An Exploration of Mamba for Speech Self-Supervised Models

Tzu-Quan Lin[1,2†] **Heng-Cheng Kuo**[1,3†] **Tzu-Chieh Wei**[1‡] **Hsi-Chun Cheng**[1,2‡]
Chun Wei Chen[1‡] **Hsien-Fu Hsiao**[1‡] **Yu Tsao**[3] **Hung-yi Lee**[1,2,4]
[1]National Taiwan University, Taiwan
[2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica
[4]NTU Artificial Intelligence Center of Research Excellence (NTU AI-CoRE)
tzuquanlin@gmail.com r11946023@ntu.edu.tw

Abstract

While Mamba has demonstrated strong performance in language modeling, its potential as a speech self-supervised learning (SSL) model remains underexplored, with prior studies limited to isolated tasks. To address this, we explore Mamba-based HuBERT models as alternatives to Transformer-based SSL architectures. Leveraging the linear-time Selective State Space, these models enable fine-tuning on long-context ASR with significantly lower compute. Moreover, they show superior performance when fine-tuned for streaming ASR. Beyond fine-tuning, these models show competitive performance on SUPERB probing benchmarks, particularly in causal settings. Our analysis shows that they yield higher-quality quantized representations and capture speaker-related features more distinctly than Transformer-based models. These findings highlight Mamba-based SSL as a promising and complementary direction for long-sequence modeling, real-time speech modeling, and speech unit extraction. The codebase is available at https://github.com/hckuo145/Mamba-based-HuBERT.

1 Introduction

In recent years, Transformer-based models and their multi-head self-attention mechanisms have achieved remarkable success across various domains (Vaswani et al., 2017; OpenAI et al., 2024; Devlin et al., 2019; Dosovitskiy et al., 2021; Radford et al., 2022). However, their quadratic computational complexity with respect to sequence length results in high deployment costs. As an alternative, Mamba adopts a Selective State Space architecture that retains content selection capabilities while reducing computational complexity to linear (Gu and Dao, 2023; Zhu et al., 2024; Huang et al., 2025; Lenz et al., 2025). In language modeling tasks, Mamba not only outperforms Transformers of similar scale but also rivals models with twice the number of parameters. Nevertheless, its application in the speech domain has yet to achieve comparable success. Most prior studies evaluate Mamba on a single downstream task, where it usually lags behind Transformer variants (Jiang et al., 2025b,a; Gao and Chen, 2024); those that insert auxiliary blocks can regain accuracy, but probably by forfeiting Mamba's hallmark linear-time scaling (Li et al., 2024)—and none provides a unified, cross-task evaluation. These limitations motivate a systematic exploration of Mamba-based SSL models that can be pretrained on large unlabelled speech and rapidly adapted, through lightweight fine-tuning, to a diverse set of tasks.

In this paper, we draw inspiration from the self-supervised training procedure of HuBERT (Hsu et al., 2021) and systematically train Mamba-based self-supervised speech models, termed Mamba-based HuBERT. The study begins with an in-depth analysis of their characteristics in automatic speech recognition (ASR) tasks, validating two key advantages: faster inference and linear scaling in sequence length. We quantify the MACs per second and real-time factor (RTF) of Mamba-based and Causal Transformer-based HuBERT models as input lengths increase from 5 seconds to 5 minutes. The results show that Mamba's computational cost remains nearly constant regardless of sequence length, while the computation and memory demands of the Causal Transformer grow rapidly, leading to out-of-memory (OOM) errors beyond 80 seconds.

These findings confirm that the Mamba architecture enables speech SSL models to handle long-context audio effectively, offering a more efficient alternative to Transformers. Therefore, we further fine-tune an External Bidirectional Mamba (ExtBiMamba) model (Zhang et al., 2024) for long-

[†‡]Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10337–10350

context ASR, where entire speeches are processed without sentence segmentation. In this setting, the word error rate (WER) is reduced from 13.37% to 11.08%, while the Transformer model of the same size fails to run due to memory limitations.

Beyond long-context modeling, Mamba's inherent causal nature demonstrates potential for streaming ASR. Under the constraint of using only past information, a 78M parameter Mamba-based HuBERT achieves a WER of 15.77%, outperforming a 94M parameter Causal Transformer-based HuBERT (16.66%), achieving better accuracy with fewer parameters.

After validating its potential in ASR tasks, we further investigate the generalizability of Mamba-based HuBERT models. Using phone purity (Hsu et al., 2021) and canonical correlation analysis (CCA) (Hotelling, 1936), we analyze the learned representations and find that Mamba-based HuBERT captures phonetic and speaker-related features more distinctly than its Transformer-based counterpart. We also evaluate four representative downstream tasks from the SUPERB benchmark (Yang et al., 2021). In the causal setting, Mamba-based HuBERT surpasses Transformer-based HuBERT in phonetic and speaker-related tasks, while slightly lagging in others, though still achieving higher overall performance. In the bidirectional setting, the small-size ExtBiMamba-based HuBERT outperforms Transformer-based HuBERT on nearly all tasks, while the base-size variant underperforms across the board, indicating room for improvement in scalability.

In conclusion, our contributions are as follows:

- We find that Mamba's inherent causal architecture makes it particularly well-suited for building causal speech SSL models, outperforming its Transformer-based counterpart
- We show that Mamba-based HuBERT models offer advantages when fine-tuned for long-context ASR and streaming ASR.
- We find that Mamba-based HuBERT models produce quantized representations with higher phonetic quality, which is beneficial for spoken language models that take SSL units as input.
- To the best of our knowledge, this is the first comprehensive exploration of Mamba-based HuBERT models as both speech foundation models and feature extractors.

2 Related Works

2.1 Speech Representation Learning and SUPERB

Over the past few years, self-supervised learning (SSL) has been adopted for speech representation, pretraining models on large amounts of unlabeled audio to extract reusable latent knowledge. Representative methods include wav2vec 2.0 (Baevski et al., 2020), which uses masking and contrastive learning and approaches fully supervised ASR performance with only minutes of labeled data, and HuBERT, which employs k-means pseudo-labels to further improve recognition and generation tasks. These frameworks greatly reduce the need for manual annotation and enable a single model to be reused across speaker identification, emotion recognition, and other downstream tasks.

