A Comparative Study of Pre-trained Speech Encoders and Training Objectives for Large-Scale Indic Spoken Language Identification

Agneedh Basu*, Pavan Kumar J*, Sujith P*, Visruth Sanka*, Nihar Desai*, Prasanta Ghosh †
* AI & Robotics Technology Park (ARTPARK), IISc, Bangalore, India
† Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

***Abstract*—Spoken language identification (LID) for Indian languages is a challenging problem due to the large number of languages, significant phonetic overlap among related varieties, and the scarcity of labeled data for many low-resource languages. In this work, we present a systematic comparative study of two pre-trained speech encoders—Whisper and FastConformer—combined with a linear classifier for large-scale Indic LID spanning 42 languages across four linguistic families. We evaluate both encoders in frozen (linear probing) and fine-tuned settings, and compare three training objectives: cross-entropy (CE), supervised contrastive loss with cross entropy (CE + supCon), and hierarchical softmax (HSM). Models are trained on the Vaani dataset and evaluated in a cross-corpus setting on Vaani-Test (held-out), FLEURS, and Kathbath, providing insights into domain generalization. The frozen FastConformer encoder achieves over 90% macro accuracy on FLEURS and Kathbath without any task-specific adaptation, substantially outperforming Whisper on out-of-domain benchmarks, while fine-tuned Whisper yields stronger in-domain performance. HSM consistently outperforms CE and CE+SupCon for both encoders across all benchmarks, with the largest gains on out-of-domain test sets. CE+SupCon degrades FastConformer's cross-corpus generalization, suggesting that the contrastive objective over-specializes representations to in-domain conditions. Per-family analysis shows that Central Indo-Aryan varieties are the hardest to discriminate, with Hindi–Urdu and the Sadri–Chhattisgarhi–Surgujia cluster being the dominant confusion pairs.**

***Index Terms*—spoken language identification, Indic languages, Whisper, Conformer, FastConformer, contrastive learning, hierarchical softmax, cross-corpus evaluation**

I. Introduction

India is home to over 1,600 mother tongues [1], with 22 constitutionally recognized languages [2] spanning three major families: Indo-Aryan, Dravidian, Sino-Tibetan ([3], [4]). Automatic spoken language identification (LID) is a critical front-end component for multilingual speech pipelines, enabling ASR, machine translation, and voice-based services to route input to the correct language-specific model. Indic LID poses distinct challenges: closely related language varieties – such as Hindi, Bhojpuri, Chhattisgarhi, and Sadri – exhibit strong phonological and lexical overlap, making fine-grained discrimination difficult; many target languages are low-resource, with limited training data available; existing benchmarks differ substantially in recording conditions, speaking style, and speaker demographics, making cross-corpus generalization a practical concern.

Recent advances in self-supervised and weakly supervised speech representation learning have yielded powerful pre-trained encoders [5], [6] that can be adapted to downstream tasks with minimal supervision. However, it is unclear which encoder architecture is better suited for Indic LID, and how much the choice of training objective—standard cross-entropy (CE), metric learning via contrastive loss, or hierarchically-structured softmax—affects downstream performance and generalization.

In this paper, we address these questions through a controlled empirical study. Specifically, we make the following contributions:

- We benchmark Whisper and FastConformer encoders, both frozen and fine-tuned, as feature extractors for Indic LID across 42 languages.
- We compare three training objectives – cross-entropy, contrastive loss, and hierarchical softmax – on the same encoder backbones and report results across three test benchmarks (Section III-B).
- We conduct a cross-corpus evaluation to assess domain generalization (Section V).
- We provide a per-family error analysis to understand the role of linguistic proximity in LID confusion (Section VI).

