AUTOMATIC SPEECH RECOGNITION IN SANSKRIT

Sanskrit, as a low-resource language, presents significant challenges for ASR development due to limited speech data and complex phonological structures. This research addresses these fundamental obstacles to build effective speech recognition systems for this ancient language.

AUTHORS

Devansh Abhay Dhok
Suraj Jaiswal
Prof. R.M Hegde

AFFILIATIONS

Office of Outreach Activities, Indian Institute of Technology Kanpur

CHALLENGES

- **Limited Speech Data**: Scarce availability of digitized Sanskrit speech corpora compared to major world languages
- **Complex Phonetic Inventory**: Extensive set of 48+ distinct phonemes including retroflex, aspirated, and nasalized sounds
- **Sandhi Rules:** Morphophonological changes that alter sounds at word boundaries, creating acoustic variability
- **Prosodic Complexity:** Distinctive stress patterns, pitch accents, and rhythmic structures crucial for meaning
- **Lack of Standardization:** Absence of unified pronunciation standards and evaluation benchmarks for Sanskrit ASR systems

DATASETS USED

- **Sanskrit Speech Data**: 8 hours of labeled Sanskrit speech from **IndicSUPERB** Kathbath dataset [3] representing typical low-resource language constraints for ASR development
- **Hindi Baseline Data**: 22 hours of Hindi speech from **IndicSUPERB** Kathbath [3] used for DSN baseline training and cross-lingual transfer learning comparison

APPROACHES USED

Approach 1: Domain Separation Networks (DSN)

- *Cross-Domain Architecture:* DSN with private and shared encoders to separate domain-specific and domain-invariant speech representations between Hindi (source) and Sanskrit (target)
- *Adversarial Training:* Domain classifier with gradient reversal layer to learn domain-invariant features while maintaining senone classification accuracy on source domain

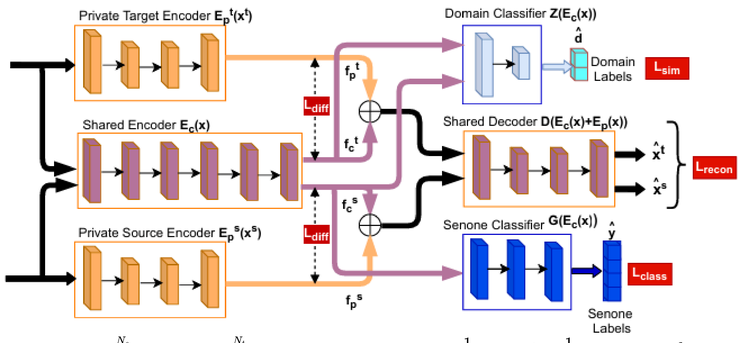
Approach 2: Whisper Fine Tuning

- *Transfer Learning:* OpenAI Whisper-small model fine-tuned on Sanskrit data, leveraging pre-trained multilingual representations for low-resource ASR adaptation
- *Parameter Strategy:* Selective fine-tuning with frozen encoder layers and trainable decoder components to prevent overfitting on limited Sanskrit data

RESULTS/FINDINGS

ANALYSIS

DOMAIN SEPARATION NETWORKS



$$L_{\text{recon}} = \sum_{i=1}^{N_s} \|\mathbf{x}_i^s - \hat{\mathbf{x}}_i^s\|^2 + \sum_{i=1}^{N_t} \|\mathbf{x}_i^t - \hat{\mathbf{x}}_i^t\|^2$$

$$L_{SIMSE} = \frac{1}{k}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 - \frac{1}{k^2}\left((\mathbf{x} - \hat{\mathbf{x}}) \cdot \mathbf{1}_k\right)^2$$

$$L = L_{\text{class}} + \beta L_{\text{sim}} + \gamma L_{\text{diff}} + \delta L_{\text{recon}}$$

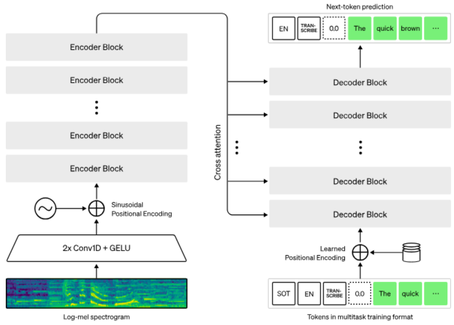
$$L_{diff} = ||F_c^{sT} F_p^s||_F^2 + ||F_c^{tT} F_p^T||_F^2$$

Architecture Components:
Domain Separation Networks model both private and shared components of speech representations across Hindi (source) and Sanskrit (target) domains. The architecture includes private encoders (E_p^t, E_p^s) that extract domain-specific features, and a shared encoder (Ec) that learns domain-invariant representations. A shared decoder reconstructs inputs while a senone classifier maps features to phonetic labels.