SUPERB (Yang et al., 2021) is designed to systematically evaluate the generality and reusability of such pretrained models. By freezing the pretrained backbone and fine-tuning only lightweight task heads, researchers can evaluate ASR, intent classification, speaker verification, emotion recognition, and more within a unified pipeline, thus focusing on representation quality while lowering experimental overhead.

2.2 Mamba

Mamba is a state space model (SSMs) whose discrete-time formulas are expressed as follow:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad y_t = Ch_t \tag{1}$$

where $h(t)$ is the state vector, $\overline{A}$ is the state transition matrix, $\overline{B}$ regulates the interaction between input and state, and C maps the state to the output.

Since $\overline{A}$ and $\overline{B}$ are discrete-time parameters obtained from an underlying continuous system, they are not learned directly through back-propagation. Instead, they are typically approximated via **Zero-Order Hold (ZOH)** as follows:

$$\begin{aligned} \overline{A} &= \exp(\Delta A), \\ \overline{B} &= (\Delta A)^{-1}(exp(\Delta A) - I) \cdot \Delta B \end{aligned} \tag{2}$$

where A and B are the continuous-time equivalents of $\overline{A}$ and $\overline{B}$, and Δ represents the discretization step. In practice, we parameterize and train the continuous matrices A and B. Then, at each forward pass, they are converted to their discrete forms $\overline{A}$ and $\overline{B}$ via ZOH. This transformation allows SSMs to be efficiently applied in discrete-time

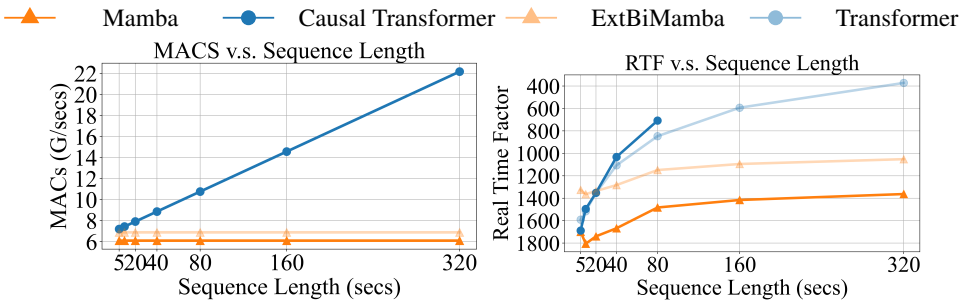

Figure 1: MACs (G/sec) and Real-Time Factor (RTF) of different HuBERT models at varying sequence lengths.

tasks while preserving their capability to capture long-range dependencies.

To enhance the context-aware capability of SSMs, Gu et al. introduced a **selection mechanism**, which dynamically adjusts system parameters based on the input signal:

$$B = f_B(x), \quad C = f_C(x), \\ \Delta = \text{Broadcast}_D(f_\Delta(x)) \tag{3}$$

where f_B, f_C, and f_Δ are parameterized linear projection. Specifically, $f_\Delta(x)$ is a one-dimensional linear projection, and Broadcast_D replicate this value D times to match the shape of h_t. This selection mechanism allows the model to flexibly adapt its state transition and output mapping according to the current input, improving its adaptability for long-sequence modeling.

2.3 Application of Mamba

Mamba has been applied to various speech-related tasks. In speech separation and enhancement (Jiang et al., 2025b,a; Li et al., 2024; Chao et al., 2024), Mamba has been integrated into U-Net architectures to leverage its linear-time complexity for long-sequence modeling while maintaining low computational cost. However, many of these works report only comparable or even slightly worse performance than original Transformer-based baselines, and often require hybrid designs involving both Mamba and Transformer components.

Notably, in causal settings, Mamba-based HuBERT models often achieve the best results among all tested architectures. Mamba has also shown strong potential in streaming ASR (Fang and Li, 2025), where its linear computational complexity is well-suited for low-latency applications. By incorporating mechanisms such as lookahead, Unimodal Aggregation (UMA), and Early Termination (ET), Mamba-based streaming models can significantly reduce latency compared to conventional causal Transformer and Conformer models, without sacrificing recognition accuracy.

In the context of self-supervised learning, Zhang et al. (Zhang et al., 2025) utilize Mamba-based SSL models as a tool to analyze Mamba's behavior in other speech processing tasks. In contrast, our work investigates the potential of Mamba-based SSL models as a speech foundation model and feature extractor, with comprehensive studies on its fine-tuning performance, probing results, and representational properties.

Moreover, Mamba has been employed in the Self-Supervised Audio Mamba (SSAM) framework to learn general-purpose audio representations (Yadav and Tan, 2024). SSAM takes advantage of Mamba's efficient state-space modeling to capture long-range dependencies and produce robust, scalable audio embeddings with both low compute cost and strong contextual modeling. However, this work focuses primarily on general audio and does not include experiments specific to speech.

3 Experiments

To construct our models, we replace the Transformer blocks in HuBERT with Mamba blocks, while preserving the feature extraction modules. Specifically, the input waveform is first processed by the standard 7-layer CNN feature encoder (with a 25 ms receptive field and 20 ms frame shift) followed by a convolution-based positional encoder. The Transformer blocks follow FAIR's original PyTorch implementation (Paszke et al., 2019), while the Mamba blocks adopt the implementation provided by Gu & Dao (Gu and Dao, 2023). By default, the Mamba blocks do not include a feedforward MLP module. If an MLP is added after the Mamba layer, we explicitly denote the model as Mamba+MLP.

We follow the HuBERT Base training

pipeline (Hsu et al., 2021) to train our Mamba-based HuBERT models. In the first iteration, MFCC features are used as targets, and the model is trained for 250k steps. In the second iteration, we restart training from scratch using the sixth-layer output from the first iteration as targets, and train for another 400k steps. All pre-training hyperparameters, except for the batch size, follow the default settings of HuBERT. Specifically, models are pre-trained on the full 960-hour LibriSpeech dataset. We use the Adam optimizer with a linear warm-up for the first 8% of training updates, followed by a linear decay scheduler. To ensure stability, the peak learning rate is set to 5e-5 for BiMamba Base models , and 5e-4 for all other configurations. The reported results are based on a single training run, without averaging across multiple trials. Due to limited computational resources, we train on a single NVIDIA V100 GPU and adopt a per-GPU batch size that is eight times larger than that used in the original HuBERT paper, which employed 32 GPUs. As a result, the total audio duration seen per batch is approximately one-fourth of the original setting. For a fair comparison, we apply the same training setup to Transformer-based HuBERT models trained from scratch.

