



## Assessing the Performance and Efficiency of Mamba ASR in Low-Resource Scenarios

Rodolfo Zevallos<sup>1</sup>, Martí Cortada-Garcia<sup>1</sup>, Sarah Solito<sup>1</sup>, Carlos Mena<sup>1</sup>, Alex Peiro-Lilja<sup>1,2</sup>, Javier Hernando<sup>1,3</sup>

<sup>1</sup>LangTech Lab, Barcelona Supercomputing Center, Spain; <sup>2</sup>Centre de Llenguatge i Computació, Universitat de Barcelona, Spain; <sup>3</sup>Universitat Politècnica de Catalunya (UPC), Spain

(rodolfo.zevallos, marti.cortada, sarah.solito, carlos.hernandez, alexandre.peiro)@bsc.es, javier.hernando@upc.edu

### Abstract

Mamba, a state space model-based architecture, is emerging as a strong alternative to Transformer models, showing equal or superior performance in sequence generation, including speech. However, analyses have focused mainly on high-resource scenarios. This paper explores Mamba's potential in ASR for low-resource scenarios. We compare the Transformer-based Conformer and its state-space counterpart, ConMamba, across nine languages with varying training data. Our results show that ConMamba achieves similar WER to Conformer for short-context inputs but significantly improves performance on long-context inputs, reducing WER by up to 50% on average. Additionally, ConMamba enhances efficiency, requiring 40–45% less training time, using 50% less memory, and accelerating inference by 63–70%, making it a more effective ASR solution across different data availability scenarios.

**Index Terms:** speech recognition, Mamba, low-resource scenarios

### 1. Introduction

Recent advancements in Automatic Speech Recognition (ASR) have been driven by models based on Transformers [1], demonstrating remarkable performance compared to other models. However, these advancements have largely relied on vast amounts of labeled data and substantial computational power, limiting their applicability to low-resource scenarios (LRSs) [2, 3] and devices with memory and computational power constraints.

To address these limitations, recent research has explored alternative approaches to improve ASR performance in LRS. One of the most significant advancements has been Wav2Vec 2.0 [4], which introduced self-supervised learning to reduce the need for manually transcribed data. OpenAI developed Whisper [5], a multilingual Transformer model trained on a vast audio corpus. Meta introduced SeamlessM4T [6], a multi-modal Transformer architecture that integrates ASR with text and speech translation across 101 languages.

Despite the achievements of Transformers in ASR, their self-attention mechanism poses efficiency challenges, particularly when handling long-context inputs, due to its quadratic complexity concerning input length. To address these limitations, in recent years alternative architectures have been explored, with Mamba standing out as a promising approach. Based on Selective State Space Models (SSMs) [7], Mamba represents a paradigm shift by replacing attention mechanisms with optimized memory structures and parametric convolutions, allowing for more efficient sequence modeling with linear computational complexity in multiple scenarios.

Prior research of Mamba in ASR has established its effectiveness in high-resource scenarios (HRSSs), yet its performance in LRS remains underexplored. To cover this gap, we assess the performance and efficiency of a Mamba-based ASR architecture over a selected group of languages in LRSs: Catalan, Basque, Galician, Spanish, Guarani, Quechua, Maltese, Icelandic, and Faroese. The selection of languages in this study is based on two key criteria: our available resources and the evaluation of ASR in low-data scenarios. We include Iberian and Latin American languages, which we already work with, alongside low-resource European languages such as Maltese, Icelandic, and Faroese, expanding linguistic diversity. Spanish serves as a central reference for most languages, except for these three, where no direct relationship exists. Although some languages, such as Spanish, Catalan, and Icelandic, have a lot of audio resources available, our study focuses on restricted-data conditions.

To ensure comparability, we limit experiments to 80, 12-18, and 5 hours of audio, except for Guarani, where data is even scarcer. This approach allows us to evaluate ASR not only in LRSs [4, 8] but also in constrained training conditions, reflecting real-world challenges in developing models for underrepresented languages.