II. Related Work

Early deep learning approaches to Indic LID used log-Mel spectrograms as input to CNN-based encoders with a linear classification head. Godbole et al. [7] demonstrated over 98% accuracy on seven Indian languages from the IIIT-H Indic speech database using this pipeline, and a related study [8] extended this to CNN+ANN classifiers on the same corpus, reporting accuracies above 99% while showing that CNN feature maps produce interpretable language-specific patterns. Despite these high numbers, both works evaluate on in-domain data only, leaving cross-corpus generalization unexplored. A survey of Indic LID [9] reviews existing systems across 23 studies and highlights the limited language coverage and dataset diversity of prior work. Self-supervised representations have more recently been applied to Indic LID. Ramesh et al. [10] explored phonotactic representations for seven Indian languages using wav2vec-based self-supervised features and senone classifier hidden representations, combined with linear, BiLSTM, or CNN classifiers. Self-supervised features consistently outperformed

supervised counterparts, achieving approximately 94% accuracy. Cross-corpus evaluation has emerged as a key challenge for Indic LID. Dey et al. [11] conducted a preliminary cross-corpus study across three Indian language corpora (IIITH-ILSC, LDC South Asian, IITKGP-MLILSC) using MFCC features with a TDNN architecture, showing that normalization via CMS and RASTA-filtering substantially improves out-of-domain LID performance. Extending this, [12] investigated cross-corpus Indic LID using ECAPA-TDNN with MFCC features, proposing domain diversification and generalization strategies to reduce the train–test domain gap. Domain-adaptive objectives have also been explored. Muralikrishna et al. [13] combined centroid classification loss with cross-entropy for LID in unseen target domains, achieving 94% on eight Indian languages.

Most closely related to our work, Shrivastava et al. [14] fine-tuned Whisper-large-v3 for joint ASR and LID on five Indian languages by incorporating language tokens into the ASR token sequence during training, reducing diarization error rate from 24.04% (vanilla Whisper) to 18.03% on the DiSPLACE 2024 challenge. Unlike that work, we study encoder representations in isolation as feature extractors for standalone LID, evaluated on a much broader set of 42 Indian languages across Vaani, FLEURS, and Kathbath. To our knowledge, LID has not been attempted on this scale of diverse Indian languages.

III. System Description

A. Model Architecture

Our system follows a simple encoder–classifier architecture: a pre-trained speech encoder maps an input utterance to a sequence of frame-level representations, which are aggregated via a pooling layer to obtain a fixed-length embedding vector $\mathbf{z}$. This embedding is then passed to a linear classification head. The encoder is either held frozen or fine-tuned end-to-end.

(a) Whisper: We use the encoder of `openai / whisper-medium` [5] which is trained for ASR on 90+ languages and produces a 1024-dimensional frame-level representation. This encoder alone comprises 350M parameters. (b) FastConformer: We use the encoder from `ARTPARK-IISc / Vaani-FastConformer-Multilingual` [15] which is trained on 60+ Indic languages, with an embedding dimension of 1024. The encoder consists of 430M parameters.

Self-Attention-pooling [16] over frames yields a fixed-length utterance embedding. A single linear layer projects the utterance embedding to logits over $L = 42$ target languages. In the *frozen* setting, only the final two layers are trained (linear probing). In the *fine-tuned* setting, encoder weights are updated jointly with these two layers.

B. Training Objectives

a) Cross-Entropy (CE): Standard multi-class cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\sum_{l=1}^{L} y_l \log \hat{p}_l, \tag{1}$$

where $y_l \in \{0, 1\}$ is the ground-truth indicator for language l and $\hat{p}_l$ is the predicted softmax probability.

b) Combined CE + SupCon Loss (CE + SupCon): Standalone supervised contrastive loss (SupCon) [17] optimizes the embedding space but provides no direct gradient to the classification head. Conversely, CE trains the output layer but does not explicitly encourage discriminative clustering in the embedding space. To benefit from both, we combine the two objectives:

$$\mathcal{L}_{\text{CE+SupCon}} = \lambda\, \mathcal{L}_{\text{SupCon}} + (1 - \lambda)\mathcal{L}_{\text{CE}}, \tag{2}$$

where λ is a weighting hyperparameter. The SupCon term is:

