VCNAC: A Variable-Channel Neural Audio Codec for Mono, Stereo, and Surround Sound

Florian Grötschla* *Amazon AGI, ETH Zürich* fgroetschla@ethz.ch
Arunasish Sen *Amazon AGI* arusen@amazon.com
Alessandro Lombardi *Amazon AGI* loaless@amazon.co.uk
Guillermo Cámbara *Amazon AGI* gcambara@amazon.com
Andreas Schwarz *Amazon AGI* asw@amazon.de

***Abstract*—We present VCNAC, a variable-channel neural audio codec. Our approach features a single encoder and decoder parametrization that enables native inference for different channel setups, from mono speech to cinematic 5.1 channel surround audio. Channel compatibility objectives ensure that multi-channel content maintains perceptual quality when decoded to fewer channels. The shared representation enables training of generative language models on a single set of codebooks while supporting inference-time scalability across modalities and channel configurations. Evaluation using objective spatial audio metrics and subjective listening tests demonstrates that our unified approach maintains high reconstruction quality across mono, stereo, and surround audio configurations.**

***Index Terms*—Neural audio codec, multi-channel codec, surround audio codec**

I. Introduction

Neural audio codecs enable a range of audio generation and processing applications, including speech synthesis and understanding. These applications operate across different channel configurations: mono for speech, stereo for music, and 5.1 surround for movies. Surround audio systems like 5.1 use six channels (front left/right, center, low-frequency effects, and rear left/right) to create immersive spatial audio experiences. However, codecs such as EnCodec [1] use fixed channel architectures that can only process audio with a predetermined number of channels. This limitation creates significant challenges when modeling multiple channel configurations within a single application. Current approaches require either training separate codecs for each desired channel configuration, each with its own distinct latent space, or processing all audio using the maximum number of channels the application may encounter. The latter approach leads to substantial computational inefficiencies when most data contains fewer channels than the maximum supported configuration. Existing multi-channel approaches have been limited to at most two channels and typically adapt single-channel designs by modifying only the input and output convolutional layers to accommodate the target number of channels. The remainder of the architecture processes all channels jointly and splits them into a fixed number of channels which does not directly generalize to varying channel requirements across different applications.

We propose a neural audio codec that addresses these limitations through a variable-channel architecture capable of dynamically processing different numbers of input and output channels within the same encoder-decoder framework. Our approach uses shared codebooks to enable a unified latent space that represents audio content regardless of channel configuration, spanning mono, stereo, and surround formats. The shared latent space enables language model training on unified codebooks with inference-time scalability across different channel configurations. Our evaluation demonstrates good reconstruction quality across all channel configurations while achieving significant bitrate reductions compared to existing approaches (7.9 kbit/s vs. 14-16 kbit/s). A MUSHRA [2] study confirms perceptual quality gains.

*Work done during an internship at Amazon.

II. Related Work

Neural Audio Codecs. Neural audio codecs have primarily targeted mono audio processing, with early architectures establishing foundational approaches. SoundStream [3] introduced Residual Vector Quantization (RVQ) for neural audio compression, building on VQ-VAE [4]. EnCodec [1] and DAC [5] extended this foundation using RVQ with strided/transposed convolutions and adversarial training. While traditional codecs like Opus [6] natively support multi-channel configurations, neural approaches typically adapt single-channel architectures by modifying input/output layers while processing all channels through shared representations. EnCodec [1] represents the only stereo-capable neural audio codec using this approach, while Stable Audio Open VAE [7] demonstrates similar stereo adaptations with continuous representations. Specialized approaches exist for specific scenarios: SpatialCodec [8] encodes reference and side channels separately for microphone arrays, while BANC [9] decomposes binaural speech into clean speech and room impulse responses. However, these specialized approaches have limited applicability to general-purpose multi-channel processing.

