A Mamba-Based Model for Automatic Chord Recognition

Chunyu Yuan, Jiyeoung Sim, Johanna Devaney

[1] *CUNY Graudate Center, Brooklyn College*

Abstract. We present a Mamba-based model named BMACE (Bidirectional Mamba-based network, for Automatic Chord Estimation), which utilizes selective structured state-space models in a bidirectional Mamba layer to effectively model temporal dependencies. Our model achieves high prediction performance comparable to state-of-the-art models, with the advantage of requiring fewer parameters and lower computational resources than existing models.

1 Introduction

Inspired by the bidirectional Transformer, we propose a lightweight bidirectional Mamba-based network specifically designed for chord estimation/recognition: BMACE (Bidirectional Mamba-based network for Automatic Chord Estimation). The Mamba architecture was first introduced in late 2023 [1] and has been gaining rapid momentum since its release. It has been applied to some speech [2–4] and some MIR [5, 6] tasks, but not yet for ACE. Mamba distinguishes itself from other models by eschewing the usual attention and MLP blocks for a more streamlined approach. This results in a model that is not only lighter and faster but also uniquely capable of scaling linearly with sequence length, an achievement that sets it apart from its predecessors. Central to Mamba's design are its Selective-State-Spaces (SSM): these are recurrent models that selectively process information based on the current input, effectively filtering out irrelevant data to focus on what is most critical for efficient processing. Additionally, Mamba simplifies its architecture by replacing the complex attention and MLP blocks in Transformers with a single, unified SSM block, enhancing inference speed and reducing computational load. Mamba incorporates hardware-aware parallelism, using a specially designed parallel algorithm that optimizes recurrent operations for improved hardware efficiency, potentially boosting performance even further.

2 Model Architecture

Figure 1 presents the structure of our bidirectional Mamba network. Bidirectional Mamba blocks and fully-connected layers are the main modules in the network. It processes a 10-second audio signal as a Constant Q Transform (CQT) feature. The model integrates a fully-connected layer into the input, which then proceeds to two Mamba blocks with opposite masking directions, represented as dotted boxes in Figure **??**. The outputs from these blocks are concatenated and passed through a fully-connected layer to maintain the input's original dimensions. We added residual operation in the blocks and layers to increase the information entropy.

Our model is implemented with Pytorch[7] framework and trained the instance node at Lambda that has a single NVIDIA RTX A6000 GPU (24 GB), 14vCPUs, 46 GiB RAM and 512 GiB SSD. Our model was trained and validated on the MARL annotations of the uspop2002 dataset[8]. Each 10-second audio signal was processed with a 5-second overlap between consecutive signals. The signals were sampled at 22,050 Hz and analyzed using a Constant Q Transform (CQT) that covered 6 octaves starting from C1, with 24 bins per octave and a hop size of 2048. The CQT features were then converted to log amplitude using the formula $S_{\log} = \ln(S + \epsilon)$, where S represents the CQT feature, and ϵ is an extremely small number. This was followed by the application of global z-normalization, using the mean and variance derived from the training data.

References

[1] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[2] X. Jiang, C. Han, and N. Mesgarani, "Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation," *arXiv preprint arXiv:2403.18257*, 2024.

[3] K. Li and G. Chen, "Spmamba: State-space model is all you need in speech separation," *arXiv preprint arXiv:2404.02063*, 2024.

[4] C. Quan and X. Li, "Multichannel long-term streaming neural speech enhancement for static and moving speakers," *arXiv preprint arXiv:2403.07675*, 2024.

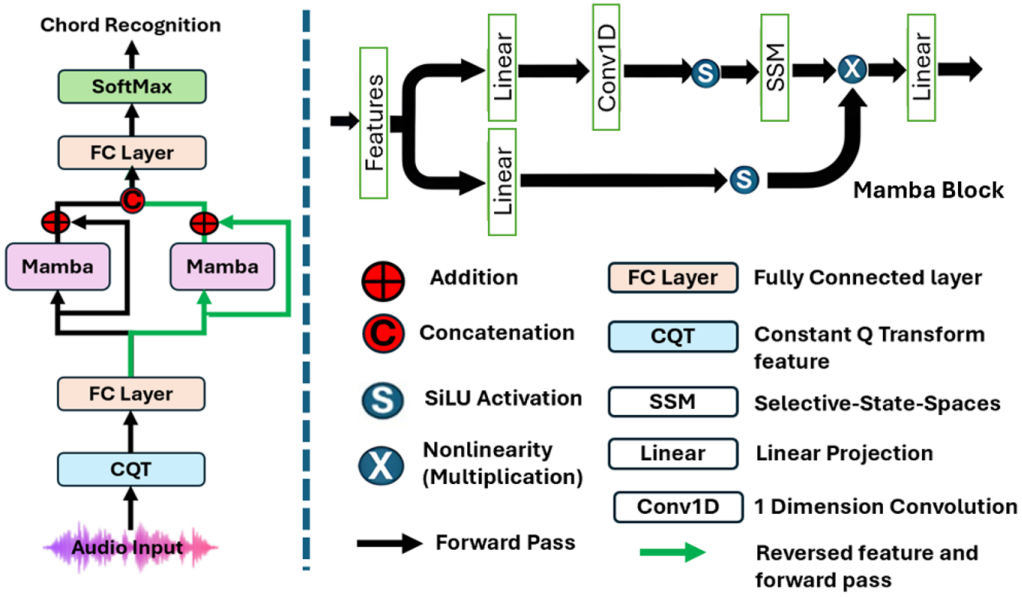

Figure 1. Architecture of BMACE. The model utilizes the Mamba block for improved feature processing. The Mamba block employs selective-state-spaces (SSM), SiLU activation, and 1D convolutions (Conv1D) for feature transformation. The figure highlights forward pass operations, feature reversal, addition, and concatenation mechanisms in the respective models, with fully connected (FC) layers leading to SoftMax output for chord recognition.

[5] J. Bai, Y. Fang, J. Wang, and X. Zhang, "A two-stage band-split mamba-2 network for music separation," *arXiv preprint arXiv:2409.06245*, 2024.

[6] J. Chen, T. Xie, X. Tang, J. Wang, W. Dong, and B. Shi, "Musicmamba: A dual-feature modeling approach for generating chinese traditional music with modal precision," *arXiv preprint arXiv:2409.02421*, 2024.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[8] A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, pp. 63–76, 2004.