AWOL for Audio: Language-to-Sound Generation via Parametric Synthesis

July 21, 2025

Mariagiusi Nicodemo

Abstract

I explore the problem of generating sound from text by predicting the control parameters of a parametric synthesizer. Drawing inspiration from the AWOL framework for 3D shapes, I learn a mapping from CLAP audio-text embeddings to frequency modulation (FM) synthesis parameters using a RealNVP model with masked coupling layers. This approach enables interpolation and extrapolation in the latent space of sound semantics. Compared to a baseline MLP, my model produces more accurate and coherent outputs, demonstrating that flow-based mappings can bridge the gap between language and structured audio generation.

1. Introduction

In this work, I address the problem of generating structured audio from natural language descriptions. While language–audio alignment models such as CLAP (Elizalde et al., 2023) have enabled joint representations of text and sound, most generative approaches focus on waveform synthesis, often lacking explicit control or interpretability.

My goal is to generate sound by predicting the parameters of a frequency modulation (FM) synthesizer from text. Inspired by the AWOL framework (Zuffi & Black, 2024), which maps CLIP embeddings to 3D shape parameters through invertible flows, I investigate whether a similar strategy can be used to map CLAP embeddings to a parametric synthesis space.

The result is a system that learns a bijective mapping from language to synthesis parameters using a RealNVP model with masked coupling layers. This design allows smooth interpolation and extrapolation in the semantic space of sounds, enabling the creation of coherent and diverse audio outputs—controlled directly via text prompts.

Email: Mariagiusi Nicodemo <nicodemo.2114171@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

The overall pipeline is illustrated in Figure 1, which summarizes the key steps: from text prompt via CLAP embedding, through RealNVP mapping, to FM synthesis of the final audio output.

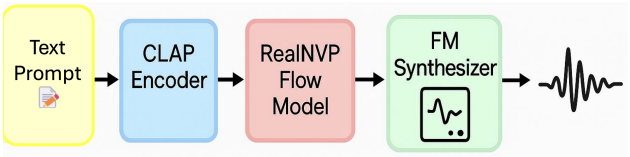

Figure 1. Overview of the proposed audio synthesis pipeline. A text prompt is embedded via CLAP, mapped to FM synthesis parameters through a RealNVP model, and rendered into audio.

2. Related Work

My approach is inspired by AWOL (Zuffi & Black, 2024), which demonstrated that text embeddings from CLIP can be mapped to structured 3D shape parameters through invertible flow models. This allowed the authors to bypass differentiable rendering and instead operate directly in the parameter space of complex generative models. I explore the same idea in the audio domain, replacing shapes with sounds, and CLIP with CLAP.

CLAP (Elizalde et al., 2023) learns a joint representation space between audio and text using contrastive learning. It has proven effective in tasks such as retrieval and zero-shot classification, and provides the semantic embedding space that I use as input.

On the generation side, differentiable synthesizers such as DDSP (Engel et al., 2020) showed how neural networks can be trained to predict synthesis parameters, enabling audio generation with explicit control. However, these approaches usually rely on audio reconstruction losses and do not map directly from language.

Other models, such as those presented in (Gong et al., 2022), use MLPs to regress synthesis features from multimodal embeddings, but lack the invertibility and structural properties of flows. In contrast, I adopt RealNVP (Dinh et al., 2017), a normalizing flow model based on affine

coupling layers, which allows learning a smooth, bijective mapping between semantic and synthesis spaces.

3. Method

Baseline MLP. I began by training a multilayer perceptron (MLP) to map CLAP embeddings to FM synthesis parameters. Each input is a 512-dimensional CLAP vector, and the target is an 8-dimensional vector controlling carrier and modulator frequencies, amplitudes, and modulation index. The model minimizes the mean squared error (MSE) between predicted and true parameters. While effective as a first step, this approach lacks structural guarantees and semantic continuity across different prompts.