We evaluate models under two settings: causal and bidirectional. The causal setting includes Mamba and causal Transformer models, where the latter uses a standard MHSA mechanism with a lower-triangular mask applied to the attention map. For the bidirectional setting, our experiments include External Bidirectional Mamba (ExtBiMamba), Inner Bidirectional Mamba (InnBiMamba), and the standard Transformer. The definitions of ExtBiMamba and InnBiMamba follow (Zhang et al., 2024). Both BiMambas adopt a two-branch design: one branch processes the sequence in its original order, while the other processes a time-reversed version; the backward output is reversed again and summed with the forward output. ExtBiMamba realizes these branches as two fully independent Mamba encoder layers with distinct input and output projections, whereas InnBiMamba shares these projections and only duplicates the convolutional and SSM modules. For clarity, we provide an illustration of these differences in Appendix A. Based on prior work showing stronger performance, ExtBiMamba is adopted as the default Mamba variant. We further perform an ablation study to evaluate the impact of using InnBiMamba.

4 Fine-tuning Results on ASR

We begin by evaluating the performance of fine-tuning Mamba-based HuBERT models for automatic speech recognition (ASR) in both bidirectional and causal settings. All fine-tuning setups follow wav2vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). Before presenting the results, we first highlight the computational efficiency of Mamba-based HuBERT models when handling long sequences.

4.1 Computational Efficiency Across Sequence Lengths

Mamba-based HuBERT models leverage the linear-time complexity of State Space Models with respect to sequence length, offering a computationally efficient alternative to Transformer-based approaches, whose self-attention mechanism incurs quadratic complexity.

To validate this advantage, we measure the Multiply-Accumulate operations (MACs) and Real-Time Factor (RTF) (Feng et al., 2023; Lin et al., 2022) across a range of sequence lengths: 5, 10, 20, 40, 80, 160, and 320 seconds. Notably, the reported MACs are measured in **MACs/second**, ensuring that the values directly reflect computational efficiency relative to input duration. Each RTF value is averaged over 10 runs to ensure stability, and all RTF measurements are conducted on a single NVIDIA RTX A6000 48 GB GPU.

Figure 1 (left panel) shows that the MACs for Mamba-based HuBERT models remain nearly constant across all sequence lengths. In contrast, the MACs for Transformer-based HuBERT models increase sharply with longer sequences, highlighting the computational overhead of attention mechanisms.

A similar trend is observed for RTF (Figure 1, right panel). Although all models exhibit increasing RTF with longer sequences, Mamba-based HuBERT models consistently maintain lower RTF values, particularly for long sequences. Notably, Causal Transformer begins to encounter out-of-memory (OOM) errors beyond a sequence length of 80 seconds, due to the additional computation required by the causal attention mask. This result underscores the superior efficiency of Mamba-based HuBERT models in long-context processing, making it a strong candidate for fine-tuning tasks that

require extensive contextual information, such as long-context ASR.

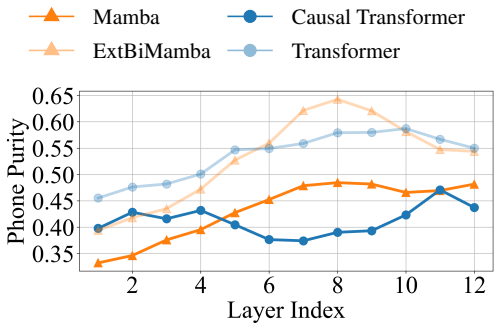

Figure 2: Layer-wise phone purity of HuBERT models.

Table 1: Application of fine-tuning long-context ASR model from Mamba-based HuBERT models, evaluated in terms of WER%(↓). Both models contain approximately 94.7M parameters.

Model	Utterance level	Document level
Transformer Base	11.86	OOM
ExtBiMamba Base	13.37	**11.08**

4.2 Fine-tuning for Long-Context ASR

The efficiency of Mamba-based HuBERT models in handling long sequences makes them particularly well-suited for fine-tuning long-context ASR models, where entire documents are processed as single inputs—unlike conventional utterance-level ASR, which operates on short, segmented speech. To demonstrate this advantage, we fine-tune a long-context ASR model based on bidirectional Mamba-based HuBERT models.

We conduct experiments on the TEDLIUM3 dataset (Hernandez et al., 2018), using its train split for fine-tuning, dev split for validation, and test split for evaluation, with talks longer than 20 minutes removed from all sets. A 12-layer convolutional encoder is appended to the SSL models, and the entire model is jointly fine-tuned using the CTC loss.

Table 1 presents the results of fine-tuning different models. At the utterance level, fine-tuning ExtBiMamba yields a WER of 13.37%, which is higher than the 11.86% achieved by fine-tuning Transformer. However, when ExtBiMamba is fine-tuned and evaluated on full documents, its WER significantly improves to 11.08%, demonstrating the benefits of leveraging broader context. Notably, Transformer-based HuBERT models fail to process document-length inputs due to out-of-memory (OOM) errors, further underscoring Mamba's practical advantage.

To further understand the benefits of long-context ASR, we performed paired t-tests comparing WER across different experimental settings. For ExtBiMamba, document-level outperformed utterance-level with p-value = 0.001. These results demonstrate that Mamba's performance improves significantly when leveraging full document context and confirm statistically significant differences between the two approaches. Qualitatively, long-context ASR achieves more consistent recognition of rare or challenging words, particularly those that appear multiple times across a document. In contrast, utterance-level ASR often produces inconsistent transcriptions for the same term, highlighting the value of global context in long-context processing.

Table 2: Application of fine-tuning causal ASR model from Mamba-based HuBERT models.

Model	Parameters M	ASR WER%↓
Causal Trans. Base	94.7	16.66
Mamba Base	78.2	**15.77**

4.3 Fine-tuning for Causal ASR

We further examine another scenario of fine-tuning an ASR model: causal ASR, where the causal speech SSL models are fine-tuned to predict the current token using only past information. For this task, we use the LibriSpeech 100-hour set for training and the test-clean set for evaluation. This setting is critical for streaming ASR, where real-time transcription is required.