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_p / \tau}}{\sum_{a \neq i} e^{\mathbf{z}_i \cdot \mathbf{z}_a / \tau}}, \tag{3}$$

where $\mathbf{z}_i$ is the ℓ_2-normalized embedding of sample i, $P(i)$ the set of same-language positives in the batch, and τ a temperature hyperparameter. The combined loss thus simultaneously shapes the embedding space via $\mathcal{L}_{\text{SupCon}}$ and supervises the classification layer via $\mathcal{L}_{\text{CE}}$.

c) Hierarchical Softmax (HSM): We structure the 42 languages into a four-level linguistic tree: Root → Family → Sub-family → Language, with families Indo-Aryan, Dravidian, Sino-Tibetan, and European. Each internal node is associated with a linear classifier that predicts among its children; the language probability is the product of branch probabilities along the root-to-leaf path [18]. Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the embedding for sample i, and y_i its label. Let the path to class y_i in the hierarchy be: $\mathcal{P}(y_i) = \{(n_1, c_1), \ldots, (n_{L_i}, c_{L_i})\}$ where n_ℓ is a node and c_ℓ is the index of the correct child at that node. The conditional probability at node n_ℓ is $P(c_\ell \mid n_\ell, \mathbf{z}_i) = \mathrm{softmax}(W_{n_\ell}\mathbf{z}_i)_{c_\ell}$. The hierarchical likelihood is

$$P(y_i \mid \mathbf{z}_i) = \prod_{\ell=1}^{L_i} P(c_\ell \mid n_\ell, \mathbf{z}_i)$$

The hierarchical softmax loss is:

$$\mathcal{L}_{\text{HSM}} = -\sum_{\ell=1}^{L_i} \log P(c_\ell^{(i)} \mid n_\ell^{(i)}, \mathbf{z}_i)$$

IV. Experimental Setup

A. Datasets

Vaani [15] is a large-scale naturalistic Indic speech corpus. From this corpus, we curate a balanced subset by selecting 10 hours of speech from each of the 42 languages (see Appendix VIII-A for the full list), maximizing diversity in districts and speakers. The resulting data is split into train, validation, and test sets in an 8:1:1 ratio, ensuring that speakers do not overlap across splits.

a) Evaluation: We adopt a cross-corpus evaluation protocol: models are trained exclusively on Vaani and tested on three benchmarks to assess in-domain and out-of-domain performance – **Vaani-Test**: held-out portion of Vaani (in-domain); 13 languages from **FLEURS** [19]; and 11 from **Kathbath** [20].

TABLE I
MACRO ACCURACY (%) FOR FROZEN VS. FINE-TUNED ENCODERS WITH CE LOSS. **BOLD** = BEST PER ENCODER.

Encoder	Setting	Vaani-Test	FLEURS	Kathbath
Whisper	Frozen	56.0	61.9	57.7
	Fine-tuned	**71.8**	**72.7**	**68.3**
FastConformer	Frozen	67.4	**94.2**	**90.9**
	Fine-tuned	**67.6**	89.9	87.4

B. Implementation Details

All models are optimized using AdamW [21] with linear learning rate warmup following [22]. For full fine-tuning, we use a learning rate of $1e{-}5$, while for linear probing, a higher learning rate of $1e{-}4$ is used. Training is performed with a batch size of 4 and gradient accumulation over 24 steps, resulting in an effective batch size of 96. The combined CE + SupCon loss uses a temperature $\tau = 0.07$ and weighting $\lambda = 0.5$. All experiments are conducted on an NVIDIA L40 GPU.

C. Evaluation Metrics

We report macro-averaged accuracy (recall) over all languages present in each test set. Macro averaging treats each language equally regardless of test-set size, which is appropriate given the large variance in per-language data volumes.

