NEXT-SCALE AUTOREGRESSION ON SPECTROGRAMS FOR SOUND GENERATION

Eleonora Ristori[*†] **Luca Bindini**[*] **Paolo Frasconi**
AI Lab, DINFO
Università degli Studi di Firenze
Florence, Italy
{eleonora.ristori, luca.bindini, paolo.frasconi}@unifi.it

ABSTRACT

Research on audio generation has progressively shifted from waveform-based approaches to spectrogram-based methods, which more naturally capture harmonic and temporal structures. At the same time, advances in image synthesis have shown that autoregression across scales, rather than tokens, improves coherence and detail. Building on these ideas, we introduce MARS (Multi-channel AutoRegression on Spectrograms), which, to the best of our knowledge, is the first adaptation of next-scale autoregressive modeling to the spectrogram domain. MARS treats spectrograms as multi-channel images and employs *channel multiplexing* (CMX), a reshaping strategy that reduces spatial resolution without information loss. A shared tokenizer provides consistent discrete representations across scales, enabling a transformer-based autoregressor to refine spectrograms from coarse to fine resolutions efficiently. Experiments on a large-scale dataset demonstrate that MARS performs comparably or better than state-of-the-art baselines across multiple evaluation metrics, establishing an efficient and scalable paradigm for high-fidelity sound generation.

1 INTRODUCTION

Audio generation has made remarkable strides in recent years, driven primarily by advances in generative models such as generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs). Two primary approaches dominate the field: the first one generates waveforms directly in the time domain, while the other synthesizes spectrograms in the frequency domain, followed by an inverse short-time Fourier transform (ISTFT) to reconstruct the raw audio. There are many different methods that operate in the time domain, such as Variational Auto-Encoder (VAE) models (Peng et al., 2020), GAN models (Yamamoto et al., 2020; Binkowski et al., 2020), and DDPM models (Kong et al., 2021), which adapt traditional generative models for the generation of waveforms. The second trend is also characterized by the application of several generative models for the generation of spectrograms. GANSynth (Engel et al., 2019), for example, applies GANs for this task while EDMSound (Zhu et al., 2023) uses DDPMs. These trends suggest that reusing and adapting models from other domains to audio generation is an effective way to enhance performance, creating high-quality audio that preserves fine-grained details.

In parallel, the field of generative modeling for images has undergone rapid innovation. Most notably, Visual AutoRegressive (VAR) models starting from the first work by Tian et al. (2024) and followed by several refinements such as ImageFolder by Li et al. (2025), have emerged as a new state-of-the-art in image synthesis. These models redefine autoregression by predicting the *next-scale* rather than the next token, progressively increasing image scale at each step. Recent work shows they produce high-resolution, semantically consistent images by capturing both local texture and global structure. These strengths are highly relevant to the challenges faced in audio generation, particularly when working with spectrograms, where high-quality samples are obtained by preserving even the smallest frequency contributions. A recent work by Qiu et al. (2024) introduced

[*]These authors contributed equally to this work. [†]Corresponding author.

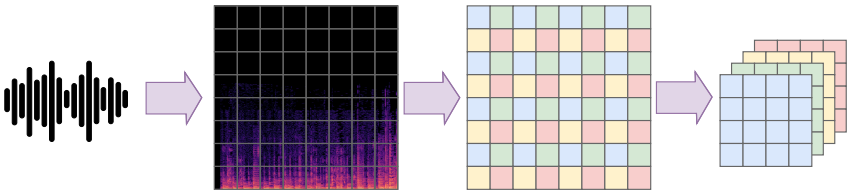

Figure 1: Audio preprocessing pipeline for tokenizer input preparation and channel multiplexing (CMX) for reducing input resolution.

next-scale prediction for audio generation in the waveform domain through a scale-level audio tokenizer and autoregressive modeling over raw-audio tokens.

Motivated by these developments, in this work we propose MARS (Multi-channel AutoRegression on Spectrograms), an autoregressive framework that adapts next-scale prediction to the spectrogram domain. MARS treats spectrograms as multi-channel images and leverages a shared tokenizer to learn consistent discrete representations across scales. The autoregressive model then progressively predicts higher-resolution tokens conditioned on coarser ones, enabling hierarchical refinement of the generated spectrogram.

A central component of our approach is a novel preprocessing strategy called *channel multiplexing (CMX)*, which reduces spatial resolution while redistributing information along the channel dimension. This technique allows us to preserve spectral fidelity while keeping computational cost manageable, making the method scalable to long and wide bandwidth audio recordings. By decoupling resolution from information density, CMX ensures that MARS can exploit the inductive biases of modern neural architectures optimized for multi-channel data.

