INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 10, October - 2026

DOIs:10.2015/IJIRMF/202610027 --:-- Research Paper / Article / Review

Domain-Specific Noise-Robust Hindi Automatic Speech Recognition Using Fine-Tuned IndicWav2Vec

Dr. David Raju Kolluri[1], Bhuvana Teja Kotti[2]
[1] Associate Professor, HITAM, Hyderabad, India
[2] UG Scholar, Department of Computer Science and Engineering (Data Science), Hyderabad Institute of Technology and Management, Hyderabad, India
Email: davidrajuk.cse@hitam.org, bhuvana16teja@gmail.com

Abstract: *Automatic speech recognition (ASR) systems transform spoken utterances into written text, supporting applications across sectors such as agriculture, financial services, and digital customer assistance. Recognition accuracy may decline when speech contains environmental noise, variations in pronunciation, differences between speakers, or specialized vocabulary. This research examines domain-oriented adaptation of the pretrained IndicWav2Vec model for Hindi speech using noisy recordings from the RESPIN dataset. Independent models were adapted for the Agriculture and Finance domains, with 20,000 utterances allocated for training and 2,500 for validation in each domain; a further 2,500 utterances were used for testing. Performance was assessed through Word Error Rate (WER) and Character Error Rate (CER) on in-domain and cross-domain speech. In the Finance experiments, the adapted system recorded WER and CER values of 25.34% and 11.02%, respectively, while the baseline produced 30.08% and 13.14%. Agriculture results showed a different pattern: CER improved from 13.26% to 12.64%, but WER increased slightly from 29.90% to 30.32%. These findings indicate that adapting Hindi ASR to a specific domain can improve recognition, although the magnitude and consistency of the gains vary across domains.*

Keywords: *Hindi Automatic Speech Recognition, IndicWav2Vec, Domain Adaptation, RESPIN Dataset, Noisy Speech, Word Error Rate, Character Error Rate.*

1. INTRODUCTION:

Computers can process spoken information only when suitable speech-processing methods convert the audio into a form they can interpret. Automatic Speech Recognition (ASR) performs this conversion by generating text from recorded speech. For Hindi-speaking users, this technology could support services involving agricultural information, banking, education, and customer assistance. Its usefulness in these applications depends on whether the words spoken by a user are represented correctly in the output transcript.

Speech recordings differ considerably in their acoustic and linguistic characteristics. A recording may contain unwanted sounds, unclear pronunciation, variations in speaking style, or words associated with a particular profession or activity. These conditions can lead to transcription errors. The challenge becomes especially relevant when the vocabulary in the audio belongs to a specific field rather than everyday conversation. Therefore, the performance of a recognition system needs to be examined using speech that reflects the conditions and terminology of its intended use.

This work focuses on Hindi recordings from two application areas: agriculture and finance. Agricultural speech can include discussions of crops, fertilizers, irrigation, and farming methods, whereas financial speech may concern loans, bank accounts, insurance, and monetary transactions. The two groups differ in the terminology that a recognition system must process. A model adapted using recordings from one group may consequently produce different results when evaluated on the other. Measuring these differences is central to the experiments reported in this paper.

The experiments use IndicWav2Vec, a pretrained model for Indian-language speech processing. Its development builds on advances in learning speech representations from audio, including the wav2vec 2.0 approach. Such pretrained representations provide a basis for subsequent task-specific development. In the present work, the Hindi model is

adapted separately using speech data from the two selected domains. This setup allows the effects of using domain-specific training data to be examined without treating Agriculture and Finance as a single, uniform speech category.
The audio used in the experiments comes from the RESPIN dataset. The RESPIN-S1.0 resource supports research involving Indian speech and linguistic variation. This study uses its noisy Hindi recordings to investigate model behaviour in agricultural and financial speech. The experiments compare the original pretrained system with the corresponding domain-adapted systems and examine performance on both matching and different domains.
Two measures are used to evaluate the generated transcripts: Word Error Rate (WER) and Character Error Rate (CER). WER measures recognition errors at the word level, while CER measures them at the character level. Using WER and CER together helps compare recognition errors at both the word and character levels. The comparison is particularly relevant when adaptation produces different outcomes for the two domains.
The study addresses four specific tasks: preparing the noisy Hindi data for the Agriculture and Finance experiments; adapting the pretrained IndicWav2Vec model separately for each domain; evaluating the adapted systems on in-domain and cross-domain data; and comparing their results using corpus-level WER and CER. These experiments provide an empirical assessment of domain-specific adaptation under the selected noisy-speech conditions.
A domain-selectable ASR application is a possible practical extension of this research. Such an application would allow a user to select a model suited to the intended domain. This should be described as future work unless the application has already been implemented and tested.