In our study, we train Conformer [9] and ConMamba [10] separately on each language and dataset size, ensuring that both models share the same configuration for a fair comparison. We evaluate our models' performance in terms of Word Error Rate (WER) on short-context and long-context input. While short-context evaluation reflects performance on brief utterances, which are common in interactive systems like voice assistants, long-context evaluation is crucial for assessing how well the models handle extended speech, maintain contextual coherence, and mitigate error accumulation. This is particularly relevant for real-world applications such as lecture transcription, meeting recordings, or media subtitling, where preserving meaning over long durations is essential. Additionally, we analyze training and inference times, as well as memory usage, to provide a comprehensive comparison of their efficiency.

### 2. Related Work

Recent research with Mamba [11] has demonstrated that Mamba is competitive not only in ASR, but in text-to-speech synthesis, and speech summarization, standing out not only for its accuracy but also for its computational efficiency. Speech Slytherin [10] evaluated variants such as Mamba-TasNet for speech separation, ConMamba for speech recognition, and VALL-M for synthesis, concluding that Mamba models or Mamba-Transformer hybrids [12] offer performance comparable or even superior to Sepformer [13], Conformer [9], and

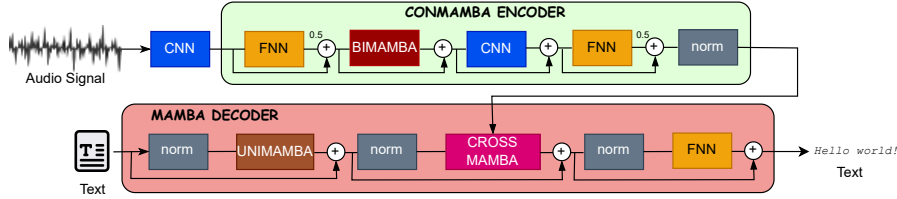


Figure 1: *The ConMamba architecture.*

VALL-E [14]. However, the most relevant aspect of these models is their lower memory consumption and higher processing speed, making them particularly suitable for ASR tasks on devices with limited resources.

More recently, Samba-ASR [15] introduced the Mamba architecture in both the encoder and decoder. Unlike Transformers, Samba-ASR utilizes structured state-space dynamics to more effectively model temporal dependencies, resulting in significant improvements across multiple standard benchmarks [16, 17]. In addition to achieving reductions in WER and character error rate, Samba-ASR excels in processing long-context speech while maintaining lower computational costs compared to traditional models. As seen in the existing related work on Mamba, specifically in ASR, has focused only on HRSs, such as English, Japanese and Mandarin [11, 10, 15].

### 3. Method

#### 3.1. Mamba

Recent advances in SSMs [7], particularly Mamba [18], offer a promising alternative with linear scaling, making them well-suited for speech recognition task [19], as they address the inherent quadratic complexity of Transformers in sequence length, enabling more efficient processing of long speech sequences. Mamba builds upon the Structured State Space Model (S4) [20], processing a one-dimensional input sequence  $x(t) \in \mathbb{R}$  and transforming it into an output sequence  $y(t) \in \mathbb{R}$  through an intermediate state  $h(t) \in \mathbb{R}$ . The model evolves according to the following equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{N \times N}$  is the evolution matrix,  $\mathbf{B} \in \mathbb{R}^{D \times N}$  is the input mapping matrix, and  $\mathbf{C} \in \mathbb{R}^{D \times N}$  is the output mapping matrix. Here,  $D$  represents the input and output dimensionality, while  $N$  defines the latent state size. Mamba employs Zero-Order Hold (ZOH) discretization to ensure stable and accurate state transitions while preserving resolution invariance and normalization.