V. RESULTS

A. Effect of Encoder and Fine-tuning

Table I compares frozen and fine-tuned settings for both encoders using CE loss. The two encoders exhibit contrasting behaviour. For Whisper, fine-tuning yields substantial gains across all benchmarks (+15.8% on Vaani-Test, +10.8% on FLEURS, +10.6% on Kathbath), indicating that Whisper's pre-trained representations, while broadly multilingual, benefit considerably from task-specific adaptation to Indic LID. For FastConformer, fine-tuning provides only a marginal improvement on Vaani-Test (+0.2%) but noticeably *degrades* cross-corpus performance (FLEURS: 94.2% $\rightarrow$ 89.9%; Kathbath: 90.9% $\rightarrow$ 87.4%), suggesting that the frozen FastConformer encoder generalizes better out-of-domain, and that fine-tuning on Vaani causes it to overfit to in-domain acoustic conditions. Notably, the frozen FastConformer outperforms the fine-tuned Whisper on FLEURS and Kathbath, highlighting the strong cross-lingual transferability of FastConformer's pre-trained representations for Indic languages.

B. Effect of Encoder and Training Objective

Table II reports macro accuracy for both fine-tuned encoders across all three training objectives, alongside two external LID baselines: Facebook MMS (FBMMS) [23] and SpeechBrain ECAPA-TDNN (Speechbrain) [24], [25]. MMS covers only 30 of the 42 target languages and ECAPA covers 13, so their Vaani-Test scores are not directly comparable to our systems.

HSM consistently outperforms CE and CE+SupCon for both encoders across all benchmarks. For Whisper, HSM achieves the best results (Vaani-Test: 74.2%, FLEURS: 73.8%, Kathbath: 75.8%), with the largest gain on Kathbath (+7.5% over CE), suggesting that the hierarchical structure aids out-of-domain generalization. For FastConformer, HSM similarly yields the best results across all benchmarks. CE+SupCon underperforms CE for FastConformer, suggesting the contrastive objective interferes with its strong pre-trained cross-lingual representations.

TABLE II
MACRO ACCURACY (%) ACROSS TEST BENCHMARKS FOR EACH TRAINING OBJECTIVE. **BOLD** = BEST PER BENCHMARK.

Model	Objective	Vaani-Test	FLEURS	Kathbath
Whisper	CE	71.8	72.7	68.3
	CE+SupCon	72.4	68.1	71.0
	HSM	**74.2**	**73.8**	**75.8**
FastConformer	CE	67.6	89.9	87.4
	CE+SupCon	64.5	82.7	79.8
	HSM	**67.7**	**91.4**	**90.0**
FBMMS	—	33.7†	95.1	91.3
Speechbrain	—	65.1‡	91.3	87.9

† 30/42 languages supported ‡ 13/42 languages supported

Among the external baselines, FastConformer + HSM matches SpeechBrain ECAPA on FLEURS (91.4% vs. 91.3%) and outperforms it on Kathbath (90.0% vs. 87.9%) despite covering over three times as many languages. MMS achieves 95.1% on FLEURS but only over its supported 30 languages; its low Vaani-Test score (33.7%) reflects the many Vaani-specific low-resource languages absent from its inventory.

C. Cross-corpus Generalization

A consistent pattern across both tables is that FastConformer generalizes substantially better to out-of-domain benchmarks than Whisper, regardless of training objective. The frozen FastConformer achieves 94.2% on FLEURS and 90.9% on Kathbath, compared to 61.9% and 57.7% for the frozen Whisper. Even after fine-tuning, FastConformer with HSM (FLEURS: 91.4%, Kathbath: 90.0%) far outperforms the best Whisper result (FLEURS: 73.8%, Kathbath: 75.8%). This suggests that FastConformer's pre-training produces more domain-robust acoustic representations for Indic languages, while Whisper's representations are more sensitive to domain shift despite its larger pre-training corpus. The CE+SupCon objective degrades cross-corpus performance for FastConformer (FLEURS: 82.7% vs. 89.9% for CE), indicating that the contrastive loss may over-specialize the embedding space to Vaani's acoustic conditions.

