BiMamba2 Masked Discrete-Unit Prediction for Multilingual Speech Representation

Prakriti Subedi, Howard Prioleau, Saurav K. Aryal PhD
AI4PC Lab, Institute for Human Centered Artificial Intelligence

INTRODUCTION

Speech technology has made remarkable progress, but most of the world's languages remain underserved. State-of-the-art self-supervised models like **HuBERT** and **wav2vec 2.0** rely on large curated corpora that do not exist for the majority of languages, creating stark performance gaps across linguistic and acoustic conditions.

The Unsupervised Speech in the Wild (UPS) Challenge at **Interspeech 2026** addresses this directly. Models train entirely on unlabeled audio and are evaluated on three frozen-representation probes with no fine-tuning: **Language Identification, ASR, and Speaker Clustering.**

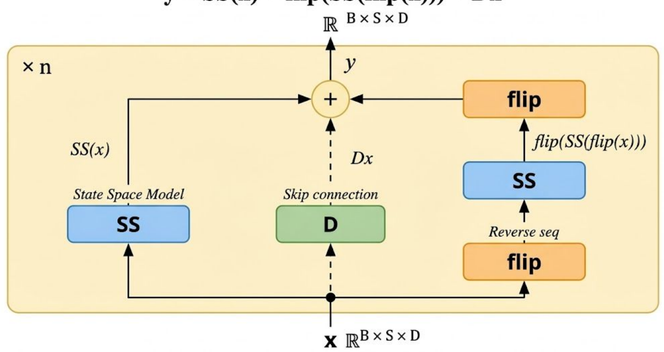
We propose **BiMamba2**, a bidirectional state-space encoder trained with masked discrete-unit prediction and zero labeled data, covering **67 languages.**

DATA PIPELINE

- Trained on unlabeled multilingual speech from the **MLCommons UPS dataset**
- Applied **voice activity detection (VAD)** to keep speech-rich segments and remove low-speech clips
- Converted audio into **log-mel spectrograms** for model input
- Used **language-aware sampling** to improve balance across languages during training

BIMAMBA2 ARCHITECTURE

$$y = SS(x) + flip(SS(flip(x))) + Dx$$



Bimamba2 layer: *forward path + backward path + skip connection*

BIMAMBA2 BACKBONE

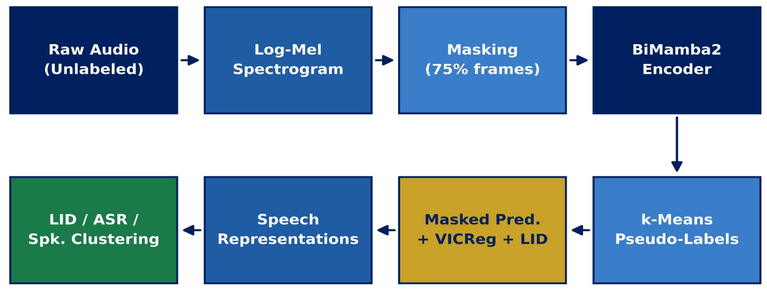
- Built a **bidirectional Mamba2 encoder** for efficient long-sequence speech modeling
- The model processes audio in both **forward and backward directions,** capturing past and future context
- A **skip path** preserves input information and helps stabilize training
- This design is well suited for **frozen speech representations**, where full-context embeddings are beneficial

METHOD

Overview

- Used **HuBERT-style** masked prediction to learn from unlabeled speech
- Masked valid frames and trained the model to recover discrete pseudo-labels
- Generated pseudo-labels with offline k-means clustering on **log-mel features**
- Added VICReg regularization to improve stability and representation diversity
- Added an auxiliary **language ID loss** to encourage multilingual structure in the embeddings

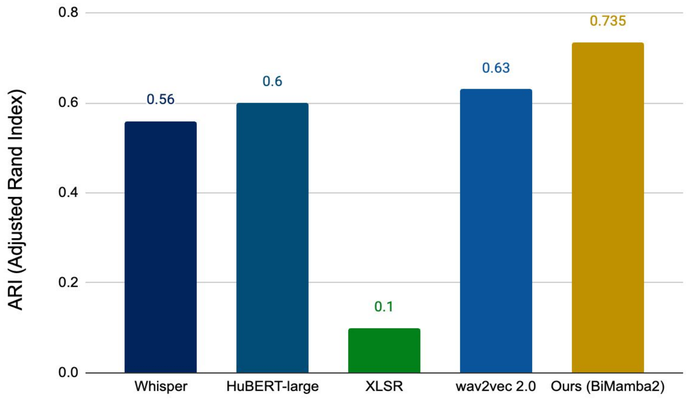
Model specs: 12-layer BiMamba2 | 47.88M params | ~250 h | 67 languages



End-to-end training pipeline *from unlabeled multilingual audio to speech representations evaluated on LID, ASR, and speaker clustering.*

RESULTS

Speaker Clustering: ARI vs Baselines



BiMamba2 achieved the highest official speaker clustering score, reaching ***ARI = 0.735*** *and outperforming* ***all compared baselines****. This result suggests the model learned strong speaker-discriminative multilingual speech representations from unlabeled audio.*

Model Scaling Results

Model	Macro-F1	CER	ARI
Ours (d=512, 8L)	0.052	0.996	0.291
Ours (d=768, 12L)	0.073	0.87	0.735

Increasing model size improved all official metrics, with the largest gain in speaker clustering ***(ARI: 0.291 → 0.735)****. This suggests that the larger BiMamba2 encoder learned stronger multilingual speech representations.*

CONCLUSION

- BiMamba2 learned multilingual speech representations from unlabeled audio using masked discrete-unit prediction, achieving ARI = 0.735 and outperforming all compared baselines.
- Future work will explore **iterative pseudo-label** refinement and broader multilingual evaluation.

REFERENCES

[1] T. Dao and A. Gu, "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality," ICML, 2024.
[2] "Unsupervised Speech in the Wild Challenge," Interspeech 2026.
[3] MLCommons, "Unsupervised People's Speech Dataset," 2024.
[4] W.-N. Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM TASLP, 2021

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.