Mamba is built around a structured state space model (SSM) enhanced with linear input-output projections and layer normalization. Bi-Mamba [21] extends this by identifying which weight matrices in Mamba-2 [22] can be binarized, focusing on linear modules while excluding the causal head to preserve token representation ability. By replacing standard linear layers with the FBI-Linear module [23], which includes a binarized weight matrix and two high-precision scale factors, Bi-Mamba achieves a 90% compression ratio while maintaining performance. Additionally, a bidirectional Mamba model [24] has been proposed, incorporating two parallel SSMs and causal convolutions—one for the original sequence and one for the

reversed sequence—allowing for more accurate and context-aware representations by leveraging both past and future context.

#### 3.2. ConMamba

For our experiments, we use ConMamba [10], an extension of the Mamba architecture that integrates convolutional modules within its encoding block, inspired by Conformer [9]. This design enhances the capture of local dependencies, improving the extraction of both global and local context features.

As shown in Figure 1, ConMamba consists of an encoder and a decoder, with the encoder block integrating three key components: a bidirectional Mamba module (BiMamba), a feedforward module, and a convolutional module. The processing follows a structured sequence where the input is first refined through a feedforward layer with a residual connection, followed by the application of BiMamba to capture global dependencies. A convolutional layer then enhances local feature extraction, and finally, layer normalization is applied along with an additional feedforward refinement to normalize the final projection [21]. This design allows ConMamba to effectively balance long-range dependencies and local feature learning, improving overall model performance.

Mamba lacks a native cross-attention mechanism for handling multiple variable-length inputs. To address this, CrossMamba, a unidirectional Mamba variant, concatenates key and query inputs while retaining only the latter half of the outputs to match the query length, effectively serving as a plug-in replacement for cross-attention. The CrossMamba decoder processes token representations through a structured sequence, beginning with layer normalization, followed by a unidirectional Mamba module (UniMamba) to model sequential dependencies. Another layer normalization step is applied before integrating contextual information from the encoder via the CrossMamba block. Finally, a feedforward module refines the output through additional normalization and processing, ensuring effective representation learning and cross-input integration.

## 4. Experiments

#### 4.1. Dataset

We evaluate our Mamba models for ASR using ten datasets: LibriSpeech (Spanish) [2], 3CatParla (Catalan) [25], Siminchik (Quechua) [26], OpenSLR (Basque & Galician) [27], AmericasNLP (Guarani) [28], Ravnursson (Faroese) [29], Samrómur (Icelandic) [30], and the Maltese split of Common Voice [31]. All audio was standardized to 16 kHz, 16-bit WAV. LibriSpeech-Spanish consists of 1,438.41 hours, while 3CatParla contains 218,532 speech recordings (731 hours). The latter, derived from Catalan TV broadcasts, is manually annotated and categorized by transcription quality, used for test and

Table 1: WER (%) results for ConMamba and Conformer models across different languages, training hours, and input speech duration.