To rigorously evaluate MARS, we rely on the NSynth dataset (Engel et al., 2017), a widely used benchmark for audio generation. Following the protocol of Vinay & Lerch (2022), we employ sample diversity metrics including NDB/k and Inception-based scores for pitch and instrument classes, which assess whether the samples capture the distribution of musical attributes, and embedding-based similarity metrics such as kernel Inception distance (KID) and Fréchet audio distance (FAD), which compare reconstructed and reference audio distributions in learned feature spaces. Our results demonstrate that MARS achieves competitive performance compared to leading models such as DDSP (Engel et al., 2020), DiffWave (Kong et al., 2021), and NSynth (Engel et al., 2017). Beyond strong quantitative results, MARS also provides a new perspective on autoregressive modeling for audio, highlighting that thanks to its scale-wise refinement and channel multiplexing design, it achieves this level of quality while keeping computational costs contained, offering a favorable balance between performance and efficiency.

2 Proposed Method

Training AR models consists of two stages: first, learning a tokenizer that remains consistent across all resolutions; and second, training the AR model to predict higher-resolution tokens conditioned on the lower-resolution token map.

2.1 Audio Preprocessing

We first convert each waveform into a spectrogram using a STFT, retaining only the amplitude since the phase is later (after inference) reconstructed with the Griffin–Lim algorithm (Griffin & Lim, 1984). As an example, an 8s waveform at 16 kHz with a 1024-point STFT and hop size of 256 yields a spectrogram of 512×512, already matching the largest image size attempted with VAR (Tian et al., 2024). Higher sampling rates further enlarge spectrograms, leading to excessive memory consumption. Moreover, processing such large inputs requires greater network depth; in fact, increasing input resolution from 256×256 to 512×512 in the original VAR model raised the parameter count from 310M to 2.3B.

To address this problem, which makes our experiments quickly unfeasible, we propose *channel multiplexing* (CMX), which reduces spatial dimensions by redistributing values across channels in a

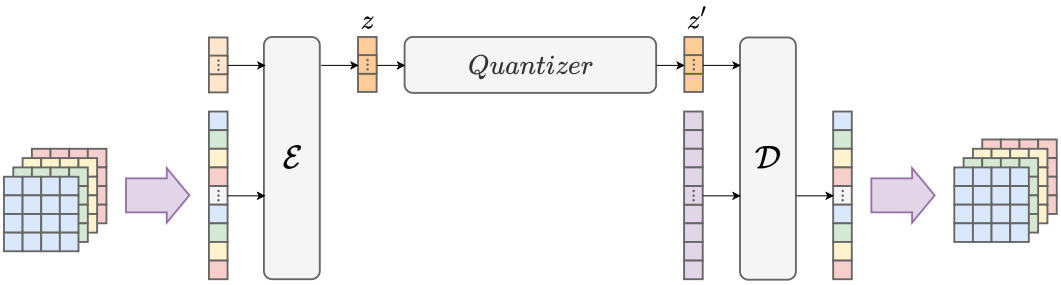

Figure 2: *Tokenizer architecture.* The tokenizer is adapted from (Li et al., 2025), improving upon the original VAR tokenizer (Tian et al., 2024). The input spectrogram is partitioned into patches of size $L \times L$ and concatenated with S learnable tokens before being processed by a transformer encoder $\mathcal{E}$, producing latent representations z. These are discretized by a vector quantizer to obtain z', which are then combined with another set of $L \times L$ learnable tokens and passed to a decoder $\mathcal{D}$ for reconstruction.

chessboard-like scheme (Figure 1). For spectrograms derived from an N-point STFT, the reshaping procedure jointly reorganizes the frequency and temporal dimensions, producing a tensor with the desired reduced spatial resolution and with a channel cardinality C scaling with N, the signal length, and sampling rate. In this way, CMX accommodates inputs of varying length and sampling rate by adjusting C, while preserving the original data information and substantially reducing spatial overhead. In addition, for our tokenizer presented in Figure 2, even without increasing network depth, enlarging the input dimension $M \times M$ leads to a parameter growth of $O(K^2 + S)$ where $K = M/L$, L is the patch size, and S the number of learnable tokens, already scaling with the square of the input dimension, while remaining nearly constant with respect to channel dimensionality C.

More generally, CMX decouples information density from time–frequency resolution. Rather than discarding or compressing data, it reorganizes it into a format naturally suited to architectures optimized for multi-channel inputs (e.g., RGB images). This enables long recordings or high-fidelity signals to be processed with bounded memory cost, and the same principle can be extended beyond audio to images, video, medical scans, or structured time series. CMX thus provides a compact, lossless representation that scales efficiently, reducing memory and compute requirements while maintaining full data fidelity.