Generative Models for Stereo Audio. While neural audio codecs have primarily focused on mono processing, audio generation applications have driven various approaches to stereo synthesis. Some systems integrate stereo capabilities directly into VAE architectures [7], [10], while others work around mono codec limitations through token-level strategies such as interleaved delay patterns that generate tokens for the different channels separately [11], [12]. Several generation models support stereo output and utilize continuous latent representations [13], [14], though many provide limited technical

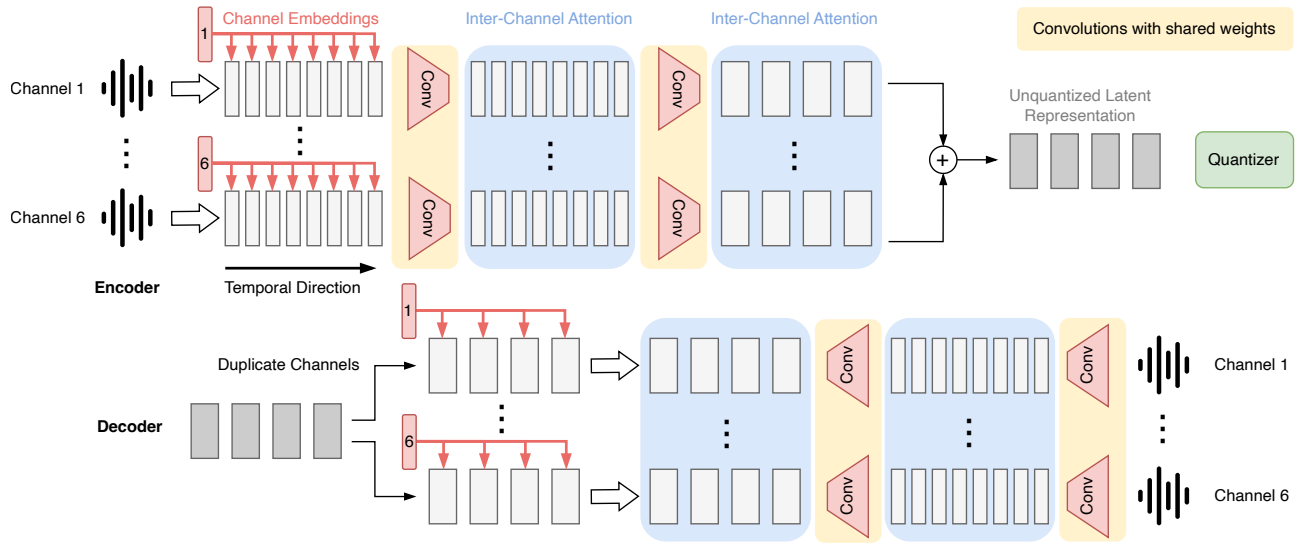

Fig. 1. Variable-channel neural audio codec architecture. Parallel channel streams with shared weights for the convolutional layers process variable input channels, fuse into unified representations for quantization, then split to target output channels. The "channel embedding" denotes the learned per-channel positional vector $\mathbf{p}_c$ added to each stream's latents to preserve channel identity. Cross-channel attention enables information exchange before fusion and after splitting. The architecture natively supports mono, stereo, and surround audio. We only show two up- and down-sampling convolutions for visualization purposes. We use five convolutions for VCNAC as tested in the experiments.

details about their multi-channel processing. Alternative approaches include pseudo-stereo generation used in Diff-A-Riff, where unified latents are diffused jointly before being split into two latent streams [15], or approaches that concatenate the two spectrograms for the input and output of the model [16]. These approaches primarily address generation via continuous representations rather than discrete tokenization.

Spatial Audio Processing. Standard reconstruction losses inadequately preserve spatial audio characteristics by treating channels independently, failing to capture critical inter-channel relationships. Spatial audio preservation benefits from specialized loss functions based on psychoacoustic principles [17]. Two primary approaches have emerged: spectral-domain methods using sum-difference STFT losses [7], [18] that separate common content from spatial differences, and binaural cue-based methods employing Interaural Level Difference (ILD) and Interaural Phase Difference (IPD) metrics that directly capture localization cues [19]. These spatial losses have been shown to correlate with human perception in binaural speech enhancement [20] and stereo-aware speech processing [21].