Language	Train Hours	Short context				Long context			
		ConMamba (S)	Conformer (S)	ConMamba (L)	Conformer (L)	ConMamba (S)	Conformer (S)	ConMamba (L)	Conformer (L)
<b>Low-resource Scenario (80h training)</b>									
Spanish	80	<b>19.11</b>	19.83	<b>18.03</b>	19.14	<b>24.04</b>	49.21	<b>22.83</b>	48.17
Catalan	80	21.81	<b>21.15</b>	19.34	<b>19.31</b>	<b>26.10</b>	53.41	<b>23.95</b>	52.83
Quechua	80	<b>26.31</b>	27.02	<b>24.79</b>	25.01	<b>30.19</b>	55.89	<b>29.15</b>	54.73
Icelandic	80	<b>22.15</b>	23.04	<b>20.98</b>	21.77	<b>29.38</b>	55.37	<b>28.74</b>	56.81
Faroese	80	<b>23.89</b>	24.67	<b>22.15</b>	23.01	<b>31.02</b>	58.27	<b>29.86</b>	57.58
<b>Average</b>	-	<b>22.65</b>	23.14	<b>21.06</b>	21.65	<b>28.15</b>	54.43	<b>26.91</b>	54.02
<b>Low-resource Scenario (8-18h training)</b>									
Basque	12	<b>43.13</b>	43.16	42.06	<b>41.75</b>	<b>49.24</b>	67.48	<b>48.80</b>	66.69
Galician	8	58.10	<b>57.92</b>	<b>56.92</b>	56.98	<b>63.62</b>	75.17	<b>62.26</b>	73.93
Maltese	18	39.72	<b>39.13</b>	<b>39.12</b>	39.41	<b>45.55</b>	63.19	<b>44.29</b>	62.46
<b>Average</b>	-	48.98	<b>48.74</b>	<b>46.03</b>	46.05	<b>52.80</b>	68.61	<b>51.78</b>	67.69
<b>Extreme Low-resource Scenario (<math>\leq 5h</math> training)</b>									
Spanish	5	<b>58.17</b>	57.82	<b>57.10</b>	57.99	<b>62.79</b>	80.10	<b>61.37</b>	78.83
Catalan	5	58.92	<b>59.27</b>	58.01	<b>58.48</b>	<b>64.02</b>	81.78	<b>63.08</b>	79.90
Quechua	5	<b>65.36</b>	66.10	64.84	<b>64.56</b>	<b>70.94</b>	90.24	<b>70.02</b>	89.10
Basque	5	<b>65.48</b>	66.06	64.92	<b>64.83</b>	<b>70.57</b>	88.49	<b>70.11</b>	89.37
Galician	5	64.25	<b>65.04</b>	<b>63.81</b>	64.15	<b>69.05</b>	86.81	<b>62.26</b>	73.93
Maltese	5	62.42	<b>62.41</b>	<b>62.05</b>	62.18	<b>67.92</b>	89.17	<b>67.03</b>	88.25
Icelandic	5	<b>61.25</b>	61.31	<b>60.29</b>	60.41	<b>68.91</b>	86.81	<b>66.81</b>	86.11
Faroese	5	<b>64.87</b>	65.02	<b>64.15</b>	64.56	<b>70.06</b>	88.15	<b>69.18</b>	87.97
Guarani	0.3	<b>82.79</b>	82.84	<b>82.24</b>	82.27	<b>86.43</b>	98.91	<b>85.99</b>	98.31
<b>Average</b>	-	<b>64.83</b>	65.09	<b>64.16</b>	64.38	<b>70.08</b>	87.83	<b>68.43</b>	85.75

dev sets. For Quechua, Siminchi provides 100 hours of transcribed radio conversations. OpenSLR offers 14 and 10 hours for Basque and Galician, respectively. AmericaNLP has only 32 minutes for Guarani, exemplifying extreme data scarcity. Common Voice includes Maltese 18 hours and 9 validated, while Faroese and Icelandic provide 109 and 145 hours, respectively.

Finally, for long-context input evaluation, we adopt SpeechMamba’s approach [32], where we define short-context input as audio segments shorter or equal to 10 seconds and long-context input as those exceeding 10 seconds. To construct long-context datasets, we merge consecutive utterances from the same speaker into segments between 10–60 seconds.

## 4.2. Model Configuration & Hyperparameters

We use the recipes<sup>1</sup> from ConMamba and Conformer, incorporating advanced configurations for robust performance. We use ConMamba (S) and Conformer (S) with 144 dimensions and 12+4 layers, and ConMamba (L) and Conformer (L) with 512 dimensions and 12+6 layers. The audio token duration is set to 40 ms. For transcriptions, a BPE tokenizer [33] is trained for each language using SpeechBrain recipe<sup>2</sup>. Convergence is optimized with a global batch size determined by the product of the batch size (16), the number of GPUs, and the gradient accumulation factor (1), ensuring stability with a maximum gradient norm of 5.0. The models are trained for 110 epochs using AdamW, with an initial learning rate of 0.0008 and a Noam scheduler with 30,000 warm-up steps. Label smoothing (0.1) is applied to the KLDiv loss to improve generalization. Training is conducted on 2x Nvidia Hopper GPUs with 32GB HBM2, and evaluation is without an external language model.

