STASE: A SPATIALIZED TEXT-TO-AUDIO SYNTHESIS ENGINE FOR MUSIC GENERATION

Tutti Chi[1] **Letian Gao**[2] **Yixiao Zhang**[3]
[1] **University of Chinese Academy of Sciences, China**
[2] **Tsinghua University, China**
[3] **Centre for Digital Music (C4DM), Queen Mary University of London, UK**
chengtopia@outlook.com, glt23@mails.tsinghua.edu.cn, yixiao.zhang@qmul.ac.uk

ABSTRACT

While many text-to-audio systems produce monophonic or fixed-stereo outputs, generating audio with user-defined spatial properties remains a challenge. Existing deep learning-based spatialization methods often rely on latent-space manipulations, which can limit direct control over psychoacoustic parameters critical to spatial perception. To address this, we introduce STASE, a system that leverages a Large Language Model (LLM) as an agent to interpret spatial cues from text. A key feature of STASE is the decoupling of semantic interpretation from a separate, deterministic signal-processing-based spatial rendering engine, which facilitates interpretable and user-controllable spatial reasoning. The LLM processes prompts through two main pathways: (i) Description Prompts, for direct mapping of explicit spatial information (e.g., "place the lead guitar at 45° azimuth, 10 m distance"), and (ii) Abstract Prompts, where a Retrieval-Augmented Generation (RAG) module retrieves relevant spatial templates to inform the rendering. This paper details the STASE workflow, discusses implementation considerations, and highlights current challenges in evaluating generative spatial audio.

1. INTRODUCTION

The landscape of AI-powered music generation has advanced rapidly, moving from general-purpose frameworks towards precision-oriented paradigms [1–3]. In parallel, immersive audio is gaining traction across industries, fueling demand for spatially enhanced auditory experiences [4, 5]. However, current spatial audio synthesis workflows—such as Dolby Atmos and Apple renderers—have been reported by music producers to provide limited operational controllability [6].

Most existing deep learning-based spatialization methods operate solely on audio inputs, without leveraging direct textual descriptions [7–9]. While latent-space manipulation techniques show potential [10–12], their black-box nature may result in information loss and limit precise control over psychoacoustic parameters critical to spatial perception [13–15]. Unlike features such as melody or emotion, spatial information can be explicitly modeled and manipulated via signal processing.

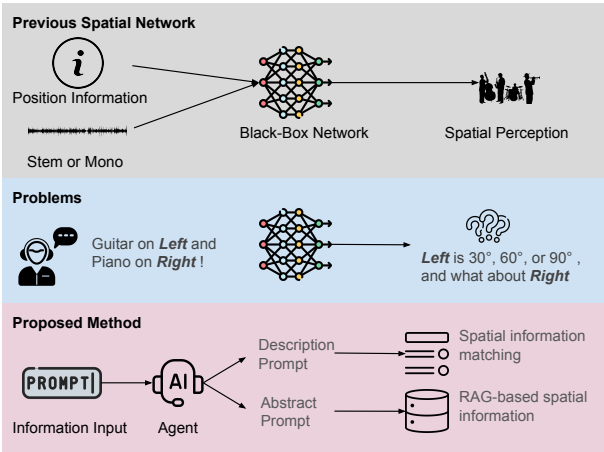

Figure 1. High-level comparison between latent-space spatialization pipelines and the proposed agentic STASE pipeline. STASE decouples prompt interpretation (LLM + RAG) from a deterministic signal-processing-based renderer to improve controllability and interpretability.

Leveraging established principles of binaural hearing—including interaural time difference (ITD), interaural level difference (ILD) [16, 17], and dynamic cues [18]—we introduce **STASE**, a hybrid neuro-symbolic framework for generating spatially dynamic music from natural language prompts. As illustrated in Fig. 1, prior spatial networks directly map position information and audio signals (stem or mono) to spatial perception via a black-box model, which can have difficulty interpreting spatial cues from descriptive language and provides limited controllability. In contrast, STASE integrates an LLM as an agent to process the input prompt and route it into one of two pathways: (*i*) *Description Prompts*, enabling direct spatial information matching when precise positional details are given; and (*ii*) *Abstract Prompts*, where a Retrieval-Augmented Generation (RAG) module retrieves relevant spatial knowledge before rendering. This decoupling of semantic interpre-

© T. Chi, L. Gao, and Y. Zhang. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** T. Chi, L. Gao, and Y. Zhang, "STASE: A Spatialized Text-to-Audio Synthesis Engine for Music Generation", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

tation from deterministic signal-processing-based spatial rendering supports interpretable and user-controllable spatial reasoning.