VI. ANALYSIS

A. Per-family Performance

Table III breaks down macro accuracy by language family and sub-family on Vaani-Test set for both encoders with HSM loss. The results reveal a striking contrast between language groups. Sino-Tibetan languages achieve the highest accuracy (97.1%), which is somewhat surprising given the limited training data for languages such as Wancho, Garo, and Chakma;

TABLE III
PER-FAMILY MACRO ACCURACY (%) ON VAANI-TEST FOR BOTH ENCODERS WITH HSM LOSS.

Language Family / Sub-family	Whisper	FastConformer
Indo-Aryan (all)	67.6	58.9
Central	58.7	47.5
Eastern	84.8	88.0
Western	76.8	75.9
Northern	79.9	56.0
Dravidian	85.9	84.3
Sino-Tibetan	97.1	95.4
European (English)	92.3	92.2
Overall	**74.2**	67.7

this likely reflects that these languages are acoustically highly distinct from the Indo-Aryan and Dravidian majority, making them easier to separate despite fewer training samples. English similarly achieves 92.3%, consistent with it being the only European language and therefore trivially distinct from all others. Dravidian languages also perform strongly at 85.9%, reflecting the relative phonological distinctiveness of Tamil, Telugu, Kannada, and Malayalam from one another and from Indo-Aryan varieties.

Indo-Aryan languages as a group are the most challenging (67.6%), and the sub-family breakdown reveals where the difficulty lies. Central Indo-Aryan is by far the hardest sub-family at 58.7%, comprising closely related varieties such as Hindi, Bhojpuri, Chhattisgarhi, Sadri, Surgujia, Bajjika, and Halbi that share substantial phonological and lexical overlap. Eastern (84.8%), Western (76.8%), and Northern (79.9%) Indo-Aryan sub-families perform considerably better, likely because languages within these groups are somewhat more acoustically distinct from each other, or because they have less within-group overlap in the training data. The low Central Indo-Aryan accuracy is the primary driver of the overall Indo-Aryan deficit and represents the key remaining challenge for large-scale Indic LID.

B. Confusion Among Related Languages

Fig. 1 shows confusion matrices for Central Indo-Aryan languages under the HSM objective for both encoders. Two dominant confusion patterns emerge across both models. First, Hindi and Urdu are heavily mutually confused, which is linguistically expected given their near-identical spoken form — both share the same phonological inventory and differ primarily in script and formal vocabulary. This confusion is more severe in FastConformer, where Urdu is almost entirely misclassified as Hindi, whereas Whisper maintains a somewhat stronger Urdu diagonal.

Second, the Sadri–Chhattisgarhi–Surgujia cluster exhibits significant mutual confusion in both encoders, reflecting the high degree of phonological overlap among these closely related Central Indo-Aryan varieties. In the FastConformer matrix, Sadri shows weak diagonal activation with mass spread across Chhattisgarhi and Surgujia. Whisper handles this cluster somewhat better, producing more distinct diagonal entries for