<sup>1</sup><https://github.com/xi-j/Mamba-ASR>

<sup>2</sup><https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/Tokenizer>

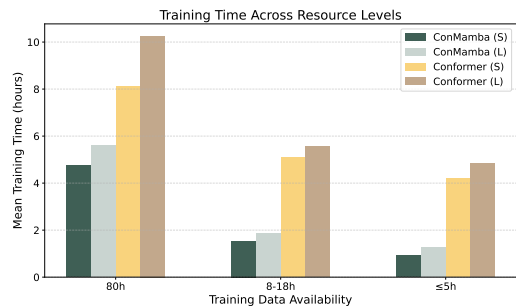


Figure 2: Comparison of training times between ConMamba and Conformer models in different languages and dataset size.

## 5. Results

### 5.1. Short Context

The results in Table 1 present the evaluation of ConMamba and Conformer using short-context audio. For languages with 80 hours of training, ConMamba (S) achieves an average WER of 22.65%, slightly outperforming Conformer (S) (23.14%,  $\downarrow 2.1\%$ ). The trend remains consistent across most languages, though differences are minimal. With 8-18 hours of training, both models perform similarly, with ConMamba (S) at 48.98% and Conformer (S) at 48.74% ( $\downarrow 0.5\%$  in favor of Conformer).

The differences remain marginal across languages. In the LRS ( $\leq 5h$ ), both models degrade, but ConMamba maintains a slightly lower WER (64.83% vs. 65.09%,  $\downarrow 0.4\%$ ). Performance varies by language, with neither model showing a consistent advantage. In extremely LRS, both models perform poorly. Overall, ConMamba slightly but not significantly outperforms Conformer in short-context scenarios (input length  $\leq 10$  seconds on average).

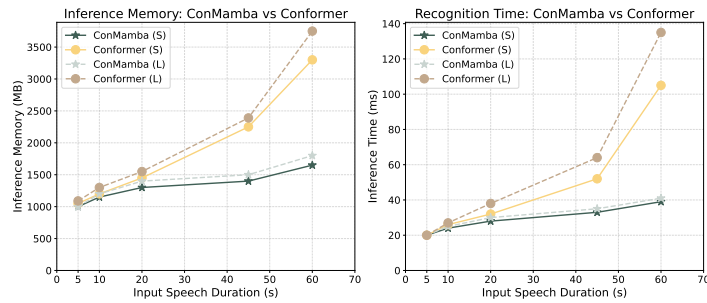


Figure 3: Average memory and speed comparisons between ConMamba and Conformer models across different languages.

## 5.2. Long Context

Table 1 shows the evaluation of ConMamba and Conformer using long-context audio, where longer-input speech affects performance in both models. With 80 hours of training, ConMamba (S) achieves a 28.15% WER, significantly outperforming Conformer (S) (54.43%,  $\downarrow 48.3\%$ ). Larger models follow the same trend, with ConMamba (L) at 26.91% vs. Conformer (L) at 54.02% ( $\downarrow 50.2\%$ ). With 8-18 hours of training, ConMamba (S) maintains a lower WER (52.80% vs. 68.61%,  $\downarrow 23.0\%$ ), and the larger models show similar gains (51.78% vs. 67.69%,  $\downarrow 23.5\%$ ). In LRS ( $\leq 5$ h), both models degrade, but ConMamba remains more robust (70.08% vs. 87.83%,  $\downarrow 20.2\%$ ). Even with extremely scarce data (0.3h), ConMamba is more stable (86.43% vs. 98.91%).