Our modular architecture combines the LLM-based prompt interpreter, a music generation module for content creation, optional source separation or monaural instrument inputs, and a dedicated spatialization engine driven by LLM-derived parameters. In our current implementation, stems are generated via a music synthesis model and source separation is not required. This modularity allows components to be independently replaced or fine-tuned, supporting both novice-friendly presets (e.g., fixed studio arrangements) and expert-level customized synthesis.

2. RELATED WORK

2.1 Spatial Audio Perception and Production

Research on spatial auditory perception, starting with Rayleigh's foundational work [16], has highlighted the roles of pinna filtering, Head-Related Transfer Functions (HRTFs), and ITDs in sound localization. Later studies demonstrated how dynamic head movements [18] and individualized HRTFs [19] improve both localization precision and subjective immersion. Despite these advancements, a study involving music producers noted that current spatial audio tools often lack flexibility and face challenges with playback consistency across different devices [6].

2.2 Intelligent Music Generation Systems

The field of music AI has seen rapid progress, particularly with the emergence of LLMs, leading to systems such as MusicGen [2] and MusicLM [3]. These models have been further refined for specific tasks such as instruction following [20] or temporal control [21]. While many AI models now generate stereo audio (e.g., MusicGen [2], Jen-1 [22], Stable Audio [23]), they generally lack the precise spatial control needed for detailed spatial rendering and immersive experiences.

2.3 Spatial Audio Generation Techniques

Stereo panning is a fundamental music production technique, deeply integrated into Digital Audio Workstations (DAWs) and primarily based on amplitude panning using principles like the "sine-cosine" pan law [24]. Beyond panning, the broader field of spatial sound synthesis and recording is an active research area. Recent neural network methods have emerged, such as BinauralGrad [7] for binaural signal conversion and methods for learning spatial cues from video [8]. Other approaches, like AudioLDM [9], manipulate audio effects via latent spaces, while TAS [10] spatializes monaural audio through similar latent manipulations. More recently, ImmerseDiffusion [11] generates Ambisonics from text, and Sun et al. [12] proposed a language-driven stereophonic audio generation framework. Our work builds upon these, specifically focusing on achieving precise text-to-spatial-audio control for music.

3. METHODOLOGY

STASE is an LLM-driven spatial audio synthesis framework designed to generate musical compositions with user-specified spatial attributes from natural language prompts. The system architecture transforms textual inputs into a fully rendered, multi-track spatial audio mix through a modular pipeline.

As depicted in Figure 2, STASE operates in four sequential stages. First, it ingests a free-form natural language prompt. A RAG module searches a preset library for semantically relevant spatial configurations to inform the generation process. Second, the *Conductor Agent*—a core reasoning module—fuses the raw prompt with the retrieved presets. Driven by an LLM, this agent outputs a structured plan: standardized music descriptions, exact spatial parameters (azimuth, distance, etc.), and concise mixing directives. Third, a music generation model synthesizes individual audio tracks (stems) based on the Conductor Agent's descriptions. Finally, the spatial renderer applies the agent-derived parameters—using one of panning, ITD/ILD, or HRTF for localization, plus reverberation—to the stems, yielding the finished spatial mix.

3.1 Prompt Interpretation and Adaptation

The essence of STASE lies in translating diverse natural-language prompts into precise, actionable spatial parameters via an LLM. As illustrated in Figure 2, the process starts when the user supplies a description of the desired sonic scene, optionally augmented by presets, instrument sets, and the RAG engine. The LLM, acting as a *Conductor Agent*, processes both the raw prompt and any RAG-retrieved results to produce a music description and a structured spatial perception map. If the prompt contains explicit spatial cues (e.g., "place the lead guitar at 45° azimuth, 10 m distance"), the LLM directly parses these into quantifiable parameters such as azimuth, elevation, and distance; if the prompt is more abstract (e.g., "a grand orchestral arrangement"), the LLM queries the RAG module to retrieve semantically relevant spatial templates (e.g., "symphony orchestra stage setup") and maps these to default parameter sets. The music description is then forwarded to the *Music Agent* to generate multitrack audio, ensuring that spatial placement decisions are tied to clearly defined sound sources. Finally, the spatialization engine applies either user-specified coordinates or template-driven defaults to the multitrack audio, incorporating techniques such as panning, HRTF-based binaural rendering, Room Impulse Response (RIR) convolution, and artificial reverberation to produce the final spatial audio. To accommodate varying user expertise and prompt specificity, STASE selects the spatialization approach:

- **Precise Spatialization:** When the input prompt contains explicit spatial cues (e.g., "place the lead guitar at 45° azimuth, 10 meters distance"), the LLM directly extracts and applies these values for accurate source positioning.