Table 2 presents the results of fine-tuning SSL on causal ASR, where no lookahead is employed for all models. Mamba Base (78.2M parameters) achieves a word error rate (WER) of 15.77%, outperforming the Causal Transformer Base (94.7M parameters) with a WER of 16.66%. This result is notable because Mamba achieves superior ASR performance with approximately 17% fewer parameters.

5 In-depth Analysis of Learned Representation

Following the exploration of fine-tuning, this section delves deeper into the characteristics of the learned representations.

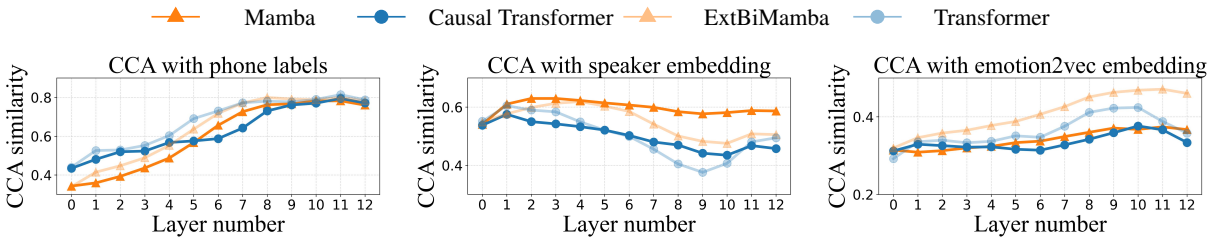

Figure 3: Layer-wise analysis results for different HuBERT models. Each plot below shows the CCA similarity to different label types: phone labels, speaker embedding, and emotion embedding.

5.1 Phone Purity of Quantized Representations

The quality of quantized SSL representations is crucial for various applications, for example, the input token of spoken language modeling (Lakhotia et al., 2021; Hassid et al., 2023; Arora et al., 2025). To evaluate the quality of quantized representations, we extract features from different layers of the SSL models, perform k-means clustering, and then calculate the phone purity of these clusters with respect to ground-truth phone labels. The phone purity is computed following the definition provided in the HuBERT paper (Hsu et al., 2021), and can be regarded as a measure of how well the quantized representations correlate with underlying phonemes.

Figure 2 illustrates the phone purity across different layers for Mamba-based and Transformer-based HuBERT models. Except for Causal Transformer, phone purity typically increases from the initial layers, reaches a peak in the mid-to-late layers, and then may slightly decline towards the final layers.

In the causal setting, Mamba consistently demonstrates higher peak phone purity than the Causal Transformer, possibly indicating better encoding of phonetic information. In the bidirectional setting, an interesting phenomenon emerges: while ExtBiMamba performs worse than the Transformer in utterance-level ASR fine-tuning, its peak phone purity slightly surpasses that of the Transformer in certain mid-to-late layers. This suggests that the quality of quantized representations does not necessarily align with fine-tuning performance.

We report results with $k = 100$ for the k-means clustering here, while additional results with $k = 500$ and $k = 1000$ are provided in Appendix B, showing overall trends consistent with the $k = 100$ case.

5.2 Similarity of Representations with Speech Attributes

Layer-wise analysis (Pasad et al., 2021; Lin et al., 2024a; Pasad et al., 2024; Lin et al., 2024b, 2025a) is a common approach to examine how speech attributes are distributed across different layers of speech SSL models. To better understand the property of the learned representations, we use Canonical Correlation Analysis (CCA) (Hotelling, 1936) to assess the similarity between representations across different layers and various speech attributes, including phoneme labels, speaker embeddings, and emotion embeddings. For phoneme labels, we convert them into one-hot vectors and compute CCA with mean-pooled phone-level representations. For speaker and emotion embeddings, we compute CCA using mean-pooled representations over the time axis.

Figure 3 presents the CCA similarity scores. For CCA with phone labels (left panel), all models show a clear trend where similarity increases with layer depth, peaking at the final or near-final layers. Interestingly, while achieving similarly high peak similarity, Mamba-based HuBERT models exhibit a steeper increase across layers, starting from lower similarity in early layers.

For CCA with Resemblyzer speaker embeddings (Wan et al., 2018) (middle panel), all models exhibit higher similarity in the early layers. Notably, Mamba-based HuBERT models generally maintain higher similarity with speaker embeddings than Transformer-based models, indicating that they capture speaker-related features more distinctly. We have also computed similarity with speaker embeddings from TitaNet (Koluguri et al., 2022) and ECAPA-TDNN (Desplanques et al., 2020); the trends are mostly similar to those observed with Resemblyzer.

For CCA with emotion2vec embeddings (Ma et al., 2024) (right panel), CCA similarity increases towards the mid-to-late layers for all models.

Table 3: SUPERB probing results for causal speech self-supervised models.

Size	Model	Parameters M	PR PER%↓	SID ACC%↑	ER ACC%↑	IC ACC%↑	SUPERB$_S$
Base	Causal Transformer	94.7	13.87	60.04	**63.33**	**94.23**	805.44
	Transformer + Causal Mask	94.7	25.57	**74.57**	58.77	80.07	734.68
	Mamba + MLP	94.7	11.72	73.48	61.72	89.62	**823.15**
	Mamba	78.2	**11.68**	73.07	61.98	87.45	817.94
Small	Causal Transformer	23.5	15.91	58.27	**62.07**	**87.92**	767.70
	Transformer + Causal Mask	23.5	24.05	66.61	60.45	81.86	735.49
	Mamba + MLP	33.0	**15.28**	**68.42**	61.63	82.39	**777.68**
	Mamba	23.5	17.34	66.06	60.82	85.13	766.94

ExtBiMamba and Transformer tend to show higher similarity scores compared to the causal models. This suggests that emotion-related information might be captured more effectively with access to bidirectional context. Compared to Transformer-based models, Mamba-based HuBERT models exhibit a steadily increasing CCA similarity across layers, without a noticeable decline in the final layers.

