Jointly Improving Dialect Identification and ASR in Indian Languages using Multimodal Feature Fusion

Saurabh Kumar, Amartyaveer, Prasanta Kumar Ghosh

SPIRE LAB
Department of Electrical Engineering
Indian Institute of Science (IISc), Bangalore, India

Interspeech 2025

Overview

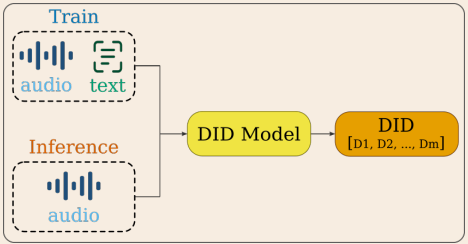
Overview

Motivation: The Challenge of Dialectal Variation

- Automatic Speech Recognition (ASR) and Dialect Identification (DID) are critical for linguistically diverse country like India.
- Significant dialectal differences in pronunciation, vocabulary, and grammar pose a major challenge for ASR systems [Udupa et al., 2023].
- Accurately identifying a speaker's dialect can allow an ASR system to adapt, mitigating errors and improving performance.
- Though DID is similar to language identification (LID), DID is inherently difficult as dialects of the same language share many phonetic and lexical characteristics.

Training–Test Modalities for DID

Train ↓ / Test →	Audio	Text	Audio+Text
Audio	Weaker	–	–
Text	–	Weaker	–
Audio+Text	Preferred (this work)	–	Limited



Key insight:

- **Audio-only / Text-only** are *suboptimal*: they fail to use complementary acoustic/lexical cues. In our prior work [Amartyaveer et al., 2025], these underperformed *ASR-based DID* trained with **Audio+Text**.
- **Audio+Text (Train & Test)** is *limited in practice*: many under-resourced Indian languages lack transcribed data.

Focus in this work:

- **Train:** use **Audio+Text** to learn richer, complementary representations.
- **Test:** require **Audio-only** to enable scalable, transcript-free deployment.

The Problem: A Performance Trade-Off

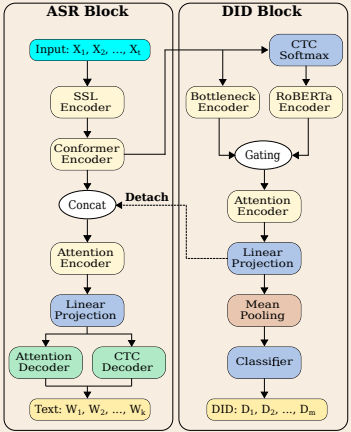
- Joint ASR–DID models often face a trade-off: gains in one task can degrade the other.
- **Prior findings:**
  - Japanese: best DID setup degraded ASR performance [Imaizumi et al., 2022].
  - Irish: improved DID slightly reduced ASR performance [Lonergan et al., 2024].
  - Indian (8 langs.): multimodal features boosted DID but slightly degraded ASR [Amartyaveer et al., 2025].
- **Common issue:** Most ASR-based DID methods prepend/append dialect ID to training text. This false context significantly degrades ASR performance.
- **Our goal:** Use dialect information as *probabilistic features*, and improve **both** ASR and DID.

Our Contributions

- A **novel joint framework** that improves both ASR and DID by fusing features with a gating mechanism:
  - **Speech Features:** A Bottleneck Encoder [Srinivas et al., 2021] refines Conformer outputs.
  - **Text Features:** A RoBERTa encoder [Liu et al., 2019] processes ASR-generated CTC embeddings.
- The fused dialect embedding enhances the ASR encoder directly, an approach that avoids prepending dialect tokens and improves robustness against DID errors.
- Achieved state-of-the-art results on **8 Indian languages** (33 dialects): **81.63%** DID accuracy, **4.65%** CER, and **17.73%** WER.
- Publicly released code and models to encourage further research [1].