III. Methodology

A. Architecture

Our approach extends the DAC encoder-decoder architecture [5] to support variable-channel configurations within a unified model (Figure 1). The key innovation lies in decoupling channel-specific processing and fusing representations only for the quantization stage.

Variable-Channel Processing Pipeline. Rather than concatenating all channels into the first fixed-size convolutional layers [1], we employ a three-stage approach: (1) *parallel channel processing* where each input channel is processed through separate but weight-shared convolutional streams, (2) *fusion* into a unified bottleneck representation for quantization, and (3) *splitting* back to the target number of output channels for reconstruction. This design enables dynamic adaptation to different channel counts. Processing streams are only instantiated for existing channels and we avoid computational overhead that would result from padding unused channels. Each channel stream processes its input through strided and transposed (in the case of the decoder) convolutions with shared parameters, which work in an identical fashion to existing neural audio codecs like DAC. Channel identity is preserved through learnable positional embeddings that are added to each stream's latent embeddings, i.e., $\mathbf{z}_c = \mathbf{h}_c + \mathbf{p}_c$ where $\mathbf{h}_c$ is the latent for channel c and $\mathbf{p}_c \in \mathbb{R}^D$ is a learned channel embedding. They are necessary to identify the channel that the stream is embedding and crucial when the channel embeddings are joined for quantization, as there would otherwise be no way to distinguish them. The positional embeddings are initialized as orthogonal vectors with small magnitude ($\sigma = 0.01$), which we found to be essential for training stability.

Fusion and Splitting Strategy. Channel streams are fused by adding up embeddings with the same temporal location at a configurable depth in the encoder hierarchy. We place fusion after the final strided convolution (following [22]), which maximizes channel-separate processing while ensuring sufficient interaction for cross-channel dependencies. The fused representation undergoes standard RVQ quantization and creates a unified latent space independent of input channel configuration. During decoding, the dequantized representation is split into the required number of output streams by duplicating the embedding streams. To identify the different channels, a

TABLE I
SINGLE-CHANNEL SPEECH EVALUATION RESULTS ON LIBRITTS. BEST RESULT IN **BOLD**, SECOND-BEST UNDERLINED. BITRATES ARE IN KBIT/S.

Codec	Bitrate	PESQ ↑	SI-SDR ↑	SI-SNR ↑	Mel ↓	STFT ↓
Opus	12	3.93	10.6	10.6	0.772	0.059
DAC	8	3.94	10.2	10.3	**0.440**	0.058
EnCodec	12	3.39	6.8	6.8	0.494	0.105
SNAC	0.98	2.25	-0.3	-0.3	0.500	0.134
VCNAC (concat)	7.9	3.45	7.6	7.6	0.494	0.070
VCNAC (no att)	7.9	3.07	7.1	7.1	0.512	0.076
VCNAC	7.9	**4.16**	**11.3**	**11.3**	0.452	**0.051**

TABLE II
STEREO MUSIC EVALUATION ON FMA-SMALL SUBSET. BEST RESULT IN **BOLD**, SECOND-BEST UNDERLINED. BITRATES IN KBIT/S.

Codec	Bitrate	SI-SDR ↑	SI-SNR ↑	Mel ↓	STFT ↓	ΔIPD ↓	ΔILD ↓
Opus	12	6.1	6.1	1.362	0.439	**0.77**	**0.16**
DAC	8.9	6.7	6.7	0.469	0.414	1.38	0.25
EnCodec	12	8.7	8.7	0.528	0.359	1.00	0.17
SNAC	4.8	4.2	4.2	0.478	0.465	1.53	0.26
VCNAC (concat)	7.9	8.6	8.6	0.479	0.343	1.18	0.17
VCNAC (no att)	7.9	8.4	8.4	0.471	0.349	1.19	0.20
VCNAC	7.9	**9.1**	**9.1**	**0.453**	**0.335**	1.17	0.18

second set of learned positional embeddings is added to the embeddings of the respective channels. The splitting occurs symmetrically to fusion, which implies equal amounts of channel-separated processing in both encoder and decoder. Each output stream then reconstructs its target channel through weight-shared transposed convolutions, again similar to how other neural audio codecs process their embeddings.

