MambaWave: Neural Raw Audio Upsampling with Mamba and Diffusion Models

September 16, 2024

Simone Facchiano, Sapienza Universty of Rome

Abstract

Recently, Mamba has been recognized as one of the most promising frameworks for operating on very long sequences with long-range dependencies. At the same time, Diffusion Models remain the state of the art in many generative task applications, ranging from the vision to the audio domain. This paper presents MambaWave, the first known model combining Mamba and Diffusion Models for the audio upsampling task. The code is public and can be visited at the following link: `https://github.com/simonefacchiano/MambaWave.git`.

1. Introduction

Neural audio upsampling refers to the process of generating high sampling rate audios from low sampling rate signals. Several works already applied deep neural networks to audio super-resolution (Kuleshov et al., 2017; Lim et al., 2018; Kim & Sathe, 2019), but papers using Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021) are still few. Furthermore, although Diffusion Models have already been combined with *State Space Models*, particularly Mamba (Gu & Dao, 2024), in the vision domain in recent months, hybrid techniques of this type have not yet seen applications in modeling audio signals. One reason for this poor exploration is that the generation of raw audios is challenging because of the *high dimensionality* of waveforms. In fact, high-resolution signals count tens of thousands of samples per second, and are characterized by a series of long-distance relations in multiple timescales.

The proposed model, named **MambaWave**, draws inspiration from the *DiffWave* architecture (Kong et al., 2021) but is enhanced by incorporating Mamba layers, creating a hybrid approach. The goal of this work is to synthesize 48 kHz waveforms from lower resolution 24 kHz inputs, and to achieve this the model was trained on paired audio samples coming from the VoiceBank speech database (Valentini-Botinhao, 2017). MambaWave is the first known work combining Diffusion Models with Mamba for the audio upsampling task.

Email: Simone Facchiano <facchiano.1919922@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

2. Related work

This section provides an overview of the main foundational work in this paper. Diffusion Models and SSMs, particularly Mamba, are therefore discussed.

2.1. Diffusion Models, DiffWave and NU-Wave

Diffusion Models (DMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021) are generative models that iteratively transform Gaussian noise into structured samples by learning from a data distribution. DMs involve a non-learnable *forward diffusion process*, where Gaussian noise is gradually added to data, and a *reverse process*, that aims to recover the original data by reversing the diffusion steps. However, since the exact reverse transition probabilities are intractable, a neural network is used to approximate this reverse process. These *unconditional models* offer limited control over the generation process (Ho & Salimans, 2022). To improve control, the process can be adapted to condition on inputs like text, images, or audios, allowing for more specific outputs.

DMs have demonstrated considerable efficacy in audio generation, a domain in which numerous contributions have established significant advancements over time. Among these, DiffWave (Kong et al., 2021) stands out as a particularly pivotal development, employing DMs to facilitate raw audio synthesis. The DiffWave architecture has since been replicated in multiple works, including NU-Wave and NU-Wave2 (Lee & Han, 2021; Han & Lee, 2022) for audio upsampling. Furthermore, it has been integrated with SSMs (discussed in Section 2.2) to model ECGs (Alcaraz & Strodthoff, 2023a) and Time Series (Alcaraz & Strodthoff, 2023b).

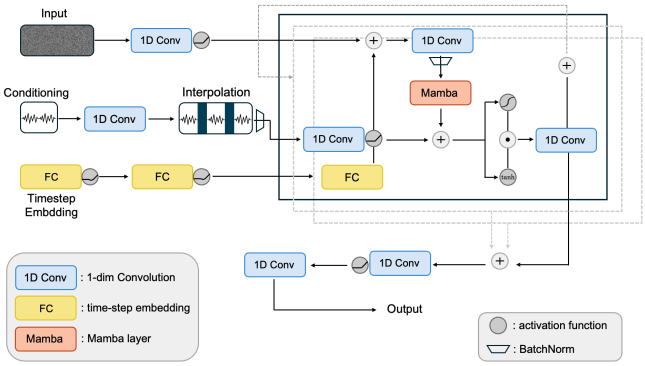

Figure 1. Architecture of MambaWave. Inspired by that of DiffWave, it relies on a stack of Residual Blocks containing a Mamba layer.

2.2. Mamba, SSMs and SaShiMi

At the end of 2023, a new framework called **Mamba** (Gu & Dao, 2024) began to be recognized by the scientific community as a potential alternative to the Transformer architecture (Vaswani et al., 2017). Mamba took inspiration from the classic State Space Models (SSMs) introduced by Kalman in the 1960s in the field of control engineering and recently rediscovered for sequence modeling with the introduction of Structured SSMs (**S4**) (Gu et al., 2022). Based on its hybrid structure of CNNs and RNNs, Mamba not only scaled quasi-linearly with sequence length, but showed an exceptional ability to model long-range dependencies in data, up to million-length sequences. This makes Mamba particularly suitable for tasks where understanding distant relationships in the data is crucial, such as modeling raw audio. In fact, audio generation by waveform modeling was precisely one of the first applications of S4 in the *SaShiMi* model from the paper "It's Raw! Audio Generation with State-Space Models" (Goel et al., 2022).