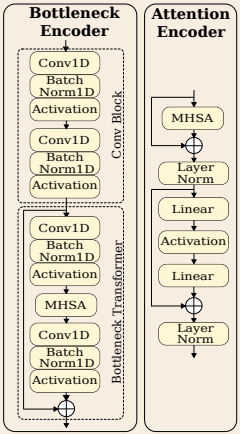
[1] https://github.com/labspire/respin_did_interspeech25.git

Proposed Method: ASR-BN-ROB Architecture



- **ASR-BN-ROB:** Joint ASR–DID using Bottleneck (BN) [Srinivas et al., 2021] encoder for speech features and RoBERTa (ROB) [Liu et al., 2019] encoder for text features.
- **ASR Block:** Conformer-based CTC+Attention ASR with SSL features.
- **DID Block:** BN captures acoustic/temporal cues from ASR output; ROB processes CTC embeddings for lexical/semantic cues.
- **Fusion:** Gating mechanism adaptively combines BN and ROB features.
- **Integration:** Fused probabilistic dialect features fed back into ASR encoder (not as tokens) to avoid degradation from DID errors.
- **Stability:** Gradient detachment from DID to ASR ensures stable joint training.

Bottleneck and Attention Encoder Architecture



Bottleneck Encoder:

- Extracts dialectal cues from Conformer output.
- Uses stacked Conv1D, BatchNorm, activations; MHSA further captures long-range context.
- Based on [Srinivas et al., 2021], with 1D convs and no pooling to preserve temporal info.

Attention Encoder:

- Refines fused speech-text features using MHSA, LayerNorm, residual FFN with up/down projections.
- Outputs frame-level embeddings; mean pooled for DID classification.

DID Block: Gating and Attention Fusion

Gating:

$$\mathbf{G} = \sigma\big(\mathbf{W}_g[\mathbf{H}_{\text{bottleneck}}, \mathbf{H}_{\text{rob}}] + \mathbf{b}_g\big)$$

Fusion:

$$\mathbf{H}_{\text{fused}} = \mathbf{G} \odot \mathbf{H}_{\text{rob}} + (1-\mathbf{G}) \odot \mathbf{H}_{\text{bottleneck}}$$

Refinement:

$$\mathbf{H}_{\text{fused}} \xrightarrow{\text{Attention Encoder}} \text{Dialect embedding}$$

Notes:

- $\mathbf{H}_{\text{bottleneck}}$: BN (speech) features
- $\mathbf{H}_{\text{rob}}$: ROB (text) features from CTC embeddings
- $[\cdot, \cdot]$: concat along feature dim
- $\mathbf{W}_g, \mathbf{b}_g$: learnable weights, bias
- $\sigma(\cdot)$: element-wise sigmoid
- $\odot$: element-wise multiplication

Joint Optimization Objective

Goal: Train ASR and DID jointly using a multi-task loss.

Total Loss:

$$L = \underbrace{\lambda_{CTC}\mathcal{L}_{CTC} + (1 - \lambda_{CTC})\mathcal{L}_{ATT}}_{\text{ASR Loss}} + \underbrace{\gamma_{CE}\mathcal{L}_{CE}}_{\text{DID Loss}}$$

- $\mathcal{L}_{CTC}$: CTC loss for faster convergence and alignment.
- $\mathcal{L}_{ATT}$: Attention decoder loss for richer context modeling.
- $\mathcal{L}_{CE}$: Cross-entropy loss for dialect classification.
- λ_{CTC}, γ_{CE}: Control ASR-DID loss balance.

Overview

Experimental Setup: Dataset

Dataset

- We use a subset of the **RESPIN** [RESPIN, 2025] dataset, following the same data split as in [Amartyaveer et al., 2025].
- Covers **8 Indian languages**:
  - Bhojpuri (bh), Bengali (bn), Chhattisgarhi (ch), Kannada (kn), Magahi (mg), Maithili (mt), Marathi (mr), and Telugu (te).
- A total of **33 distinct dialects** across these languages.
- Approximately 140-175 hours of training data in a read-speech setting.

Lang	bh	bn	ch	kn	mg	mr	mt	te
#Dialects	3	5	4	5	4	4	4	4

Table: Number of dialects per language in the RESPIN dataset.