Inter-Channel Attention. To enable information exchange between parallel streams before fusion and after splitting, we incorporate adapted Transformer Audio AutoEncoder (TAAE) blocks [23]. Embeddings from different channels are interleaved in time, allowing attention to capture both temporal and cross-channel dependencies. Positional encoding inside TAAE blocks is based solely on temporal position, while channel identity is maintained through the learned channel embeddings. We employ a lightweight configuration with a single attention layer, 4 heads, and a sliding window attending to 2 elements left and right temporally, plus all channel embeddings from other streams within this window. Channel masking during training handles variable batch compositions where samples may have different channel counts.

Implementation Details. Our architecture processes 48kHz audio using 5 convolutional layers with strides (2, 4, 5, 6, 8), achieving 1920× total downsampling and 25Hz frame rate. The total number of parameters is approx. 160M, with the decoder allocated twice the parameters of the encoder, following established practices for neural audio codecs [5]. The 16-dimensional latent undergoes RVQ with 26 codebooks (first: 16384 entries, remaining: 4096 each), yielding 7.85 kbit/s total bitrate at 25Hz token rate. We apply the rotation trick during quantizer training [24] for improved vector utilization.

B. Losses

We generally follow established practices for neural audio codec training, especially DAC [5], [25] and use a combination of multi-scale mel-based reconstruction losses for all channels, together with a discriminator for GAN-like training. We further extract additional audio representations with mid/side processing and downmixing of the audio. For stereo content, we compute mid-side decompositions where the mid channel $M = L + R$ represents monophonic downmixing and the side channel $S = L - R$ captures spatial differences. Multi-scale mel-spectrogram losses are applied to these derived channels. This formulation simultaneously optimizes monophonic compatibility (through the mid channel) and spatial relationship preservation (through the side channel), following established practices in stereo audio processing [7]. For surround configurations, we extend this approach by extracting mid/side representations for the front and rear channels and additionally using standardized downmixing procedures that follow ITU-R BS.775-4 specification [26]. Reconstruction losses are computed between reference and predicted downmixed signals, ensuring spatial audio characteristics remain preserved during reduced-channel rendering. These compatibility losses are applied selectively based on batch composition: mid-side extraction occurs only for stereo and surround samples, while surround downmixing applies to 6-channel content. This enables training on mixed-channel batches without unnecessary computational overhead. We employ the same multi-scale discriminators operating at different temporal resolutions to improve perceptual quality that were introduced by DAC [5]. Adversarial training uses a hinge loss formulation. We use one shared discriminator that operates on single channel inputs and apply it to both original channel audio and derived mid/side downmixed representations. We weigh all extracted and original channels equally for both the mel-based reconstruction losses and the discriminator losses.

IV. EXPERIMENTS

A. Training Data

Due to limited high-quality 5.1 surround datasets, we simulate synthetic surround training content by combining mono speech with stereo music/sound effects into 6-channel configurations (L, R, C, LFE, Ls, Rs). Speech populates the center channel (70% probability, gains 0.4–1.0), primary stereo fills front channels (gains 0.5–1.0), and secondary stereo drives surround channels (80% probability, gains 0.3–0.8). Cross-channel bleed and LFE synthesis (4th-order Butterworth lowpass, 80–120 Hz) model realistic mixing. We use 70% simulated surround data mixed with mono speech and stereo music. Training runs 250k steps with batch size 8 (1.28s chunks). Data sources include LibriTTS [27], LibriVox [28], and general audio datasets.