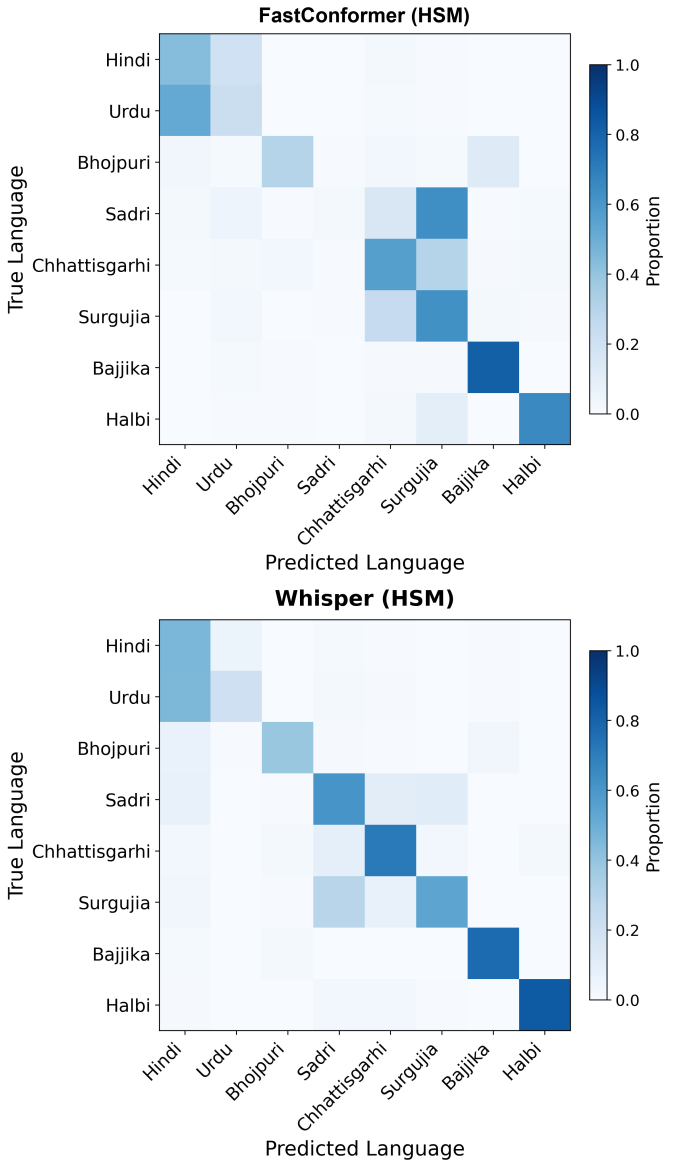

Fig. 1. Confusion matrices for Central Indo-Aryan languages: Whisper + HSM vs. FastConformer + HSM. (Darker diagonal = better discrimination.)

Sadri and Chhattisgarhi. In contrast, Bajjika and Halbi are well-identified by both encoders, likely due to sufficient and acoustically distinctive training samples for these languages. Bhojpuri also shows moderate accuracy in both models, though FastConformer exhibits some confusion between Bhojpuri and Bajjika.

VII. CONCLUSION

We presented a systematic comparison of pre-trained speech encoder architectures and training objectives for large-scale Indic spoken language identification across 42 languages. Our cross-corpus evaluation on Vaani-Test, FLEURS, and Kathbath shows that the FastConformer encoder substantially outperforms Whisper on out-of-domain benchmarks, with the frozen FastConformer achieving over 90% macro accuracy

on FLEURS and Kathbath without any task-specific fine-tuning, while fine-tuning Whisper yields stronger in-domain performance. Among training objectives, HSM consistently outperforms both CE and CE+SupCon for both encoders across all benchmarks, with the most pronounced gains on Kathbath. CE+SupCon degrades FastConformer's cross-corpus generalization, suggesting that the contrastive objective over-specializes representations to in-domain acoustic conditions. Per-family analysis reveals that Central Indo-Aryan varieties are the hardest to discriminate (58.7%), with Hindi–Urdu and the Sadri–Chhattisgarhi–Surgujia cluster accounting for the majority of errors, owing to their high degree of phonological overlap. Sino-Tibetan and Dravidian languages are comparatively well-identified despite differing resource levels, highlighting the importance of inter-language acoustic distinctiveness. Future work will explore targeted data augmentation for low-accuracy Central Indo-Aryan varieties and language-family-aware training strategies.

REFERENCES

[1] R. C. Nigam, *Language Handbook on Mother Tongues in Census of India, 1971*. New Delhi: Office of the Registrar General, India (Ministry of Home Affairs), 1972, accessed: 2025-07-22. [Online]. Available: https://language.census.gov.in/eLanguageDivision_VirtualPath/eArchive/pdf/28.pdf

[2] GoI, "Official Indian Languages," https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution, accessed: 2025-07-22.