6 Downstream Performance Comparison

After evaluating the fine-tuning performance and representation characteristics, we further conduct downstream probing experiments on SUPERB (Yang et al., 2021). MiniSUPERB (Wang et al., 2023b) demonstrates that evaluating a subset of SUPERB tasks can sufficiently reflect the ranking observed in the full evaluation. Hence, we select four tasks for evaluation: phoneme recognition (PR), speaker identification (SID), emotion recognition (ER), and intent classification (IC). Additionally, we select several representative models for a full SUPERB evaluation to provide a more comprehensive comparison, with results presented in Appendix E. We follow the default evaluation settings of SUPERB. Following prior works (Feng et al., 2023; Shi et al., 2023a,b, 2024; Wang et al., 2023a; Lin et al., 2025b), we report the widely used overall SUPERB score (SUPERB$_S$) computed from these four tasks for easier comparison (see Appendix C for details on the score computation).

6.1 Probing Results in Causal Setting

Table 3 presents the performance in the causal setting. In the Base-size category, Mamba-based HuBERT models demonstrate strong performance. Mamba + MLP (94.7M parameters) achieves the highest SUPERB$_S$ score (823.15) and a competitive PR of 11.72%. The standard Mamba (78.2M parameters) also performs well, with a PR of 11.68% and a SUPERB$_S$ score of 817.94. Both Mamba variants outperform Causal Transformer (94.7M parameters), which records a PR of 13.87% and a SUPERB$_S$ score of 805.44. While Causal Transformer excels in ER (63.33%) and IC (94.23%), Mamba-based HuBERT models consistently achieve better phoneme recognition and overall representation quality.

For Small-size models, Mamba + MLP (33.0M parameters) achieves the highest SUPERB$_S$ score (777.68), along with strong PR (15.28%) and SID (68.42%). The standard Mamba (23.5M parameters) performs comparably to the Causal Transformer (23.5M parameters), with similar SUPERB$_S$ scores (766.94 vs. 767.70). This suggests that even at smaller scales, Mamba-based HuBERT models maintain competitive performance.

In addition, we include a naive baseline, Transformer + Causal Mask, which directly applies a causal mask to a Transformer trained in a bidirectional setting. However, this approach performs significantly worse than the Causal Transformer, highlighting the importance of pre-training with causal behavior.

The probing results on SUPERB are consistent with our observations on phone purity in Section 5.1, where Mamba's representations demonstrate superior phonetic information compared to those of causal Transformer.

6.2 Probing Results in Bidirectional Setting

Table 4 compares the performance in the bidirectional setting. For Base-size models, Transformer (94.7M parameters) consistently outperforms ExtBiMamba (94.3M parameters) across all metrics. The Transformer achieves a PR of 7.49% and a SUPERB$_S$ score of 868.93, while ExtBiMamba records a PR of 10.65% and a SUPERB$_S$ score of 815.38.

Table 4: SUPERB probing results for bidirectional speech self-supervised models.

Size	Model	Parameters M	PR PER%↓	SID ACC%↑	ER ACC%↑	IC ACC%↑	SUPERB$_S$
Base	Transformer	94.7	**7.49**	**75.77**	**62.36**	**97.44**	**868.93**
	ExtBiMamba	94.3	10.65	68.31	61.24	91.7	815.38
Small	Transformer	23.5	12.21	67.29	60.46	**91.83**	802.63
	ExtBiMamba	23.2	**11.38**	**69.22**	**62.34**	86.69	**809.18**

Table 5: Ablation on bidirectional Mamba architectures. The definitions of ExtBiMamba and InnBiMamba follow those in prior work (Zhang et al., 2024).

Size	Model	Parameters M	PR PER%↓	SID ACC%↑	ER ACC%↑	IC ACC%↑	SUPERB$_S$
Base	ExtBiMamba	94.3	10.65	68.31	61.24	91.7	815.38
	InnBiMamba + MLP	94.7	**9.00**	**70.67**	61.07	**91.56**	825.17
	InnBiMamba	82.9	9.62	70.79	**62.36**	91.43	**832.31**
Small	ExtBiMamba	23.2	**11.38**	**69.22**	**62.34**	86.69	**809.18**
	InnBiMamba	25.1	14.44	60.77	60.84	**87.00**	767.60

In contrast, the Small-size models exhibit a different trend. ExtBiMamba (23.2M parameters) surpasses the Transformer (23.5M parameters) in several metrics: PR (11.38% vs. 12.21%), SID (69.22% vs. 67.29%), ER (62.34% vs. 60.46%), and SUPERB$_S$ score (809.18 vs. 802.63). Although the Transformer retains an advantage in IC (91.83% vs. 86.69%), ExtBiMamba shows a clear edge in most tasks, indicating its effectiveness at smaller scales.

In summary, the probing performance of Mamba-based HuBERT models in the bidirectional setting is scale-dependent. ExtBiMamba underperforms the Transformer at the base scale but demonstrates a competitive advantage at the small scale. This might indicate that the scalability of Mamba in the bidirectional setting remains a challenge.

Interestingly, despite having higher phone purity as shown in Section 5.1, ExtBiMamba does not outperform Transformer in PR probing results. We attribute this to the fact that high phone purity reflects strong intra-class consistency (frames of the same phone falling into dominant clusters) but does not guarantee the inter-class discriminability required for a linear probe to separate different phonemes. Furthermore, the ranking of SSL models can be influenced by the architecture of downstream heads (Zaiem et al., 2023), and disentanglement of specific attributes does not always correlate with downstream performance (Plachouras et al., 2025).

6.3 Ablation on Bidirectional Mamba Architectures

In response to the initial results, we investigated alternative bidirectional Mamba designs, specifically comparing ExtBiMamba and InnBiMamba. The results are summarized in Table 5.

For Base-size models, InnBiMamba outperforms ExtBiMamba. Notably, InnBiMamba + MLP (94.7M parameters) achieves a PR of 9.00% and a SUPERB$_S$ score of 825.17, surpassing ExtBiMamba (94.3M parameters, PR 10.65%, SUPERB$_S$ 815.38). Even the standard InnBiMamba (82.9M parameters) performs exceptionally well, with a PR of 9.62% and the highest SUPERB$_S$ score of 832.31. It also excels in SID (70.79%) and ER (62.36%).