3. Method

The architectural design of MambaWave draws upon the principles established by DiffWave, but focuses solely on conditional generation on low-resolution audio. A 48 kHz input audio signal is noised according to the forward diffusion process discussed in Section 2.1 and passed through a one-dimensional convolution. Concurrently, a 24 kHz conditioning audio is processed by a second convolution layer and brought to the same length as the input audio (i.e., doubled) through the use of parameter-free interpolation, and then normalized through BatchNorm. Subsequently, the time-step is processed by two fully connected layers. From here, the three components enter the Residual Block. The choice of activation functions is fairly arbitrary, but guided by the empirical results observed during the model validation phase. Although the use of `tanh` may be more natural than `ReLU` at some points in the architecture (and indeed, sometimes resulting in a lower training loss), this produced a more muffled output audio, lacking the high frequencies. Within the Residual Layer there is onea Mamba block, which is the core of the entire architecture. The original goal was to insert it after the sum with audio conditioning, but this resulted in a strong numerical instability probably due to hardware limitations. An attempt was also made with a double Mamba block, but without success. The complete architecture of MambaWave is illustrated in Figure 1.

4. Results

The performance of the model was evaluated in terms of generation time, signal-to-noise ratio (SNR), and log-spectral distance (LSD). Two versions of the model were analyzed. The first was trained using MSE loss, the second using L1 loss. In both cases the training time was around 5 hours (5 epochs), a learning rate scheduling was used and gradient clipping as a regularization mechanism was implemented. The number of convolution channels (32) and the number of residual layers (11) was kept fixed in both cases. In fact, reducing these hyperparameters led to a slight performance degradation, while it was impossible to increase their values due to hardware limitations. For reference, NU-Wave uses a channel number of 64 and 30 residual layers. The results obtained are shown in Table 1:

Model	Gen. Time (s) ↓	SNR ↑	LSD ↓
MambaWave (*MSE*)	25.54	0.367	2.14
MambaWave (*L1*)	25.29	-0.43	2.52

Table 1. Results of the two versions of MambaWave. The version trained by minimizing MSE is significantly better than the version with L1.

5. Discussion and conclusion

MambaWave claims to be the first work combining Mamba and Diffusion Models for the audio upsampling task. Despite this, the results obtained are still far from the performance of works such as NU-Wave and NU-Wave2. In the next stages of the work, the focus will be on making the code more hardware-efficient so that the number of convolutional and residual layers can be increased. Also, it will be interesting to observe the impact of a second Mamba layer inserted later in the Residual Block.

References

Alcaraz, J. M. L. and Strodthoff, N. Diffusion-based conditional ecg generation with structured state space models, 2023a. URL https://arxiv.org/abs/2301.08227.

Alcaraz, J. M. L. and Strodthoff, N. Diffusion-based time series imputation and forecasting with structured state space models, 2023b. URL https://arxiv.org/abs/2208.09399.

Goel, K., Gu, A., Donahue, C., and Ré, C. It's raw! audio generation with state-space models, 2022. URL https://arxiv.org/abs/2202.09729.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2024.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

Han, S. and Lee, J. Nu-wave 2: A general neural audio upsampling model for various sampling rates. In *Interspeech 2022*, interspeech$_2$022, *pp.* 4401–4405.$ISCA$, $September$2022.doi :. URL http://dx.doi.org/10.21437/Interspeech.2022-45.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Kim, S. and Sathe, V. Bandwidth extension on raw audio via generative adversarial networks, 2019. URL https://arxiv.org/abs/1903.09027.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis, 2021. URL https://arxiv.org/abs/2009.09761.

Kuleshov, V., Enam, S. Z., and Ermon, S. Audio super resolution using neural networks, 2017. URL https://arxiv.org/abs/1708.00853.

Lee, J. and Han, S. Nu-wave: A diffusion probabilistic model for neural audio upsampling. In *Interspeech 2021*, interspeech$_2$021, *pp.* 1634–1638.$ISCA$, $August$2021.10.21437/$interspeech$.2021 − 36.URL.

Lim, T., Yeh, R., Xu, Y., Do, M., and Hasegawa-Johnson, M. Time-frequency networks for audio super-resolution. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018 - Proceedings*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 646–650, United States, September 2018. Institute of Electrical and Electronics Engineers Inc. ISBN 9781538646588. 10.1109/ICASSP.2018.8462049. Publisher Copyright: © 2018 IEEE.; 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018 ; Conference date: 15-04-2018 Through 20-04-2018.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021. URL https://arxiv.org/abs/2102.09672.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. URL http://arxiv.org/abs/1503.03585.

Valentini-Botinhao, C. Noisy speech database for training speech enhancement algorithms and tts models, 2016 [sound], 2017. URL https://doi.org/10.7488/ds/2117. Accessed: Month Day, Year.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.