2.2 Tokenizer

The tokenizer architecture and training procedure are adapted from ImageFolder (Li et al., 2025), extending the original VAR tokenizer (Tian et al., 2024). The overall design is illustrated in Figure 2. The tokenizer is trained using a composite objective that combines three terms:

$$\mathcal{L} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{VQ}\mathcal{L}_{VQ} + \lambda_{ad}\mathcal{L}_{ad}. \tag{1}$$

Here, $\mathcal{L}_{recon}$ is an L_2 reconstruction loss that ensures fidelity between the reconstructed and ground-truth spectrograms; $\mathcal{L}_{VQ}$ is the vector quantization loss, which aligns the encoder outputs with the nearest codebook entries; and $\mathcal{L}_{ad}$ is an adversarial loss applied through a PatchGAN discriminator (Isola et al., 2017), encouraging reconstructions to be indistinguishable from real spectrograms.

2.3 Autoregressive Model

Once the tokenizer is trained, the autoregressive model is trained following (Tian et al., 2024) to conduct the next-scale AR modeling for a faster inference speed. The AR model is designed as a transformer-based architecture that progressively predicts tokens across different scales, effectively reducing the autoregressive sequence length to the number of scales and accelerating generation. During training, the model learns to condition predictions of fine-scale tokens on coarser ones, enabling a hierarchical refinement of the output. This multi-scale training strategy not only improves efficiency but also enhances consistency across scales, leading to higher-quality generated samples and reduced inference cost compared to standard autoregressive approaches.

Table 1: Reconstruction results according to the objective metrics reported in (Vinay & Lerch, 2022). The best results are highlighted in bold, and the second best are underlined. The last (grayed) row reports metrics on generated audio (i.e., after the autoregressive module), and values are not directly comparable against those in the previous four rows.

Model	NDB/k (↓)	PKID (↓)	IKID (↓)	PIS (↑)	IIS (↑)	MSE (↓)	MAE (↓)	FAD (↓)
Diffwave	0.74	0.0093	0.0021	2.3814	**5.6477**	0.0291	0.1369	7.9488
DDSP	<u>0.20</u>	<u>0.0053</u>	<u>0.0020</u>	**3.3224**	<u>5.3371</u>	**0.0130**	**0.0666**	**1.1519**
NSynth	0.74	0.0101	0.0024	2.3238	4.6364	0.0329	0.1224	4.0590
MARS (ours)	**0.19**	**0.0035**	**0.0015**	<u>2.9602</u>	5.2047	<u>0.0143</u>	<u>0.0915</u>	<u>1.6429</u>
MARS (generated)	0.15	0.0066	0.0017	3.2220	4.5357	0.0201	0.0795	1.8833

3 Experimental Evaluation

3.1 Experimental Setup

We tested our autoregressive model using NSynth (Engel et al., 2017), which is a dataset of over 300,000 musical notes, each with a unique pitch, timbre, and envelope. The samples are monophonic signals of approximately 4s and with a sampling rate of 16 kHz. We converted them into spectrograms using a 1024-point STFT and rearranged them using CMX into tensors of dimension $256 \times 256 \times 2$, splitting frequencies among two channels.

The tokenizer transformer encoder follows a DINOv2-base architecture and is initialized with random weights. The tokenizer employs a codebook of size 16,384. We trained the tokenizer for 400 epochs following the training settings proposed in Li et al. (2025), with loss weights set to $\lambda_{recon} = \lambda_{VQ} = 1$ and $\lambda_{ad} = 0.5$. The autoregressive model was then trained for 350 epochs following Tian et al. (2024). We ran our experiments on a single RTX 5000 Ada GPU with 32GB of RAM.

3.2 Evaluation Metrics

To assess our model, we follow the evaluation protocol proposed by Vinay & Lerch (2022) on the NSynth dataset. We evaluate our tokenizer for audio reconstruction using complementary metrics targeting several aspects of audio quality. Reconstruction quality is measured via MSE and MAE on mel-spectrograms. Sample diversity and semantic consistency are assessed using NDB/k (Richardson & Weiss, 2018), Pitch Inception Score (PIS), and Instrument Inception Score (IIS) (Nistal et al., 2021). Perceptual and distributional similarity between reference and reconstructed audio is evaluated with PKID, IKID (Nistal et al., 2021), and Fréchet Audio Distance (FAD) (Kilgour et al., 2019).

We additionally evaluate our autoregressive (AR) model for audio generation using the same metrics; in this case, MSE and MAE are computed against the closest spectrogram in the NSynth test set for each generated sample.