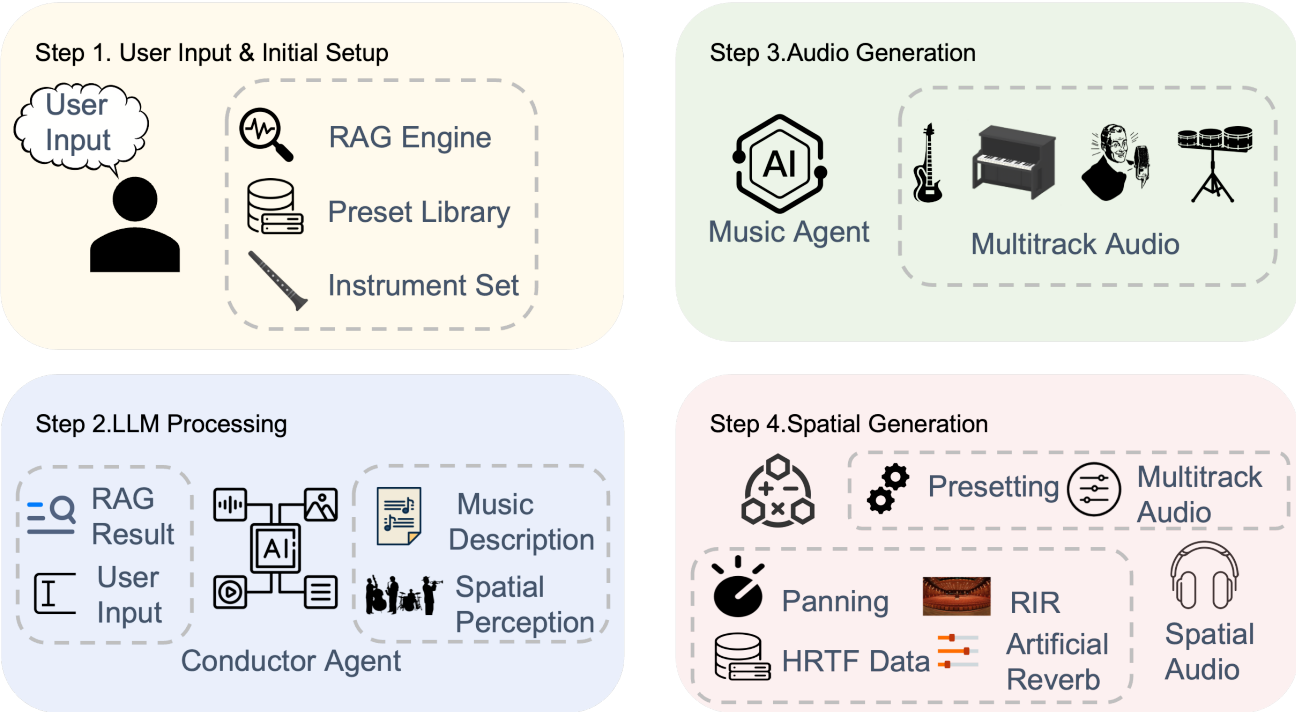

Figure 2. STASE workflow: prompts are fused with template knowledge (RAG), transformed by a Conductor Agent into a structured plan (music description, spatial map, mix notes), synthesized into stems, and rendered by a deterministic signal-processing chain (per-source localization uses one of: panning, ITD/ILD, or HRTF; plus reverb/RIR). The agent routes inputs via Description vs. Abstract pathways.

- **Templated Spatialization:** For more abstract or less precise prompts (e.g., "a grand orchestral arrangement"), the LLM utilizes a RAG approach. It retrieves pre-defined spatial templates (e.g., "symphony orchestra stage setup") semantically closest to the input, providing default spatial parameters for coherent layouts.