2. LITERATURE REVIEW:

Research on automatic speech recognition has explored ways to obtain useful information from audio recordings, including recordings that do not have transcripts available during the initial learning stage. The wav2vec 2.0 approach uses this principle to learn representations from speech audio, which can then be applied to recognition tasks. For speech technologies involving Indian languages, IndicWav2Vec provides a pretrained model that can serve as a starting point for further development.
Apart from the model itself, the choice of speech data also affects the types of conditions that can be investigated. RESPIN-S1.0 contains speech resources representing Indian languages and differences in spoken-language use. Such resources are relevant when examining how speech systems behave across varied linguistic settings. The approaches and resource considered in this study are summarized in Table 1.

Table 1. Overview of Pretrained Speech Recognition Approaches in Related Literature

Approach	Description	Relevance to Speech Recognition
wav2vec 2.0	Learns representations from audio recordings without requiring transcripts during the pretraining stage.	Provides a basis for training and evaluating speech recognition systems.
IndicWav2Vec	Provides pretrained speech representations intended for use with Indian languages.	Offers a starting point for speech recognition experiments involving languages such as Hindi.
RESPIN-S1.0 corpus	Contains speech recordings that reflect linguistic and dialectal diversity in India.	Enables researchers to investigate speech technology using data from varied linguistic conditions.

The present work builds on these existing resources by examining a pretrained Hindi IndicWav2Vec model with noisy recordings from the Agriculture and Finance domains. The model's baseline performance is compared with the results obtained after separate domain-specific adaptation. Evaluation covers both in-domain and cross-domain speech, using Word Error Rate (WER) and Character Error Rate (CER) to measure transcription errors. This experimental comparison investigates whether adapting the model to a particular domain changes recognition performance and whether the changes are consistent across both domains.

3. RESEARCH METHOD:

The experiments were organized to measure how a pretrained Hindi speech recognition model behaves before and after adaptation to two subject areas. The procedure covered data selection, transcript preparation, baseline testing,

independent model training, and evaluation across Agriculture and Finance speech. The following subsections describe the experimental setup and the measures used to compare recognition performance.

3.1 Dataset Description

Noisy Hindi recordings from the RESPIN dataset were selected for the experiments. The selected speech was organized into two groups according to domain: Agriculture and Finance. Each group was used in a separate model-training experiment so that the effect of domain-specific data could be examined independently.

The experiments covered two speech categories: Agriculture and Finance. Each category contributed 25,000 Hindi utterances, giving a combined experimental set of 50,000 utterances. Table 2 shows how these recordings were divided for model development and assessment.

Table 2. Distribution of Speech Samples by Domain

Speech domain	Training set	Validation set	Test set	Total
Agriculture	20,000	2,500	2,500	25,000
Finance	20,000	2,500	2,500	25,000
Overall	**40,000**	**5,000**	**5,000**	**50,000**

The training partition was used to adapt the pretrained recognition model to each subject area. Validation recordings provided a means of checking model performance during training, while the test partition was kept for the final assessment. Both domains followed the same partition sizes.

3.2 Preparation of Speech and Text Data

The dataset metadata was used to link every audio recording with its Hindi transcription. These links were checked before the data was passed to the recognition system, ensuring that the waveform and its reference text were associated correctly.

Several text-handling operations were applied before model processing. Unicode normalization was used to make text encoding consistent, and redundant spaces were standardized. Punctuation was adjusted only where required by the processing pipeline. Hindi text remained in Devanagari script so that the original script was preserved in the training targets and evaluation references.

The preparation workflow included loading the metadata, resolving audio paths, matching recordings to transcripts, checking the available records, and preparing text for the model's tokenizer. Audio processing was carried out with the processor supplied for the pretrained IndicWav2Vec model.

The same preparation approach was followed across the relevant data partitions. Reference transcripts from the test set were kept available so that the system's predictions could be evaluated against the expected text.

3.3 Baseline Assessment

Before adapting the model to either domain, its existing recognition performance was measured on the selected Agriculture and Finance test sets. This initial experiment established a point of comparison for the subsequent domain-specific training experiments.

The baseline system used the pretrained model ai4bharat/indicwav2vec-hindi. Each test recording was processed to obtain a predicted Hindi transcript. The predictions were compared with their corresponding reference transcripts, and WER and CER were computed independently for the two domains.

Table 3. Baseline Recognition Performance

Test domain	WER (%)	CER (%)
Agriculture	29.90	13.26
Finance	30.08	13.14

The values in Table 3 describe recognition performance before the additional training stage. They serve as the baseline for examining the changes observed after adaptation to Agriculture and Finance speech.