B. Evaluation Setup

We evaluate reconstruction quality across three modalities: (1) 500 mono speech samples from LibriTTS test set [27]; (2) 150 stereo music tracks from FMA-small [29]; and (3) four Creative Commons movies (Sintel, Big Buck Bunny, Tears of Steel, Elephants Dream) [30] for 5.1 surround evaluation

TABLE III

CHANNEL-WISE EVALUATION OF 5.1 SURROUND AUDIO RECONSTRUCTION QUALITY AND SPATIAL PRESERVATION METRICS, WITH LFE CHANNEL OMITTED DUE TO SPARSE ACTIVITY. BEST VALUES IN **BOLD**, SECOND-BEST VALUES ARE UNDERLINED. BITRATES IN KBIT/S.

Codec	Bitrate	Front L/R						Center			Rear L/R						
		SI-SDR ↑	SI-SNR ↑	Mel ↓	STFT ↓	ΔIPD ↓	ΔILD ↓	SI-SDR ↑	SI-SNR ↑	Mel ↓	SI-SDR ↑	SI-SNR ↑	Mel ↓	STFT ↓	ΔIPD ↓	ΔILD ↓
Opus	12	-9.01	-9.01	2.182	0.213	**1.29**	0.22	-9.08	-9.01	1.961	-11.42	-11.38	1.482	0.070	**1.30**	**0.19**
DAC	16	4.30	4.30	0.516	0.117	1.84	0.29	**1.26**	**1.44**	0.570	2.93	3.07	0.506	0.037	1.80	0.32
EnCodec	9	3.33	3.34	0.625	0.121	1.41	**0.21**	-0.62	-0.61	0.779	2.51	2.56	0.561	0.036	1.48	0.20
SNAC	14.4	4.83	4.82	0.468	0.117	1.89	0.28	0.99	0.99	**0.437**	**4.35**	**4.37**	**0.413**	**0.035**	1.88	0.26
VCNAC (concat)	7.9	5.40	5.40	0.452	0.102	1.66	0.21	-3.06	-3.05	0.915	-4.60	-4.53	0.589	0.040	1.55	0.21
VCNAC (no att)	7.9	4.91	4.91	0.422	0.104	1.72	0.25	-5.49	-5.49	0.786	-1.78	-1.78	0.426	0.036	1.78	0.25
VCNAC	7.9	**5.72**	**5.72**	**0.419**	**0.100**	1.72	0.23	-1.59	-1.56	0.788	-0.44	0.01	0.427	0.037	1.73	0.22

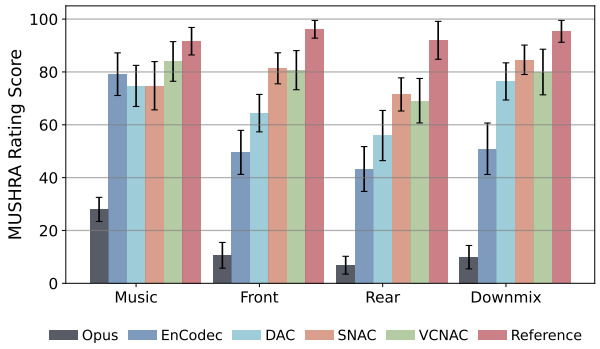

Fig. 2. MUSHRA quality ratings by test category. Mean ratings ± 95% CI for audio codecs across music, front/rear channels, and downmixed content. DAC and SNAC operate at almost double the total bitrate for surround data.

(split into 30 seconds chunks). We compare three model configurations trained on our synthetic dataset. The 6-channel "VCNAC (concat)" baseline incorporates our architectural and training improvements over DAC (including losses) but employs traditional channel fusion by concatenating all channels in the first convolutional layer, requiring silence padding for mono/stereo inputs with unused outputs discarded. It serves as an ablation to demonstrate that the split-channel architecture can reconstruct the audio, while allowing for variable-channel encoding and decoding at the same time. The "VCNAC (no att)" baseline adds the multi-channel processing with split embedding streams, but does not use the proposed attention mechanism. We provide it as an ablation to show that the attention mechanism is necessary for good quality reconstructions.