Overall, ConMamba reduces WER by up to 50% in long-context scenarios (input length  $> 10$  seconds), while Conformer degrades significantly as input length increases, making ConMamba the better choice.

## 5.3. Efficiency

We evaluate the training and inference efficiency of ConMamba and Conformer models in terms of memory usage and speed. For training, we measure the total time required for 110 epochs, while for inference, we assess processing time and memory usage across different input lengths. Figure 2 presents the training times (in hours) for ConMamba and Conformer in their Small (S) and Large (L) versions across different LRS categories.

The results consistently show that ConMamba trains faster than Conformer in all scenarios, demonstrating its efficiency in various data availability settings. In the 80-hour training category, ConMamba significantly reduces training time, requiring approximately 40% less time for (S) and 45% less for (L) compared to Conformer. This trend remains evident in the 8-18 hour training category, where ConMamba maintains a 35-45% reduction in training duration. Even in the extreme LRS ( $\leq 5$ h training), where total training time is lower, ConMamba continues to be more efficient, consistently requiring less time than Conformer.

Figure 3 shows that ConMamba (S) and (L) maintain more stable memory growth during inference compared to Conformer. These results, averaged across all models described in Table 1, show that as input duration increases, Conformer exhibits a sharper rise in memory usage. For the long-context input ( $> 10$ ), Conformer (L) reaches 3750 MB, while ConMamba (L) remains at 1800 MB, achieving a 52% reduction. Similarly, Conformer (S) requires 3300 MB, whereas ConMamba (S) remains at 1650 MB, reflecting a 50% reduction. While both models perform similarly on shorter inputs, the gap widens

significantly as input duration increases.

For the long-context input, Conformer (L) requires 135 ms, whereas ConMamba (L) completes inference in only 41 ms, marking a 70% improvement. A similar trend is observed in the small models, where ConMamba (S) processes long-context inputs 63% faster than Conformer (S).

To evaluate the efficiency of the models in terms of environmental impact<sup>3</sup>, we compared the  $CO_2$  emissions generated by ConMamba and Conformer in different training scenarios. ConMamba is consistently more efficient, with lower emissions across all categories. The most eco-friendly model, ConMamba (S) -  $\leq 5$ h, emits only 0.34 kg of  $CO_2$ , while Conformer (L) - 80h reaches 3.54 kg of  $CO_2$ .

## 6. Conclusion

This work is the first to present a comprehensive study on the application of Mamba in ASR for LRL, marking a significant milestone in expanding the applicability of SSMS beyond HRS.

Our results demonstrate that the Mamba-based architecture is not only competitive with traditional Transformer models but, in many cases, surpasses them in both accuracy and efficiency. In terms of WER, ConMamba achieves a significant reduction in error compared to Conformer, particularly in long-sequence scenarios. Additionally, its ability to model temporal dependencies more effectively than Transformers suggests that Mamba represents a more robust and adaptable approach to speech recognition in LRSs.

Beyond accuracy performance, Mamba’s computational efficiency is a key factor that makes it a more viable alternative for applications on devices with memory and computational power constraints. By encompassing diverse linguistic families and regions, our study lays the foundation for future multilingual ASR research. As the first to assess Mamba in ASR for LRSs, our work sets a new baseline and underscores the need to explore alternative architectures to further democratize speech recognition.

## 7. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU - NextGenerationEU within the frameworks of the project Modulos del Lenguaje and ILENIA, the latter project with reference 2022/TL22/00215337.