Dataset: Preserving Dialect in Read-Speech

- **All RESPIN data in this work are read-speech.**
- **Design to preserve dialectal diversity:**
  - *Sentence composition:* Native composers wrote in regional/colloquial variants.
  - *No normalization:* No orthographic or lexical standardization at composition time.
  - *Speaker selection:* Native speakers from the corresponding dialect regions.
  - *Reading protocol:* Speakers read *exactly as written*, without normalization or "standard" correction.
- **Outcome:** Dialect-specific lexical, phonetic, and syntactic cues are preserved in both text and audio, despite the read-speech setting.

Examples & Code

`https://github.com/labspire/respin_did_interspeech25`

Scan for audio & text examples

Experimental Setup: Baselines

Compared Methods

- **Base-ASR:** Conformer-based CTC+Attention ASR (SSL features), no DID component.
- **ASR-DID:** Ground-truth dialect ID prepended to training text.
- **ASR-DID-AUX:** ASR-DID with auxiliary CTC loss for DID.
- **ASR-DID-ROB (baseline):** Uses RoBERTa encoder on CTC embeddings for DID.
  - Chosen as baseline because in our recent work [Amartyaveer et al., 2025], it outperformed other ASR-DID variants in terms of DID accuracy.
- **Proposed: ASR-BN-ROB** and ablations:
  - **ASR-BN:** Bottleneck encoder only (speech features).
  - **ASR-ROB:** RoBERTa encoder only (text features).

Overview

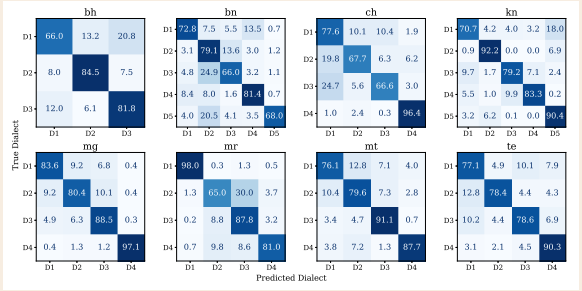
DID Performance: Accuracy (%)

Table: Our proposed model achieves the highest average DID accuracy.

System	bh	bn	ch	kn	mg	mr	mt	te	Avg
ASR-DID	74.91	72.43	77.81	80.08	86.26	82.07	82.35	77.58	79.19
ASR-DID-SC	75.54	72.64	76.60	80.96	86.19	82.72	81.94	78.89	79.44
ASR-DID-AUX	75.72	73.10	76.38	81.80	86.88	81.64	82.77	80.28	79.82
ASR-DID-ROB (baseline)	77.43	73.38	77.00	83.18	87.31	82.90	83.67	81.06	80.74
ASR-ROB	77.32	74.28	76.33	83.11	87.77	83.18	83.62	79.22	80.60
ASR-BN	77.39	73.19	78.88	82.56	87.59	81.82	**84.22**	80.82	80.81
ASR-BN-ROB (Proposed)	**78.74**	**74.46**	**79.38**	**83.55**	**87.88**	**83.66**	83.11	**82.23**	**81.63**

- The proposed ASR-BN-ROB model, which fuses both modalities, achieves the highest average DID accuracy of **81.63%**.
- This demonstrates the benefit of multimodal fusion over single-modality approaches.

DID Confusion Matrices: Baseline vs. Proposed



(a) Baseline (ASR-DID-ROB)

(b) Proposed (ASR-BN-ROB)

- **Key Insight:** Our model not only improves overall accuracy but also enhances generalization across dialects.
- This is shown by a **16.08% average reduction** in the standard deviation of dialect-wise accuracies, indicating more consistent and robust performance.

ASR Performance: CER (%) and WER (%)

Table: Proposed method improves ASR while also boosting DID.

