonic or transient sound, which would, by design, contain minute interferences between its respective partial frequencies, this would result in subtle amplitude modulations that the JTFS transformation could then detect and represent. And similarly, if one were to choose percussive sounds over textural drones, the JTFS transformation would detect the various modulations present within their transient spectral contents, as well as the inter-onset amplitude modulations caused by the individual sonic envelopes of each percussive strike.

5.3 How does one's movement within the space affect their perception of the original sound?

In general, our additive synthesis technique was noticeably sensitive to our position and movement within a spatialised environment. With each different position, individual partial frequencies were brought forth into our perceptual awareness. This sensitivity heightened our capacity to identify and focus on each partial frequency, in turn weakening our capacity to perceive the original sound as a holistic sonic entity. When one stands within the center of the space, this aural effect is subtle and in accordance purely with the position of the head. However, when one moves further out towards the edge of the space, partial frequencies quickly begin to showcase their individuality, becoming easily distinguishable as one approaches a particular speaker.

Similar to our additive synthesis technique, the DFT technique maintained an analogous relationship to our position within the space. Overall, however, these effects were markedly more subtle, as each spatialised object has a more sonically entangled relationship with the rest of the spatialised objects present within the space. As a result, movement within the space does not diminish one's ability to holistically perceive the original sonic entity, but does still promote one's capability to hear out more generalisable spectral characteristics of the original sound source.

Unlike the prior two techniques, the JTFS transformation does not engender the same spectral relationship with position and movement, as the spectral content of each spatialised sonic object is largely homogenous across the space. Instead, it is only the gestural components of a sound which present themselves in greater of lesser detail, depending on one's position and orientation. Of the three techniques, the JTFS approach also requires the most acute form of active listening to perceive the many gestural nuances emphasised by each sonic object, which, as previously noted, are largely dependent upon the diversity of gestural characteristics present in the original sound. This attentive subtlety is difficult to describe, for although the sonic landscape, in terms of dynamicity and spectral content, is just as immersive as the other two techniques, the JTFS transformation affords an extra layer of timbral intricacy.

6 Studio Evaluation

To conclude our experimentations, we also rendered these spatialisation techniques in binaural, using a set of head-related transfer functions derived from a KEMAR dummy head [24], and again employing the IRCAM Spat$\sim$ library for Max [23]. Using headphones, we explored each technique using a freely modifiable reference location to simulate our virtual movement, whilst also incorporating the same spatialisation pattern described by equation 6.

In general, our evaluations of each spatialisation technique, as described in the previous section, was retained throughout each iteration of binaural spatialisation. However, the overall immersive effectiveness of our binaural exposition was at times limited due this approach's inherent perceptual nuances. These nuances include difficulties in perceptually differentiating between sonic objects which are either in front of the listener or behind them [25], as well as *the room divergence effect* [26]. The room divergence effect describes an inability to accurately adapt our perception when experiencing virtual room acoustics, contrary to one's typical capacity to adapt to rapidly changing acoustic room conditions in real-world scenarios. Within the context of this work, the biggest difference between the binaural and real-world deployments of our spatialisation techniques was the loss of an accurate sense of distance and circularity. As a result of this, and also as the result of an absence of naturalistic, acoustic reverberation, it became easier to independently perceive each spatialised sonic object when using binaural rendering. Although these perceptual differences are noticeable, the use of binaural rendering, in combination with our spatialisation techniques, presents an additional opportunity to emphasise the perceptual grey area between holistic and analytic listening. Further audio examples of the effects of binaural rendering alongside our additive synthesis spatialisation technique have also been showcased in some of our previous works [6, 7][6].

7 Conclusion

We have here presented three techniques developed for the deconstruction and spatialisation of timbre in an immersive audio environment. Through there distinct approaches to spatialisation, each of these techniques engenders a similarly unique aesthetic, perceptual and immersive impression of a sound. Of these three techniques, our additive synthesis method acutely affords the possibility to explore one's perceptual capacity towards holistic and analytic listening. And whilst our spatialisation technique based on the discrete Fourier transform is widely applicable to the accurate spatialisation of an arbitrary sound source, our technique developed upon the joint time frequency scattering transform can be used to curate entirely original sonic landscapes. As a series of compositional tools, each of these techniques affords an artist or sound engineer the possibility to have their listeners hear the spectral and gestural characteristics of a sound from within.

8 Acknowledgements

The authors would like to thank Christian Duka[7] for arranging the experimentation session at Iklectik Art Lab. They would also like to thank Francis Devine[8] for his help in preparing figures 1, 2 & 3.

This work is supported by the Centre for Doctoral Training in Artificial Intelligence and Music at Queen Mary University of London, funded by UK Research and Innovation (UKRI) under EPSRC grant EP/S022694/1.

References

[1] Schrimshaw, W., *Immanence and Immersion: On the Acoustic Condition in Contemporary Art*, Bloomsbury Academic, New York, NY, 2017.