In contrast, for Small-size models, ExtBiMamba maintains an advantage. It achieves a PR of 11.38% and a SUPERB$_S$ score of 809.18 with 23.2M parameters, outperforming InnBiMamba (25.1M parameters, PR 14.44%, SUPERB$_S$ 767.60).

These results demonstrate that the choice of bidirectional Mamba architecture may significantly impact the performance of speech SSL model, and the optimal design may vary with model size. InnBiMamba shows strong potential at larger scales, while ExtBiMamba is more effective at smaller scales.

We hypothesize that the performance gap in ExtBiMamba Base stems from training instability, a phenomenon also reported in large-scale Mamba-based visual backbones (Suleman et al.,

2024; Shaker et al., 2025; Patro and Agneeswaran, 2024). Detailed analysis of our training dynamics supports this hypothesis (see Appendix D).

7 Conclusion

In this work, we conduct a comprehensive exploration of Mamba-based HuBERT models, including fine-tuning, representational analyses, and downstream probing. Our results demonstrate that Mamba-based HuBERT models offer advantages in long-context modeling and real-time recognition. Additionally, they can extract better quantized speech representations, which are beneficial for certain applications such as spoken language modeling. In downstream probing, Mamba-based HuBERT models generally perform better in the causal setting. However, we also observe that they suffer from limited scalability in the bidirectional setting. We believe this work provides empirical evidence and design guidance for the development of Mamba-based speech self-supervised models.

8 Limitations

This study is designed with an emphasis on reproducibility and interpretability, so we adopt relatively conservative and controlled settings for training scale and evaluation scope to clearly attribute model behavior and conclusions. Specifically, each pre-training model is trained on a single GPU. Although we maximize the amount of audio to fit within the memory limit of one GPU, the overall batch size remains smaller than in the original HuBERT setup. As supporting evidence, our Transformer-based HuBERT trained under this setting achieves lower SUPERB scores compared to the official FAIR-released model.

This choice makes the study reproducible and cost-efficient, though it may not fully exploit larger-scale regimes regarding model parameters, batch sizes, or diverse training corpora; future work can explore scaling laws and alternative curricula on cluster-level resources to verify if findings such as the bidirectional scalability challenges hold at larger scales.

Efficiency results are reported on a consistent single-GPU setup to ensure stable comparisons; while absolute numbers may vary with hardware or low-level optimizations (kernels, computation graphs), relative trends should remain valid. In long-document scenarios, standard Transformer baselines reached memory limits under our budget, reflecting the well-known quadratic scaling property; with memory optimizations, the absolute processable length could be extended for both models. These boundaries reflect the scope of our experimental design and, while they may limit certain aspects, they are not expected to alter the central conclusions of this study.

9 Acknowledgments

We thank the National Center for High-performance Computing (NCHC) of the National Institutes of Applied Research (NIAR) in Taiwan for providing computational and storage resources. Additionally, this work was supported by the Ministry of Education (MOE) of Taiwan under the project Taiwan Centers of Excellence in Artificial Intelligence, through the NTU Artificial Intelligence Center of Research Excellence (NTU AI-CoRE).

References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. 2024. An Investigation of Incorporating Mamba For Speech Enhancement. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 302–308.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *arXiv preprint arXiv:2005.07143*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Ying Fang and Xiaofei Li. 2025. Mamba for Streaming ASR Combined with Unimodal Aggregation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, and 1 others. 2023. Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103. IEEE.

Xiaoxue Gao and Nancy F. Chen. 2024. SpeechMamba: Long-Context Speech Recognition with Selective State Spaces Models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–8.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis CONNEAU, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually Pretrained Speech Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 63483–63501.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Proceedings of the 20th International Conference on Speech and Computer (SPECOM)*, pages 198–208. Springer.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hsiang-Wei Huang, Cheng-Yen Yang, Wenhao Chai, Zhongyu Jiang, and Jeng-Neng Hwang. 2025. MambaMOT: State-Space Model as Motion Predictor for Multi-Object Tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5.

Xilin Jiang, Cong Han, and Nima Mesgarani. 2025a. Dual-path Mamba: Short and Long-term Bidirectional Selective Structured State Space Models for Speech Separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xilin Jiang, Yinghao Aaron Li, Adrian Nicolas Florea, Cong Han, and Nima Mesgarani. 2025b. Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8102–8106. IEEE.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Barak Lenz and 1 others. 2025. Jamba: Hybrid Transformer-Mamba Language Models. In *The Thirteenth International Conference on Learning Representations*.

Kai Li, Guo Chen, Runxuan Yang, and Xiaolin Hu. 2024. SPMamba: State-space model is all you need in speech separation. *arXiv preprint arXiv:2404.02063*.

Tzu-Quan Lin, Hsi-Chun Cheng, Hung-yi Lee, and Hao Tang. 2025a. Identifying Speaker Information in Feed-Forward Layers of Self-Supervised Speech Transformers. In *APSIPA ASC 2025*.

Tzu-Quan Lin, Wei-Ping Huang, Hao Tang, and Hung-yi Lee. 2025b. Speech-FT: A Fine-tuning Strategy for Enhancing Speech Representation Models Without Compromising Generalization Ability. *arXiv preprint arXiv:2502.12672*.

Tzu-Quan Lin, Hung-yi Lee, and Hao Tang. 2024a. DAISY: Data Adaptive Self-Supervised Early Exit for Speech Representation Models. In *Interspeech 2024*, pages 4513–4517.

Tzu-Quan Lin, Guan-Ting Lin, Hung yi Lee, and Hao Tang. 2024b. Property Neurons in Self-Supervised Speech Transformers. In *Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 401–408. IEEE.

Tzu-Quan Lin, Tsung-Huan Yang, Chun-Yao Chang, Kuang-Ming Chen, Tzu-hsun Feng, Hung-yi Lee, and Hao Tang. 2022. Compressing transformer-based self-supervised models for speech processing. *arXiv preprint arXiv:2211.09949*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.

Badri N Patro and Vijay S Agneeswaran. 2024. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*.

Christos Plachouras, Julien Guinot, György Fazekas, Elio Quinton, Emmanouil Benetos, and Johan Pauwels. 2025. Towards a Unified Representation Evaluation Framework Beyond Downstream Tasks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Fahad Shahbaz Khan, and Rao Muhammad Anwer. 2025. Groupmamba: Efficient group-based visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023a. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Interspeech*, pages 884–888.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, and 1 others. 2023b. Findings of the 2023 ML-SUPERB challenge: Pre-training and evaluation over more languages and beyond. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. 2024. Multi-resolution HuBERT: Multi-resolution Speech Self-Supervised Learning with Masked Unit Prediction. In *The Twelfth International Conference on Learning Representations*.