<sup>3</sup>[https://huggingface.co/docs/leaderboards/open\\_llm\\_leaderboard/emissions](https://huggingface.co/docs/leaderboards/open_llm_leaderboard/emissions)

## 8. References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
- [2] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Proc. INTERSPEECH 2020 – 21<sup>th</sup> Annual Conference of the International Speech Communication Association*, 2020.
- [3] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003.
- [4] A. Baeviski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.
- [5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.
- [6] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.
- [7] N. M. Cirone, A. Orvieto, B. Walker, C. Salvi, and T. Lyons, "Theoretical foundations of deep selective state-space models," *arXiv preprint arXiv:2402.19047*, 2024.
- [8] M. Elamin, M. Omer, Y. Chanie, and H. Ndlovu, "Creating spoken dialog systems in ultra-low resourced settings," *arXiv preprint arXiv:2312.06266*, 2023.
- [9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [10] X. Jiang, Y. A. Li, A. N. Florea, C. Han, and N. Mesgarani, "Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis," *arXiv preprint arXiv:2407.09732*, 2024.
- [11] K. Miyazaki, Y. Masuyama, and M. Murata, "Exploring the capability of mamba in speech applications," *arXiv preprint arXiv:2406.16808*, 2024.
- [12] J. Team, B. Lenz, A. Arazi, A. Bergman, A. Manevich, B. Peleg, B. Aviram, C. Almagor, C. Fridman, D. Padnos *et al.*, "Jamba-1.5: Hybrid transformer-mamba models at scale," *arXiv preprint arXiv:2408.12570*, 2024.
- [13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.
- [14] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.
- [15] S. A. Gaffar Shakhadri, K. KR, and K. Basavaraj Angadi, "Samba-asr: State-of-the-art speech recognition leveraging structured state-space models," *arXiv e-prints*, pp. arXiv–2501, 2025.
- [16] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.
- [17] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman *et al.*, "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.
- [18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2024.
- [19] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, "Mamba in speech: Towards an alternative to self-attention," *arXiv preprint arXiv:2405.12609*, 2024.
- [20] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.
- [21] S. Tang, L. Ma, H. Li, M. Sun, and Z. Shen, "Bi-mamba: Towards accurate 1-bit state space models," 2024. [Online]. Available: <https://arxiv.org/abs/2411.11843>
- [22] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.
- [23] L. Ma, M. Sun, and Z. Shen, "Fbi-llm: Scaling up fully binarized llms from scratch via autoregressive distillation," *arXiv preprint arXiv:2407.07093*, 2024.
- [24] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.
- [25] C. D. Hernández Mena, C. Armentano Oller, S. Solito, and B. Külebi, "3catparla: A new open-source corpus of broadcast tv in catalan for automatic speech recognition," in *Proc. IBER-SPEECH 2024*, 2024, pp. 176–180.
- [26] R. Cardenas, R. Zevallos, R. Baquerizo, and L. Camacho, "Sim-inchik: A speech corpus for preservation of southern quechua," in *Proceedings of the LREC 2018 Workshop "Improving Social Inclusion using NLP: Tools, Methods and Resources"*, 2018.
- [27] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirshahin, and C. Riveria, "Open-source high quality speech datasets for Basque, Catalan and Galician," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds. Marseille, France: European Language Resources association, May 2020, pp. 21–27.
- [28] A. Ebrahimi, M. Mager, A. Wiemerslage, P. Denisov, A. Oncevay, D. Liu, S. Koneru, E. Y. Ugan, Z. Li, J. Niehues *et al.*, "Findings of the second americasnlp competition on speech-to-text translation," in *NeurIPS 2022 Competition Track*. PMLR, 2023, pp. 217–232.
- [29] C. H. Mena, A. Simonsen, and J. Gudnason, "Asr language resources for faroeese," in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2023, pp. 32–41.
- [30] S. Hedstrom, D. E. Mollberg, R. Thorhallsdottir, and J. Gudnason, "Samromur: Crowd-sourcing large amounts of data," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 2311–2316.
- [31] C. Mena, A. Gatt, A. DeMarco, C. Borg, L. van der Plas, A. Muscat, and I. Padovani, "Masri-headset: A maltese corpus for speech recognition," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020.
- [32] X. Gao and N. F. Chen, "Speech-mamba: Long-context speech recognition with selective state spaces models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1–8.
- [33] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725.