For external comparison, we select established neural codecs supporting high-fidelity audio ($\geq$44.1 kHz): DAC [5], EnCodec [1], and SNAC [25]. SNAC operates at a fixed bitrate per model checkpoint and does not allow for different bitrates as provided. The bitrates for surround encodings are thus higher. Multi-channel evaluation employs channel-wise encoding strategies: EnCodec's stereo checkpoint processes channel pairs (front L/R, center/LFE, surround L/R) independently, while mono codecs process channels separately. VCNAC natively encodes and decodes content with different channel numbers with the same checkpoint. We normalize total bitrates by adjusting per-channel quantization. E.g., for DAC we use 5 codebooks per channel for stereo content (8.9 kbit/s total bitrate). Evaluation metrics include SI-SNR and PESQ for speech, multi-scale mel-spectrogram distance for music, and spatial-specific measures: frequency-domain deltas (L1 distance between original and reconstructed IPD and ILD values) to assess spatial cue preservation [20]. We compute IPD based on the STFT and ILD on the mel spectrogram (2048-point FFT, 512 hop, 320 mel bins). All metrics are computed per channel and averaged.

C. Results

Speech and Music Reconstruction. VCNAC achieves strong performance across all tested modalities, demonstrating that it can encode and decode variable-channel numbers well. For speech (Table I), VCNAC outperforms all other codecs in PESQ and SI-SDR. It outperforms the fixed-channel "VCNAC (concat)" baseline and the baseline without inter-channel attention "VCNAC (no att)", which validates our design choices. The attention mechanism provides the largest gains for mono speech, where the single-stream representation benefits from the TAAE block's temporal modeling capacity even without cross-channel interaction. Stereo music results (Table II) follow the same trend.

Surround Audio. Table III shows VCNAC achieves the highest front channel reconstruction quality with spatial preservation exceeded only by EnCodec, the sole neural codec with native stereo support. Center and rear SDR and SNR values are generally worse for all codecs, partly due to lower loudness on these channels. The Mel and STFT values indicate good performance for VCNAC, confirmed by the MUSHRA study. The LFE channel has extremely sparse activity in our test data, making SI-SDR/SI-SNR unreliable; VCNAC nonetheless achieves the best SI-SNR (-7.09). One limitation is imperfect surround simulation in training data, particularly loudness balance between front/rear channels. The spatial IPD and ILD metrics show that codecs which jointly process channels have advantages, with Opus, EnCodec and VCNAC showing the best performance. Overall, objective metrics show VCNAC maintains strong performance at lower bitrates than competing models.

Subjective Quality Assessment. We conduct a MUSHRA study evaluating stereo music, front/rear channels of encoded 5.1 surround audio (presented as stereo to the participants), and a stereo downmix of encoded surround audio (following ITU-R BS.775-4 [26]). A 3.5 kHz low-pass filtered anchor

was included per ITU-R BS.1534 [2]. The 10 participants were audio technology professionals who were briefed prior to testing. They were presented with 12 samples (3 for each of the setups) with a length of 7 seconds. The samples were selected to represent a balance of music, dialogue and ambient surround, and were loudness-normalized to -23 LUFS. The results (Figure 2) demonstrate good perceptual quality of VCNAC across all content types. Despite modest rear channel SDR/SNR values, subjective evaluation confirms fair quality; many codecs struggled with rear channel reconstruction due to low loudness levels compared to front channels. Lastly, the quality of downmixed reconstructions is also good, which addresses a common requirement for surround content playback on limited-speaker devices. We note that the downmix evaluation does not fully capture spatial fidelity of the 5.1 reconstruction, as spatial cues are collapsed; a multi-channel listening test would be needed for comprehensive spatial quality assessment. Overall, for surround audio, VCNAC achieves the perceptual quality of state-of-the-art low-bitrate codecs like SNAC while operating at almost half the bitrate. This shows the advantage in coding surround channels jointly and that VCNAC can model channel interactions efficiently.

V. Conclusions

We present VCNAC, a neural audio codec that is able to process mono, stereo, and surround audio within a single architecture and parametrization. Our approach uses parallel channel streams with cross-channel attention and additional channel reconstruction losses. We eliminate the need for separate codecs or channel padding while achieving good reconstruction quality at lower bitrates than the current state-of-the-art. The unified codebook representation enables generative language models to train on the same vocabularies while supporting runtime scalability across channel configurations.