Hamid Suleman, Syed Talal Wasim, Muzammal Naseer, and Juergen Gall. 2024. Stablemamba: Distillation-free scaling of large ssms for images and videos. *arXiv preprint arXiv:2409.11867*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Haoyu Wang, Siyuan Wang, Wei-Qiang Zhang, Hongbin Suo, and Yulong Wan. 2023a. Task- Agnostic Structured Pruning of Speech Representation Models. In *Interspeech*, page 231–235.

Yu-Hsiang Wang, Huang-Yu Chen, Kai-Wei Chang, Winston Hsu, and Hung-yi Lee. 2023b. MiniSUPERB: Lightweight benchmark for self-supervised speech models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Sarthak Yadav and Zheng-Hua Tan. 2024. Audio Mamba: Selective State Spaces for Self-Supervised Audio Representations. In *Interspeech*, pages 552–556.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, and 2 others. 2024. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik

Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech*, pages 1194–1198.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2023. Speech Self-Supervised Representation Benchmarking: Are We Doing it Right? In *Interspeech*, pages 2873–2877.

Xiangyu Zhang, Jianbo Ma, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2025. Rethinking Mamba in Speech Processing by Self-Supervised Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2024. Mamba in Speech: Towards an Alternative to Self-Attention. *arXiv preprint arXiv:2405.12609*.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *International Conference on Machine Learning (ICML)*.

A Architectural Differences between ExtBiMamba and InnBiMamba

Figure 4 and 5 illustrate the architectures of ExtBiMamba and InnBiMamba to clarify their differences.

B Additional Results for Phone Purity of Quantized Representations

Figure 6 shows the layer-wise phone purity of HuBERT models with k-means clustering at $k = 500$ and $k = 1000$, complementing the $k = 100$ results in Figure 2.

C Computation of overall SUPERB score

To provide a unified comparison across tasks, we use the overall SUPERB score (SUPERB$_S$) (Feng et al., 2023). Each task's metric is linearly scaled between the performance of baseline features (FBank) and the state-of-the-art (SOTA) results. This formulation implicitly accounts for task difficulty: when FBank already performs close to SOTA, even small improvements are emphasized more; whereas for tasks with a large gap between FBank and SOTA, the same improvement is considered less significant. For tasks with multiple metrics, scores are first averaged within the task, then across all tasks. The final result is multiplied by 1000 for readability. In this work, we follow the default SUPERB evaluation protocol and compute SUPERB$_S$ on four representative tasks: phoneme recognition (PR), speaker identification (SID), emotion recognition (ER), and slot filling (SF). The formulation is:

$$\Phi_{t,j}(f) = \frac{\phi_{t,j}(f) - \phi_{t,j}(\text{FBank})}{\phi_{t,j}(\text{SOTA}) - \phi_{t,j}(\text{FBank})} \quad (4)$$

$$\text{SUPERB}_S(f) = \frac{1000}{|T|} \sum_{t \in T} \frac{1}{|M_t|} \sum_{j \in M_t} \Phi_{t,j}(f) \quad (5)$$

Here, $\phi_{t,j}(f)$ denotes the score of model f on metric j of task t, T is the set of tasks, and M_t is the set of metrics for task t. The results of FBank and SOTA are taken from the SUPERB journal extension (Yang et al., 2024).

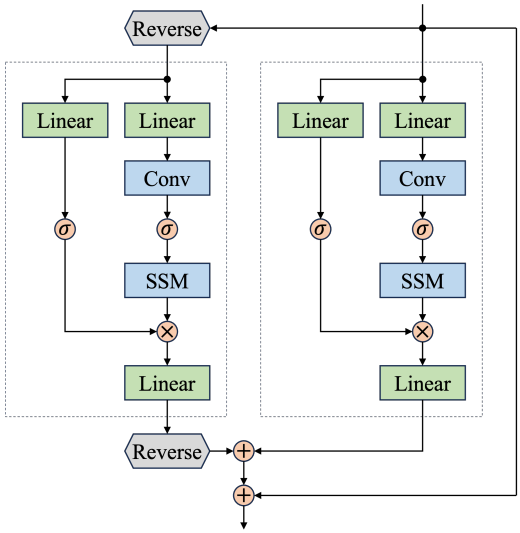

Figure 4: Architecture of ExtBiMamba, illustrating its independent forward and backward branches.

D Training Stability Analysis

We visualize the variation of loss scale at each training step for different Mamba variants in Figure 7. These variants—Mamba, InnBiMamba, and ExtBiMamba—represent a progression of increasing architectural complexity. In our training pipeline, the loss scale is halved whenever a gradient overflow occurs and doubled when no overflow is detected for several consecutive steps.

As shown in the figure, the vanilla Mamba model rarely overflows and stabilizes immediately. InnBiMamba overflows frequently during early training

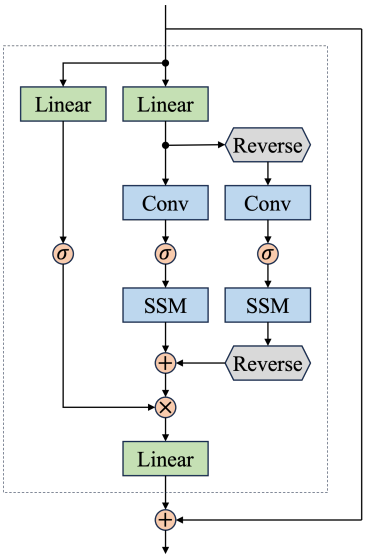

Figure 5: Architecture of InnBiMamba, where the forward and backward branches share projections but duplicate the convolutional layer and SSM module.