3.4 Adaptation for Individual Domains

Two separate model-training runs were performed, beginning from the same pretrained Hindi IndicWav2Vec checkpoint. In the first run, the model was adapted using Agriculture training utterances; in the second, the Finance training partition was used. Separate outputs were maintained for the two runs so that their subsequent evaluation results could be distinguished.

The resulting models are referred to as the Agriculture-adapted model and the Finance-adapted model. The reported training configuration is listed in Table 4.

Table 4. Configuration Used in the Training Experiments

Parameter	Setting
Pretrained checkpoint	ai4bharat/indicwav2vec-hindi
Number of epochs	2
Learning rate	5×10^{-5}
Training batch size	4
Evaluation batch size	4
Precision mode	FP16 mixed precision
Gradient checkpointing	Enabled
Optimizer	AdamW
GPU	NVIDIA Tesla T4
Training strategy	Independent runs for Agriculture and Finance

Both runs used the same reported settings, but their training data differed by domain. Keeping the training runs separate allowed the study to examine the outcomes of adapting the starting model to each subject area individually.

3.5 Evaluation Across Domains

After adaptation, each model was assessed using both the Agriculture and Finance test sets. This resulted in four model–test-set combinations. Table 5 identifies the domain relationship for each combination.

An evaluation is classified as *in-domain* when the test set belongs to the same subject area as the model's adaptation data. It is classified as *cross-domain* when the test set comes from the other subject area.

Table 5. Model and Test-Set Evaluation Matrix

Model	Agriculture test set	Finance test set
Agriculture-adapted model	In-domain	Cross-domain
Finance-adapted model	Cross-domain	In-domain

This evaluation design examines both performance on the domain represented during adaptation and recognition on the alternative domain. Comparing these outcomes helps determine whether the benefits of domain-specific training are retained when the subject matter of the test speech changes.

3.6 Evaluation Measures

Recognition quality was assessed by comparing each predicted transcript with its corresponding reference. Two measures were used: Word Error Rate (WER), which evaluates differences at word level, and Character Error Rate (CER), which evaluates differences at character level.

Word Error Rate (WER)

WER expresses the number of word-level errors relative to the total number of words in the reference transcript. The errors consist of substitutions, deletions, and insertions:

WER = (S + D + I) / N

Here, (S) is the number of substituted words, (D) is the number of deleted words, (I) is the number of inserted words, and (N) denotes the number of words in the reference transcript.

Character Error Rate (CER)

CER uses the same error categories but counts individual characters rather than complete words.

Both metrics were applied to the baseline and adapted systems under the relevant test conditions. Their values were used to compare recognition errors between the two domains and to assess how the results changed following domain-specific adaptation.

4. BLOCK DIAGRAM:

The proposed workflow processes noisy Hindi speech separately for Agriculture and Finance, followed by domain-specific model adaptation and in-domain and cross-domain evaluation.

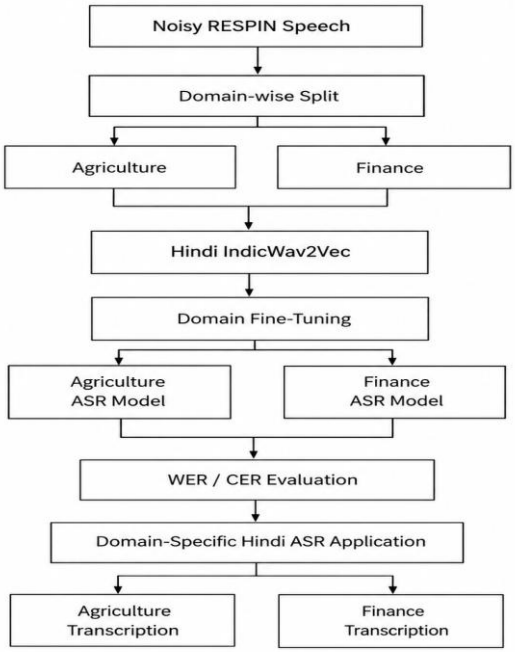


Figure 4.1 : Proposed Domain-Specific Hindi ASR Framework

The framework adapts a common pretrained Hindi ASR model independently to both domains.

5. RESULTS:

Recognition scores were obtained for the pretrained IndicWav2Vec model and both domain-adapted versions. Each system was evaluated on the Agriculture and Finance test partitions using WER and CER. The Hindi transcripts were generated using greedy decoding with connectionist temporal classification (CTC), without incorporating an external language model.

5.1 Comparative Evaluation

Table 6 presents the error measurements for the three systems across the two evaluation domains. The pretrained model provides the reference scores, while the adapted systems show the results obtained after domain-specific training.