References

[1] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[2] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," Recommendation BS.1534-3, International Telecommunication Union, 2015.

[3] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[4] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27980–27993, 2023.

[6] Jean-Marc Valin, Koen Vos, and Timothy Terriberry, "Definition of the opus audio codec," Tech. Rep., 2012.

[7] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, "Stable audio open," in *Proc. ICASSP*. IEEE, 2025.

[8] Zhongweiyang Xu, Yong Xu, Vinay Kothapally, Heming Wang, Muqiao Yang, and Dong Yu, "Spatialcodec: Neural spatial speech coding," in *Proc. ICASSP*. IEEE, 2024.

[9] Anton Ratnarajah, Shi-Xiong Zhang, and Dong Yu, "Banc: Towards efficient binaural audio neural codec for overlapping speech," in *Proc. ICASSP*. IEEE, 2025.

[10] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons, "Fast timing-conditioned latent audio diffusion," in *Proc. ICML*, 2024.

[11] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47704–47720, 2023.

[12] Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang, "Songgen: A single stage auto-regressive transformer for text-to-song generation," *arXiv preprint arXiv:2502.13128*, 2025.

[13] Mark Levy, Bruno Di Giorgi, Floris Weers, Angelos Katharopoulos, and Tom Nickson, "Controllable music production with diffusion models and guidance gradients," *arXiv preprint arXiv:2311.00613*, 2023.

[14] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf, "Moûsai: Efficient text-to-music diffusion models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 8050–8068.

[15] Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner, "Diff-a-riff: Musical accompaniment co-creation via latent diffusion models," *arXiv preprint arXiv:2406.08384*, 2024.

[16] Marco Pasini, Stefan Lattner, and György Fazekas, "Music2latent2: Audio compression with summary embeddings and autoregressive decoding," in *Proc. ICASSP*. IEEE, 2025.

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

[18] Christian J Steinmetz and Joshua D Reiss, "auraloss: Audio focused loss functions in pytorch," in *Digital music research network one-day workshop (DMRN+ 15)*, 2020.

[19] Alessandro Carlini, Camille Bordeau, and Maxime Ambard, "Auditory localization: a comprehensive practical review," *Frontiers in Psychology*, vol. 15, pp. 1408073, 2024.

[20] Vikas Tokala, Eric Grinstein, Mike Brookes, Simon Doclo, Jesper Jensen, and Patrick A Naylor, "Binaural speech enhancement using deep complex convolutional transformer networks," in *Proc. ICASSP*. IEEE, 2024.

[21] Bahareh Tolooshams and Kazuhito Koishida, "A training framework for stereo-aware speech enhancement using deep neural networks," in *Proc. ICASSP*. IEEE, 2022, pp. 6962–6966.

[22] Yunpeng Li, Kehang Han, Brian McWilliams, Zalan Borsos, and Marco Tagliasacchi, "Spectrostream: A versatile neural codec for general audio," *arXiv preprint arXiv:2508.05207*, 2025.

[23] Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu, "Scaling transformers for low-bitrate high-quality speech coding," *arXiv preprint arXiv:2411.19842*, 2024.

[24] Christopher Fifty, Ronald G. Junkins, Dennis Duan, Aniketh Iyengar, Jerry W. Liu, Ehsan Amid, Sebastian Thrun, and Christopher Ré, "Restructuring vector quantization with the rotation trick," 2025.

[25] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer, "Snac: Multi-scale neural audio codec," *arXiv preprint arXiv:2410.14411*, 2024.

[26] ITU-R, "Multichannel stereophonic sound system with and without accompanying picture," Recommendation BS.775-4, International Telecommunication Union, 2022.

[27] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[28] Jodi Kearns, "Librivox: Free public domain audiobooks," 2014.

[29] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[30] Blender Foundation, "Blender open movie collection: Sintel, Big Buck Bunny, Tears of Steel, Elephants Dream," 2006–2012, Creative Commons licensed test sequences.