[3] C. P. Masica, *The Indo-Aryan Languages*. Cambridge University Press, 1991.

[4] A. Abbi, "India as a linguistic area revisited," *Language Sciences*, vol. 28, no. 6, pp. 617–633, 2006.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, R. Pang *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of Interspeech*, 2020.

[7] S. Godbole, V. Jadhav, and G. Birajdar, "Indian language identification using deep learning," *ITM Web of Conferences*, vol. 32, p. 01010, 01 2020.

[8] H. Mukherjee, S. Ghosh, S. Sen, O. Sk, K. Santosh, S. Phadikar, and K. Roy, "Deep learning for spoken language identification: Can we visualize speech signal patterns?" *Neural Computing and Applications*, vol. 31, 12 2019.

[9] S. Dey, M. Sahidullah, and G. Saha, "An overview of Indian spoken language recognition from machine learning perspective," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, pp. 1 – 45, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247319584

[10] G. Ramesh, C. S. Kumar, and K. S. R. Murty, "Self-Supervised Phonotactic Representations for Language Identification," in *Interspeech 2021*, 2021, pp. 1514–1518.

[11] S. Dey, G. Saha, and M. Sahidullah, "Cross-corpora language recognition: A preliminary investigation with Indian languages," *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 546–550, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234357575

[12] S. Dey, M. Sahidullah, and G. Saha, "Cross-corpora spoken language identification with domain diversification and generalization," *ArXiv*, vol. abs/2302.05110, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256808424

[13] H. Muralikrishna, S. Kapoor, D. A. Dinesh, and P. Rajan, "Spoken language identification in unseen target domain using within-sample similarity loss," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7223–7227.

[14] A. Shrivastava *et al.*, "A unified approach to multilingual automatic speech recognition with improved language identification for indic languages," in *Proceedings of Interspeech*, 2024.

[15] S. Pulikodan, A. Singh, A. Basu, N. Desai, P. K. J, P. D. Bhat, R. Dharmaraju, R. Gupta, S. Udupa, S. Kumar, S. Sharma, V. Sanka, D. Tewari, H. Dhand, A. Kamat, S. Singh, S. Vashishth, P. Talukdar, R. Acharya, and P. K. Ghosh, "VAANI: Capturing the language landscape for an inclusive digital India," 2026. [Online]. Available: https://arxiv.org/abs/2603.28714

[16] P. Safari, M. India, and J. Hernando, "Self-attention encoding and pooling for speaker recognition," 10 2020, pp. 941–945.

[17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[19] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2023.

[20] T. Javed, S. Doddapaneni, A. Raman, K. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. Khapra, "Towards building ASR systems for the next billion users," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10 813–10 821, 06 2022.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[23] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, W.-N. Wang *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[24] J. Valk and T. Alumäe, "VoxLingua107: A dataset for spoken language recognition," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn-based speaker verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834.

VIII. APPENDIX

A. Languages

We consider 42 languages organized into a hierarchical taxonomy spanning multiple language families. The languages were chosen on the basis of sufficient durations being available. The Indo-Aryan family is divided into four sub-groups: Central (Hindi, Urdu, Bhojpuri, Sadri, Chhattisgarhi, Surgujia, Bajjika, Halbi), Eastern (Angika, Bengali, Assamese, Maithili, Magahi, Khortha, Odia, Sambalpuri, Surjapuri, Nepali, Nagamese), Western (Marathi, Konkani, Rajasthani, Marwari, Gujarati, Malvani), and Northern (Punjabi, Kumaoni, Garhwali, Haryanvi, Khariboli). The Dravidian family includes Telugu, Kannada, Tamil, Tulu, and Malayalam. The Sino-Tibetan family comprises Wancho, Sumi, Garo, Kokborok, and Chakma, while the European family is represented by English.