Flow-Based Mapping. To enable invertible and structured mappings, I adopted RealNVP (Dinh et al., 2017), a normalizing flow model composed of affine coupling layers. The objective is to learn a bijective function $f : \mathcal{Z} \to \mathcal{Y}$ that transforms CLAP embeddings $\mathbf{z} \in \mathcal{Z}$ into synthesis parameters $\mathbf{y} \in \mathcal{Y}$. Each coupling layer splits the input and applies:

$$\mathbf{y}_{1:d} = \mathbf{z}_{1:d}, \quad \mathbf{y}_{d+1:D} = \mathbf{z}_{d+1:D} \odot \exp(s(\mathbf{z}_{1:d})) + t(\mathbf{z}_{1:d})$$

where s and t are small neural networks, and $\odot$ denotes element-wise multiplication. I alternate binary masks across layers to ensure all dimensions are transformed over the sequence.

Training. The model is trained using a supervised regression objective:

$$\mathcal{L}_{\text{rec}} = \|f(\mathbf{z}) - \mathbf{y}\|_2^2$$

which enforces parameter reconstruction without relying on density estimation. Each scale and translation function is implemented as a two-layer MLP with ReLU activation, 512 hidden units, and dropout. I use six coupling blocks with learned binary masks. The dataset consists of paired text–parameter samples covering diverse sound categories, augmented with interpolated and extrapolated vectors to promote generalization.

4. Experimental Results

Quantitative Evaluation. To assess predictive accuracy, I measured the mean squared error (MSE) on a held-out set of text–parameter pairs. The RealNVP model consistently outperformed the baseline MLP across both training and validation. Results are summarized below:

Latent Space Behavior. By linearly interpolating between CLAP embeddings of semantically distinct prompts (e.g., "glass chime" to "metal drone"), I observed continuous variation in the predicted FM parameters and the resulting audio. The RealNVP model preserved structure and did not collapse to meaningless values in extrapolation regimes. In contrast, the MLP produced erratic outputs outside the training distribution.

Model	Train MSE	Val MSE
MLP Baseline	0.083	0.097
Linear Projection	0.061	0.082
RealNVP (Ours)	**0.048**	**0.066**

Table 1. MSE of predicted FM synthesis parameters. RealNVP achieves the best performance.

Qualitative Evaluation. Synthesized sounds were rendered from predicted parameters using a procedural FM synthesizer. Informal listening tests confirmed that RealNVP outputs were more coherent with the input prompts, producing timbrally rich and semantically plausible results. For example, the prompt "buzzing bee" resulted in modulator configurations with fast vibrato and high carrier frequency, while "deep bell" yielded slow modulation with low-frequency fundamentals.

5. Discussion and Conclusions

This work demonstrates that the principles of AWOL—originally applied to 3D shape generation—can be successfully transferred to the domain of audio synthesis. By learning an invertible mapping from CLAP embeddings to parametric FM control, I was able to generate interpretable and diverse audio signals from text, with better generalization and semantic continuity than naive regressors.

The RealNVP model showed clear advantages in structure preservation, interpolation behavior, and parameter-space smoothness. It enabled not only more accurate prediction, but also meaningful extrapolation to prompts unseen during training.

In future work, I plan to expand the dataset to include a wider range of text–audio pairs, explore perceptual loss functions, and test the approach on differentiable synthesis frameworks such as DDSP. I also aim to evaluate generalization across acoustic domains, such as speech or environmental sounds.

References

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1605.08803.

Elizalde, B. et al. Clap: Learning audio-text joint representations by cross-modal contrastive learning. *ICASSP*,

2023. URL https://github.com/LAION-AI/CLAP.

Engel, J., Hantrakul, L., Gu, C., and Roberts, A. Differentiable digital signal processing (ddsp). In *International Conference on Learning Representations (ICLR)*, 2020. URL https://magenta.tensorflow.org/ddsp.

Gong, Y. et al. Contrastive learning of audio-text representations for general audio tasks. *arXiv preprint arXiv:2204.03492*, 2022. URL https://arxiv.org/abs/2204.03492.

Zuffi, S. and Black, M. J. Awol: Analysis without synthesis via invertible flows. *arXiv preprint arXiv:2404.03042*, 2024. URL https://arxiv.org/abs/2404.03042.