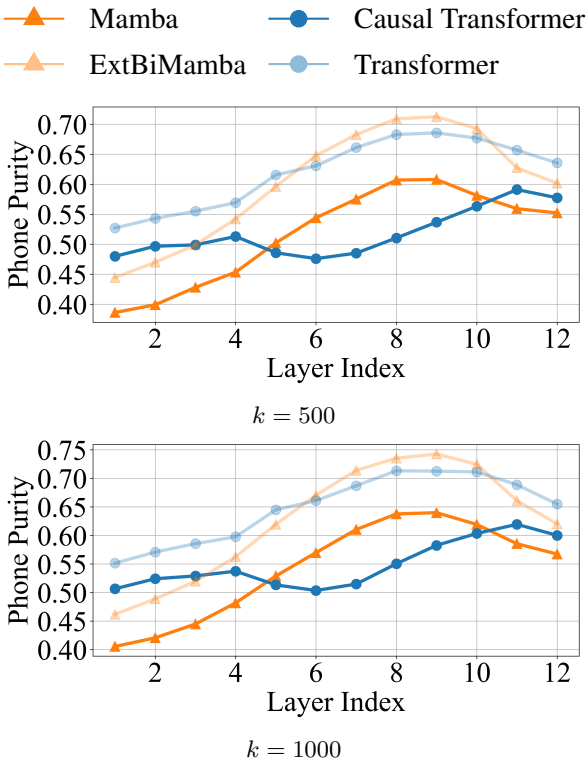

Figure 6: Layer-wise phone purity of HuBERT models with k-means clustering at $k = 500$ and $k = 1000$, as a supplement to the $k = 100$ results in Figure 2.

but gradually converges to stable loss scales. In contrast, ExtBiMamba, the most complex architecture among the three, exhibits persistent oscillations throughout pre-training, indicating continual instability.

This instability aligns with observations in prior work on Mamba-based visual backbones (Suleman et al., 2024; Shaker et al., 2025; Patro and Agneeswaran, 2024), where scaling up often leads to unstable training compared to Transformers. To ensure stability in larger-scale regimes with increased batch sizes or diverse corpora, future designs may require auxiliary mechanisms such as hybrid blocks that incorporate attention mechanisms or distillation strategies.

E Full SUPERB Evaluation

We provide the full SUPERB benchmark evaluation results for representative models to verify whether the trends observed in Section 6, based on the four-task subset ($SUPERB_S$), hold across the complete set of tasks. The detailed scores across all 10 tasks are summarized in Table 6. $\text{SUPERB}_{\text{FULL}}$ denotes an overall SUPERB score computed across the 10 tasks. Note that $\text{SUPERB}_{\text{FULL}}$ is not directly comparable to SUPERB_S reported elsewhere in the paper.

E.1 Causal Setting

The full evaluation reinforces our observation that Mamba-based architectures are well-suited for causal speech self-supervised modeling:

- Performance Advantage: Mamba+MLP Base achieves a significantly higher $\text{SUPERB}_{\text{FULL}}$ score of 651.6 compared to 608.1 for the Causal Transformer Base, demonstrating superior overall representation quality.
- Task-Specific Strengths: Mamba demonstrates clear superiority in Content-related tasks (PR, ASR, KS, QbE) and Speaker-related tasks (SID, SD).
- Limitations: Mamba exhibits a performance gap in Paralinguistic (ER) and Semantic tasks (IC, SF) compared to the Causal Transformer.

E.2 Bidirectional Setting

In the bidirectional setting, the results confirm that the Transformer remains the dominant architecture at the Base scale:

- Transformer Superiority: The standard Transformer Base outperforms ExtBiMamba Base across nearly all task categories, achieving a $\text{SUPERB}_{\text{FULL}}$ score of 772.4 versus ExtBiMamba's 683.59.

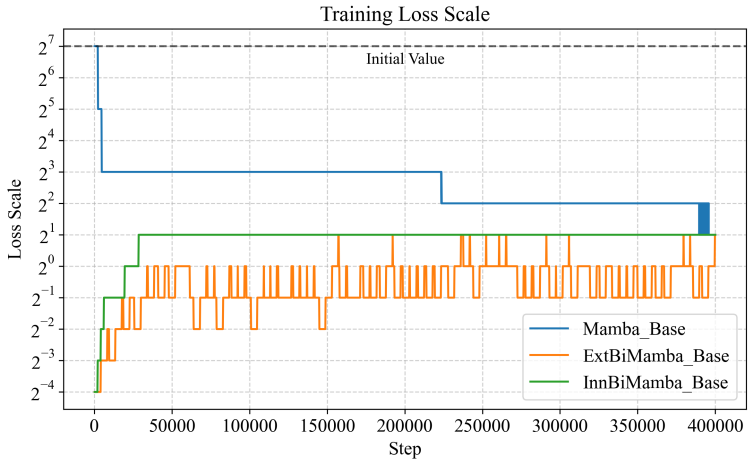

Figure 7: Training dynamics of the loss scale for different Mamba variants.

Table 6: Full SUPERB evaluation results for Causal Transformer Base, Mamba+MLP Base, Transformer Base, and ExtBiMamba Base. $\text{SUPERB}_{\text{FULL}}$ is computed across all 10 tasks. Note that $\text{SUPERB}_{\text{FULL}}$ is not directly comparable to SUPERB_S reported elsewhere in the paper.

Model	**PR**	**ASR**	**KS**	**QbE**	**SID**	**ASV**	**SD**	**ER**	**IC**	**SF**		$\text{SUPERB}_{\text{FULL}}$
	PER%↓	WER%↓	ACC%↑	MTWV↑	ACC%↑	EER%↓	DER%↓	ACC%↑	ACC%↑	F1↑	CER%↓	
Causal Trans. Base	13.87	13.55	95.13	0.0356	60.04	**7.66**	8.23	**63.33**	**94.23**	**81.91**	**35.55**	608.1
Mamba+MLP Base	**11.72**	**12.25**	**95.91**	**0.0590**	**73.48**	7.77	**7.36**	61.72	89.62	80.99	37.12	**651.6**
Trans. Base	**7.49**	**9.16**	95.01	**0.0792**	**75.77**	**5.91**	6.86	**62.36**	**97.44**	**87.18**	**26.38**	**772.4**
ExtBiMamba Base	10.65	11.41	**96.36**	0.0586	68.31	7.07	**6.76**	61.24	91.70	83.14	33.89	683.5

- Scalability Challenges: These results support our finding that while Mamba is a powerful alternative in causal settings, its scalability in bidirectional settings remains a challenge.