3.2 Deterministic Signal-Processing-Based Spatial Rendering

STASE employs a flexible, deterministic signal-processing approach for spatial rendering (as opposed to latent-space manipulation). Our spatialization module supports established signal-processing methods for sound source localization and environmental acoustic simulation:

3.2.1 Sound Source Localization

We support three mutually exclusive localization modes: (i) stereo amplitude panning for lateral placement and non-binaural playback; (ii) analytic ITD/ILD rendering (without HRTF) for controlled cue manipulation; and (iii) HRTF convolution for full three-dimensional binaural rendering. The Conductor Agent selects one mode per source based on the requested coordinates and the target output format. When HRTF is used, additional panning or explicit ITD/ILD processing is disabled to avoid double-counting cues. In non-HRTF modes, ITD is realized by fractional-delay lines aligned with head-width approximations and ILD by frequency-dependent gain shaping.

3.2.2 Reverberation

As with localization, reverberation enriches perceived space. Based on the input style, the system can use parameterized algorithmic reverberators for fine-grained control or directly convolve with Room Impulse Responses (RIRs) for specific acoustic environments or stylistic matching. RIR selection can be initialized from the template retrieved by RAG and refined by the Conductor Agent's mix notes.

4. IMPLEMENTATION DETAILS

4.1 Resources

We describe resources used in our implementation, including acoustic databases and predefined spatial configurations.

4.1.1 RIR Database

For environmental acoustic simulation, we employed single-channel RIRs extracted from established multi-channel datasets including the dEchorate database [25] and selected measurements from the OpenAIR library [26]. Our curated RIR collection includes ten distinct acoustic environments designed to match our spatial configuration templates: (1) large concert halls with natural reverberation for classical orchestras, (2) intimate studio rooms for jazz ensembles, (3) dry recording studios for controlled rock band setups, (4) small chambers for intimate music arrangements, (5) medium-sized venues for electronic performances, (6) churches with extended reverb for choir

formations, (7) recital halls for solo performances, (8) acoustically diverse spaces for world music, (9) professional recording environments, and (10) simulated outdoor spaces with minimal reflections for festival configurations. Each single-channel RIR is applied to individual audio stems during the spatial rendering process, enabling the system to match textual descriptions of acoustic environments with appropriate reverberation characteristics that complement the corresponding spatial arrangement template.

4.1.2 Spatial Configuration Templates

To systematically evaluate spatial placement accuracy, we manually defined ten distinct spatial configuration templates corresponding to common musical performance scenarios: (1) *Classical Orchestra* - traditional symphonic layout with woodwinds front, brass middle, and strings distributed; (2) *Jazz Ensemble* - intimate small group arrangement; (3) *Rock Band* - conventional stage setup with drums center-back; (4) *Chamber Music* - close-proximity classical arrangement; (5) *Electronic/DJ Setup* - electronic music performance configuration; (6) *Choir Formation* - vocal ensemble positioning; (7) *Solo Performance* - single instrument with accompaniment; (8) *World Music Ensemble* - diverse cultural instrument arrangements; (9) *Studio Recording* - controlled studio environment layout; and (10) *Outdoor Festival* - large-scale outdoor performance setup. Each template specifies precise azimuthal positions, elevation angles, and distance parameters for up to 6 simultaneous sources, providing standardized reference points for evaluation.

4.1.3 HRTF Implementation

For binaural spatialization, we employed the KEMAR HRTF database [27].

4.2 LLM and Prompting Details

We use an instruction-tuned, open-weight LLM in the 7–13B parameter range with deterministic decoding (temperature $= 0$, top-$p = 1$) and a constrained, schema-guided output format. Few-shot exemplars cover both Description and Abstract pathways. We provide model names and prompt templates with the supplementary materials.

4.3 Music Generation Module

Our prototype integrates an off-the-shelf text-to-music synthesis system to produce up to 2–6 stems depending on the prompt (e.g., drums, bass, guitar, keys, lead, pads). The module can be substituted or bypassed without changing the rest of the pipeline.

4.4 User-Provided Stems

Beyond generation, STASE accepts user-provided monaural stems. The user supplies instrument labels (or lets the Agent infer them), and the spatial renderer applies the same plan to the uploaded stems, enabling practical workflows in DAW-centric production.

4.5 Reproducibility Notes

We release the template bank, RIR list, parsing code, and prompts used for the Conductor Agent, together with random seeds and decoding settings. This enables step-by-step reproduction of routing decisions and rendering given fixed inputs.

5. RESULTS

Audio demonstrations are available on the project page: `https://chengtopia.github.io/STASE.github.io/`. We include multiple RAG-driven generations and audio samples to facilitate subjective evaluation of spatialization quality and controllability.