Training Strategy:
The network optimizes a multi-objective loss L. L_class handles senone classification on source data, L_sim ensures domain-invariant features through adversarial training with gradient reversal, L_diff enforces orthogonality between private and shared components, and L_recon maintains reconstruction fidelity. This approach enables effective cross-domain knowledge transfer from Hindi to Sanskrit.

Implementation Details:
The domain classifier uses adversarial training to distinguish between domains while the shared encoder learns to fool it, creating domain-invariant features. Hyperparameters β, γ, and δ balance the different loss components for optimal performance.

OPENAI'S WHISPER-BASE MODEL



$$P = [p_{pitch}, p_{energy}, p_{delta}, p_{delta-delta}]^T$$

$$P' = [p'_{pitch}, p'_{energy}, p'_{delta}, p'_{delta-delta}]^T \quad F = \begin{bmatrix} M \\ P' \end{bmatrix}$$

$$f(x) = \text{sgn}(x) \cdot \frac{\log(1 + \mu \cdot |x|)}{\log(1 + \mu)}$$

Model Configuration:
Whisper-base employs a transformer encoder-decoder architecture pre-trained on multilingual speech data. The model processes 80-channel mel-spectrogram features and uses attention mechanisms for robust cross-lingual transfer learning.

Fine-tuning Approach:
Selective fine-tuning strategy adapts pre-trained representations to Sanskrit while preserving multilingual knowledge. Training uses CTC loss with gradient accumulation and regularization techniques to prevent overfitting on limited Sanskrit data.

Prosodic Feature Integration:
Enhanced Whisper by incorporating prosodic features using early fusion strategy. The prosodic feature vector **P** contains pitch, energy, delta, and delta-delta components normalized to [-1,1] range, then transformed using **μ-law companding** with μ = 255. The companded features **P'** are stacked with Whisper's truncated 40-bin mel-spectrogram **M** to form the final fused feature vector **F,** maintaining 80-dimensional input compatibility while integrating both spectral and prosodic speech dynamics.

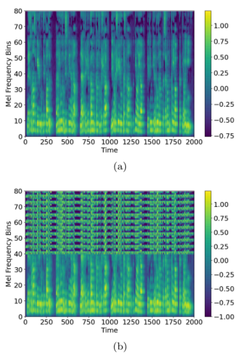


Figure : Image showing
(a) Original MFCC
(b) MFCC with prosodic feature embedding

CONCLUSION

Features	WER	CER
MFCC	62.73	20.09
MFCC + Pr	69.66	24.19

Table: Performance Evaluation of Whisper Model with different feature sets and datasets. 'Py' is Prosody

- **Key Findings and Unexpected Results:** Whisper fine-tuning established a strong baseline for low-resource Sanskrit ASR.
- However, incorporating prosodic features (pitch, energy, etc.) led to a 6.93% increase in WER, contradicting expectations of improved accuracy.
- This performance drop indicates that simple concatenation of prosodic features is ineffective for integration.
- The added features likely introduced noise or conflicted with Whisper's pre-trained representations, rather than enhancing recognition.

Research Implications: Transfer learning offers a strong foundation for low-resource ASR, but naive integration can degrade performance — underscoring the need for advanced methods like cross-modal attention to effectively leverage prosodic cues.

References:

- [1] S. Jaiswal, G. Routray, A. Rai, P. Dwivedi and R. M. Hegde, "Low Resource Verse Dataset and Prosodic Feature Integration for Sanskrit ASR," 2025 National Conference on Communications (NCC), New Delhi, India, 2025, pp. 1-6, doi: 10.1109/NCC63735.2025.10983176.
- [2] A. C S, P. A P and A. G. Ramakrishnan, "Unsupervised Domain Adaptation Schemes for Building ASR in Low-Resource Languages," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 2021, pp. 342-349, doi: 10.1109/ASRU51503.2021.9688269.
- [3] Javed, T., Bhogale, K., Raman, A., Kumar, P., Kunchukuttan, A., & Khapra, M. M. (2023). IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 12942-12950. https://doi.org/10.1609/aaai.v37i11.26521

devanshad23@iitk.ac.in
https://surge.iitk.ac.in