[2] Geier, M., Ahrens, J., and Spors, S., "Object-Based Audio Reproduction and the Audio Scene Description Format," *Organised Sound*, 15(03), pp. 219–227, 2010.

[3] Sethares, W. A., *Tuning, Timbre, Spectrum, Scale*, Springer, Heidelberg, Germany, 2005.

[4] Bregman, A. S. and Pinker, S., "Auditory Streaming and the Building of Timbre," *Canadian Journal of Psychology*, 32(1), pp. 19–31, 1978.

[5] Manning, P., *Electronic and Computer Music*, Oxford University Press, New York, NY, second edition, 2004.

[6] *chroma* by Lewis Wolstanholme:
https://lewiswolstanholme.bandcamp.com/chroma
josef by Julia Set:
https://juliaset.bandcamp.com/josef

[7] Christian Duka: https://christianduka.com

[8] Francis Devine: http://francisdevine.co.uk

[6] Wolstanholme, L., “chroma,” Music Recording & Score, 2019.

[7] Wolstanholme, L. and Devine, F., “josef: Spatiality as a Material Property of Audiovisual Art,” *International Journal of Creative Media Research (IJCMR)*, Forthcoming, 2023.

[8] van Loan, C., *Computational Frameworks for the Fast Fourier Transform*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[9] Fechner, G. T., *Elements of Psychophysics*, Holt, Rinehart and Winston, New York, NY, 1966.

[10] Andén, J., Lostanlen, V., and Mallat, S., “Joint Time–Frequency Scattering,” *IEEE Transactions on Signal Processing*, 67(14), pp. 3704–3718, 2019.

[11] Andén, J. and Mallat, S., “Deep Scattering Spectrum,” *IEEE Transactions on Signal Processing*, 62(16), pp. 4114–4128, 2014.

[12] Lostanlen, V. and Hecker, F., “The Shape of RemiXXXes to Come: Audio Texture Synthesis with Time-Frequency Scattering,” in *22nd International Conference on Digital Audio Effects (DAFx)*, Birmingham, UK, 2019.

[13] Muradeli, J., Vahidi, C., Wang, C., Han, H., Lostanlen, V., Lagrange, M., and Fazekas, G., “Differentiable Time-Frequency Scattering on GPU,” in *25th International Conference on Digital Audio Effects (DAFx)*, Copenhagen, Denmark, 2022.

[14] Vahidi, C., Han, H., Wang, C., Lagrange, M., Fazekas, G., and Lostanlen, V., “Mesostructures: Beyond Spectrogram Loss in Differentiable Time-Frequency Analysis,” *arXiv preprint arXiv:2301.10183*, 2023.

[15] Brown, J. C., “Calculation of a Constant Q Spectral Transform,” *The Journal of the Acoustical Society of America*, 89(1), pp. 425–434, 1991.

[16] Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., Zarka, J., Mallat, S., Andén, J., Belilovsky, E., Bruna, J., Lostanlen, V., Chaudhary, M., Hirn, M. J., Oyallon, E., Zhang, S., Cella, C., and Eickenberg, M., “Kymatio: Scattering Transforms in Python,” *Journal of Machine Learning Research*, 21(60), pp. 1–6, 2020.

[17] Kowalski, N., Depireux, D. A., and Shamma, S. A., “Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex. I. Characteristics of Single-Unit Responses to Moving Ripple Spectra,” *Journal of Neurophysiology*, 76(5), pp. 3503–3523, 1996.

[18] Fletcher, N. H. and Rossing, T. D., *The Physics of Musical Instruments*, Springer, New York, NY, second edition, 1998.

[19] Bilbao, S., *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*, John Wiley & Sons, Chichester, UK, 2009.

[20] Wolstanholme, L. and Devine, F., “terracotta,” in *22nd International Conference on New Interfaces for Musical Expression (NIME)*, Auckland, New Zealand, 2022.

[21] Wolstanholme, L., “kac_drumset: A Dataset Generator for Arbitrarily Shaped Drums,” Zenodo, 2022, Version: 1.1.

[22] Pulkki, V., “Virtual Sound Source Positioning Using Vector Base Amplitude Panning,” *Journal of the Audio Engineering Society*, 45(6), pp. 456–466, 1997.

[23] Carpentier, T., “A New Implementation of Spat in Max,” in *15th Sound and Music Computing Conference (SMC)*, Limassol, Cyprus, 2018.

[24] Romblom, D. and Cook, B., “Near-Field Compensation for HRTF Processing,” in *125th Convention of the Audio Engineering Society*, San Francisco, CA, 2008.

[25] Zieliński, S. K., Antoniuk, P., Lee, H., and Johnson, D., “Automatic Discrimination between Front and Back Ensemble Locations in HRTF-convolved Binaural Recordings of Music,” *EURASIP Journal on Audio, Speech, and Music Processing*, 1, 2022.

[26] Klein, F., Werner, S., Götz, G., and Brandenburg, K., “Auditory Adaptation in Real and Virtual Rooms,” *Proceedings of the International Symposium on Auditory and Audiological Research*, 7, pp. 341–348, 2019.