3.3 Experimental Results

Table 1 compares our MARS with NSynth (Engel et al., 2017), DDSP (Engel et al., 2020), and DiffWave (Kong et al., 2021). MARS achieves the best scores in NDB/k, PKID, and IKID, indicating superior sample diversity and fidelity in both pitch and timbre. It also ranks among the top methods for MSE, MAE, and FAD, reflecting accurate reconstruction and high perceptual quality. PIS and IIS, which measure similarity on sound properties rather than audio quality, remain competitive.

The last row reports metrics on generated audio, which are not directly comparable to the reconstruction metrics; nonetheless, the generated samples maintain low reconstruction error and perceptual similarity, confirming that MARS effectively preserves signal characteristics during synthesis.

3.4 EFFECTIVENESS OF CMX

We use a toy example to assess the effect of CMX. We trained the tokenizer of our AR model for 200k steps under two different settings: (i) truncated spectrograms of size $256 \times 256 \times 1$, extracted from the original $512 \times 512 \times 1$ inputs, and (ii) truncated spectrograms rearranged with CMX into $128 \times 128 \times 4$. The CMX-based setup reduced training time by a factor of $1.5\times$ while maintaining reconstruction quality close to that of the truncated baseline. In terms of reconstruction accuracy, CMX performed better, yielding lower errors, with improvements of 0.002 in MSE and 0.022 in MAE. On the other hand, perceptual quality showed a slight decline, as reflected by a 0.19 increase in FAD. Importantly, CMX enables the retention of the full frequency content, which is crucial for preserving audio quality, while remaining computationally feasible due to its substantial memory savings. By contrast, frequency truncation results in a significant loss in quality, as indicated by FAD difference of 2.16 and an MSE difference of 0.01 compared to our experiments with inputs of size $256 \times 256 \times 2$. This confirms that CMX is an effective compromise between efficiency and fidelity. Overall, these results indicate that CMX lowers memory requirements, improves reconstruction accuracy, and preserves frequency resolution, with only a minor trade-off in perceptual quality.

4 CONCLUSION

In this work, we introduced MARS (Multi-channel AutoRegression on Spectrograms), a novel framework for audio generation that adapts next-scale autoregression from image synthesis to the spectrogram domain. By introducing the channel multiplexing (CMX) technique, MARS reduces spatial resolution while preserving frequency information, enabling scalable training with a manageable memory footprint. A shared tokenizer across scales further ensures consistent discrete representations, allowing the autoregressive model to refine spectrograms hierarchically.

The experimental evaluation on the NSynth dataset demonstrated that MARS achieves competitive or superior performance compared to state-of-the-art baselines across multiple metrics, confirming both the effectiveness of CMX in balancing efficiency with fidelity and the robustness of the overall framework for high-quality audio generation.

REFERENCES

Mikolaj Binkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. In *ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Jesse H. Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In Doina Precup and Yee Whye Teh (eds.), *ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1068–1077. PMLR, 2017.

Jesse H. Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. In *ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Jesse H. Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: differentiable digital signal processing. In *ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. doi: 10.1109/TASSP.1984.1164317.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Gernot Kubin and Zdravko Kacic (eds.), *20th Annual Conference of the International Speech Communication Association,*

Interspeech 2019, Graz, Austria, September 15-19, 2019, pp. 2350–2354. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-2219.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. In *ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Javier Nistal, Stefan Lattner, and Gaël Richard. Comparing representations for audio synthesis using generative adversarial networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 161–165. IEEE, 2021.

Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. Non-autoregressive neural text-to-speech. In *International conference on machine learning*, pp. 7586–7598. PMLR, 2020.

Kai Qiu, Xiang Li, Hao Chen, Jie Sun, Jinglu Wang, Zhe Lin, Marios Savvides, and Bhiksha Raj. Efficient autoregressive audio modeling via next-scale prediction. *CoRR*, abs/2408.09027, 2024. doi: 10.48550/ARXIV.2408.09027. URL https://doi.org/10.48550/arXiv.2408.09027.

Eitan Richardson and Yair Weiss. On gans and gmms. *Advances in neural information processing systems*, 31, 2018.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Ashvala Vinay and Alexander Lerch. Evaluating generative audio systems and their metrics. In Preeti Rao, Hema A. Murthy, Ajay Srinivasamurthy, Rachel M. Bittner, Rafael Caro Repetto, Masataka Goto, Xavier Serra, and Marius Miron (eds.), *ISMIR 2022*, pp. 858–865, 2022.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020*, pp. 6199–6203. IEEE, 2020.

Ge Zhu, Yutong Wen, Marc-André Carbonneau, and Zhiyao Duan. Edmsound: Spectrogram based diffusion models for efficient and high-quality audio synthesis. *arXiv preprint arXiv:2311.08667*, 2023.