6. DISCUSSION

Evaluating text-driven spatial audio remains challenging due to the lack of standardized metrics. ITD and ILD are effective for single-source azimuthal accuracy (e.g., 30° vs 60°), but in multi-source mixes their cues interact and are hard to isolate; they also fail to capture full 3D or perceptual quality. In complex arrangements, overlap and reverberation further confound per-source analysis. Conventional semantic metrics (e.g., CLAP, T5/KL) measure content alignment yet are largely insensitive to fine-grained spatial instructions. We therefore recommend paired objective proxies together with controlled listening tests tailored to spatial attributes.

7. CONCLUSION

We presented STASE, an agentic framework that interprets prompts with an LLM and renders spatial mixes using a deterministic signal-processing chain. Decoupling semantics from rendering improves controllability and interpretability, and the modular design supports swapping LLMs, music models, and spatializers. The main limitation is reliance on separated monaural stems; performance can degrade on dense or reverberant mixtures. Future work includes standardized objective/subjective evaluation for spatial attributes and extending the interaction model to mixed monaural inputs and broader production workflows.

8. REFERENCES

[1] L. Wang, Z. Zhao, H. Liu, J. Pang, Y. Qin, and Q. Wu, "A review of intelligent music generation systems," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6381–6401, 2024.

[2] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 704–47 720, 2023.

[3] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm:

Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[4] T. Potter, Z. Cvetković, and E. De Sena, "On the relative importance of visual and spatial audio rendering on vr immersion," *Frontiers in Signal Processing*, vol. 2, p. 904866, 2022.

[5] F. Immohr, G. Rendle, A. Neidhardt, S. Göring, R. R. Ramachandra Rao, S. Arevalo Arboleda, B. Froehlich, and A. Raake, "Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal vr communication," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 2023, pp. 209–215.

[6] C. Dewey, A. Moore, and H. Lee, "Practitioners' perspectives on spatial audio: Insights into dolby atmos and binaural mixes in popular music," *AES: Journal of the Audio Engineering Society*, vol. 72, no. 7/8, pp. 504–516, 2024.

[7] Y. Leng, Z. Chen, J. Guo, H. Liu, J. Chen, X. Tan, D. Mandic, L. He, X. Li, T. Qin *et al.*, "Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 689–23 700, 2022.

[8] K. K. Parida, S. Srivastava, and G. Sharma, "Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3347–3356.

[9] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[10] Z. Li, B. Zhao, and Y. Yuan, "Tas: Personalized text-guided audio spatialization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9029–9037.

[11] M. Heydari, M. Souden, B. Conejo, and J. Atkins, "Immersediffusion: A generative spatial audio latent diffusion model," *arXiv preprint arXiv:2410.14945*, 2024.

[12] P. Sun, S. Cheng, X. Li, Z. Ye, H. Liu, H. Zhang, W. Xue, and Y. Guo, "Both ears wide open: Towards language-driven spatial audio generation," *arXiv preprint arXiv:2410.10676*, 2024.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[14] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4432–4441.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[16] L. Rayleigh, "On our perception of the direotion of a source of sound," *Proceedings of the Musical Association*, vol. 2, no. 1, pp. 75–84, 1875.

[17] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[18] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, vol. 27, no. 4, p. 339, 1940.

[19] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the improvement of localization accuracy with non-individualized hrtf-based sounds," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 821–830, 2012.

[20] Y. Zhang, Y. Ikemiya, W. Choi, N. Murata, M. A. Martínez-Ramírez, L. Lin, G. Xia, W.-H. Liao, Y. Mitsufuji, and S. Dixon, "Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning," *arXiv preprint arXiv:2405.18386*, 2024.

[21] Y.-H. Lan, W.-Y. Hsiao, H.-C. Cheng, and Y.-H. Yang, "Musicongen: Rhythm and chord control for transformer-based text-to-music generation," *arXiv preprint arXiv:2407.15060*, 2024.

[22] Y. Yao, P. Li, B. Chen, and A. Wang, "Jen-1 composer: A unified framework for high-fidelity multi-track music generation," *arXiv preprint arXiv:2310.19180*, 2023.

[23] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *Forty-first International Conference on Machine Learning*, 2024.

[24] V. Pulkki, *Spatial sound generation and perception by amplitude panning techniques*. Helsinki University of Technology, 2001.

[25] A. Caillon, R. M. Bittner, M. Lagrange, C. Singh, J. P. Bello, and G. Richard, "dechorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–19, 2021.

[26] D. T. Murphy and S. Shelley, "Openair: The open acoustic impulse response library," University of York, 2009, available online at www.openairlib.net.

[27] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy-head microphone," in *Audio Engineering Society Convention 95*. Audio Engineering Society, 1995.