System	Average CER (%)		Average WER (%)	
	CER	Rel. Impr.	WER	Rel. Impr.
Base-ASR	4.81	–	18.38	–
ASR-DID-ROB (baseline)	4.76	1.0%	18.16	1.2%
ASR-BN-ROB (proposed)	**4.65**	**3.3%**	**17.73**	**3.5%**

- **Best ASR performance:** CER 4.65%, WER 17.73%.
- **Breaks the trade-off:** Previous joint methods often degraded ASR, but our model improves both ASR and DID.
- Full language-wise results are reported in the paper.

Analysis: Impact of Incorrect DID on ASR

- A key challenge for some joint models is their reliance on prepending dialect IDs to text.
- **What happens when the DID prediction is wrong?**
  - The ASR model is conditioned on "false context".
  - This can significantly harm transcription accuracy, worsening the ASR/DID trade-off.
- We analyse ASR performance specifically on utterances where the baseline model misclassifies the dialect.
- **Hypothesis:** Our model, which does not prepend IDs, should be more robust to these errors.

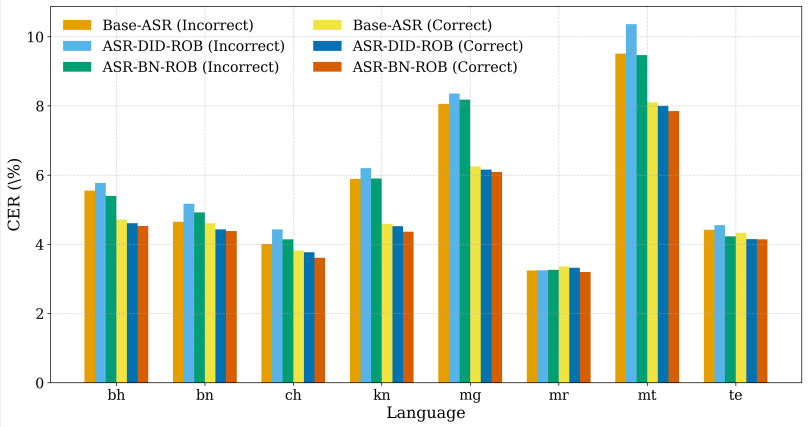
Analysis: ASR Robustness to DID Errors

Table: ASR performance on utterances with correct vs. incorrect DID predictions.

	CER (%)		WER (%)	
ASR System	**Incorrect**	**Correct**	**Incorrect**	**Correct**
Base-ASR	5.67	4.66	20.74	17.93
ASR-DID-ROB (baseline)	6.01	4.51	21.85	17.43
ASR-BN-ROB (proposed)	**5.68**	**4.46**	**20.72**	**17.14**
Relative difference (%)	**5.52**	1.24	**5.17**	1.62

- **Key takeaway:** Proposed model is more robust, especially when DID is incorrect.
- **Incorrect DID: 5.52% CER** and **5.17% WER** relative reduction over baseline.
- **Correct DID:** Consistent small gains in both CER and WER.

Language-wise CER for Incorrect vs Correct DID predictions



Observations:

- CER is consistently lower when DID predictions are correct.
- ASR-BN-ROB yields the lowest CER across most languages.
- Compared to ASR-DID, largest CER improvement observed in **mt**, **te**, and **ch** for incorrect DID.

Statistical Significance

- To confirm the robustness of our results, we performed a paired T-test at a 95% confidence interval across the 8 languages, comparing our proposed model to the ASR-DID-ROB baseline.
- The improvements are **statistically significant** for all key metrics:
  - **DID Accuracy:** $p = 0.0211$
  - **ASR CER:** $p = 0.0002$
  - **ASR WER:** $p = 0.0006$
- These results strongly validate the effectiveness of our proposed multimodal fusion approach.

Overview

Conclusion

- We presented a novel multimodal feature fusion approach that successfully improves both DID and ASR for Indian languages, mitigating the traditional performance trade-off.
- By integrating speech features (Bottleneck Encoder) and text features (RoBERTa on CTC embeddings) through an adaptive gating mechanism, our model achieves superior performance.
- Our proposed ASR-BN-ROB model establishes a new state-of-the-art, with **81.63% DID accuracy**, **4.65% CER**, and **17.73% WER**.
- The model is also more robust to incorrect dialect predictions, a key weakness in previous systems.

Future Work

- **Expand to spontaneous and real-world speech:**
  - Evaluate the fusion approach on spontaneous, conversational, and noisy speech scenarios.
  - Extend to multilingual and code-switched ASR settings.
- **Generalize across dialects and languages:**
  - Assess model robustness across a wider set of dialects and under-represented languages.
- **Explore learning paradigms:**
  - Investigate unsupervised and semi-supervised techniques to reduce label dependence.
  - Study alternative fusion strategies and upstream SSL feature encoders.

Overview

References I

Amartyaveer, Kumar, S., Sharma, S., Udupa, S., Badiger, S., Singh, A., G, D., Bandekar, J., Murthy, S., and Kumar Ghosh, P. (2025).
Improving Dialect Identification in Indian Languages Using Multimodal Features from Dialect Informed ASR.
In *Proc. ICASSP*, pages 1–5.

Imaizumi, R., Masumura, R., Shiota, S., and Kiya, H. (2022).
End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning.
APSIPA Transactions on Signal and Information Processing, 11(1):–.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019).
RoBERTa: A Robustly Optimized BERT Pretraining Approach.
ArXiv, abs/1907.11692.

References II

Lonergan, L., Qian, M., Chiaráin, N. N., Gobl, C., and Chasaide, A. N. (2024).
Low-resource speech recognition and dialect identification of Irish in a multi-task framework.
In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 67–73.

RESPIN (2025).
RESPIN: Speech Recognition in Agriculture and Finance for the Poor in India.
https://respin.iisc.ac.in/.
Accessed: August 12, 2025.

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021).
Bottleneck Transformers for Visual Recognition.
In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16514–16524.

References III

Udupa, S., Bandekar, J., Deekshitha, G., Kumar, S., Ghosh, P. K., Badiger, S., Singh, A., Murthy, S., Pai, P., Raghavan, S., and Nanavati, R. (2023).

Gated Multi Encoders and Multitask Objectives for Dialectal Speech Recognition in Indian Languages.

In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Acknowledgments

- This work was supported by the **RESPIN project**, funded by the Gates Foundation.
- We extend our thanks to the entire RESPIN team and our project partner, Navana Tech, for their valuable contributions to data collection.

Thank You!

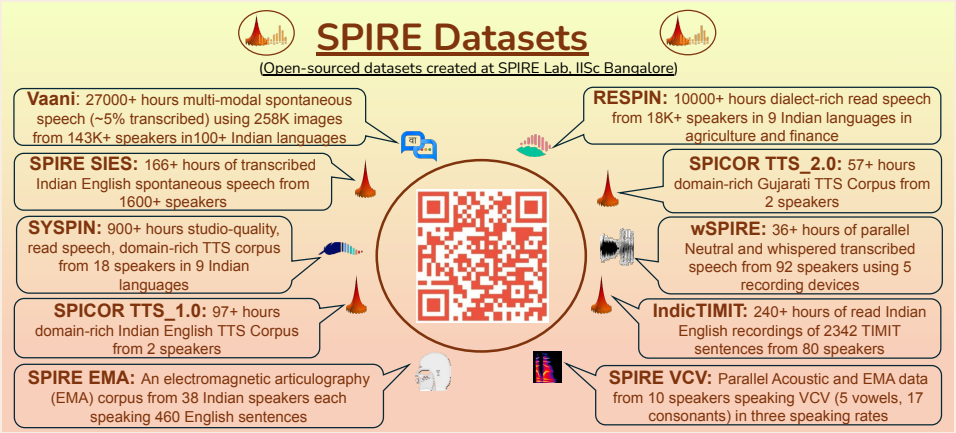
Questions or Suggestions?

Scan for code and resources

Code Repository:
github.com/labspire/respin_did_interspeech25

Contact:
saurabhk0317@gmail.com
spirelab.ee@iisc.ac.in

SPIRE Open-Source Datasets



Scan the QR code or visit **spiredatasets.ee.iisc.ac.in** for more details.