Table 6. Error measurements for the evaluated ASR systems

Model	Test domain	Evaluation category	WER (%)	CER (%)
Pretrained IndicWav2Vec	Agriculture	Baseline	29.90	13.26
Agriculture-adapted model	Agriculture	In-domain	30.32	12.64
Finance-adapted model	Agriculture	Cross-domain	29.73	12.58
Pretrained IndicWav2Vec	Finance	Baseline	30.08	13.14
Agriculture-adapted model	Finance	Cross-domain	32.07	13.07
Finance-adapted model	Finance	In-domain	25.34	11.02

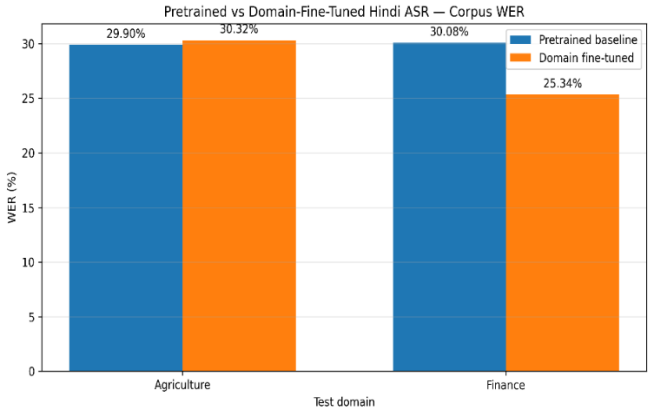


Figure 5.1: WER comparison of the evaluated ASR systems across the Agriculture and Finance test domains.

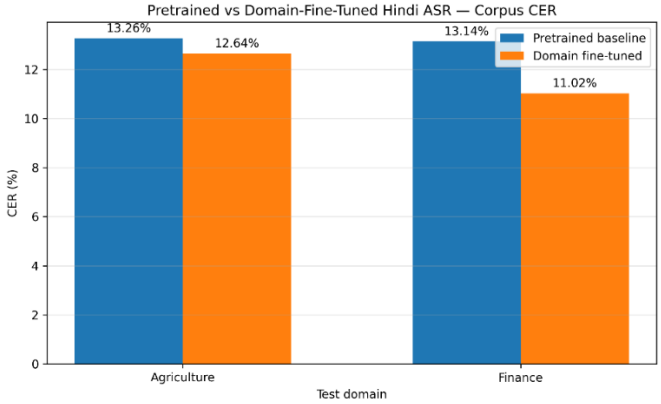


Figure 5.2: CER scores of the three ASR systems on the Agriculture and Finance evaluation sets.

5.2 Agriculture Test Results

The pretrained system produced WER and CER scores of 29.90% and 13.26%, respectively, on Agriculture recordings. The Agriculture-adapted version recorded 30.32% WER and 12.64% CER. The Finance-adapted version returned 29.73% WER and 12.58% CER on the same test partition.

5.3 Finance Test Results

On Finance recordings, the pretrained model achieved 30.08% WER and 13.14% CER. Evaluation of the Agriculture-adapted system yielded values of 32.07% and 13.07% for the two metrics, respectively. For the Finance test data, the model adapted to the Finance domain achieved a WER of 25.34% and a CER of 11.02%.

6. DISCUSSION :

The experiments showed that domain adaptation affected Agriculture and Finance speech recognition differently.

On Finance data, the adapted model reduced WER from 30.08% to 25.34% and CER from 13.14% to 11.02%, improving both evaluation measures. In contrast, Agriculture adaptation lowered CER from 13.26% to 12.64%, but WER increased from 29.90% to 30.32%.

Cross-domain testing also produced different outcomes. The Finance-adapted model achieved 29.73% WER on Agriculture speech, whereas the Agriculture-adapted model recorded 32.07% WER on Finance speech. These results indicate that adaptation benefits depend on the training and evaluation domains.

Overall, evaluating both in-domain and cross-domain performance provides a broader view of how well an adapted ASR model handles different types of speech.

7. CONCLUSION:

This research evaluated domain-specific adaptation of a pretrained Hindi speech recognition model using noisy Agriculture and Finance recordings from the RESPIN dataset. The experiments revealed different performance patterns across the two domains.

On Finance test recordings, the adapted model produced error rates of 25.34% WER and 11.02% CER, compared with the pretrained model's corresponding values of 30.08% and 13.14%. For Agriculture, CER fell from 13.26% to 12.64%, while WER rose from 29.90% to 30.32%. Recognition scores differed when each adapted model was evaluated on recordings from the other domain. Future work can investigate additional Indian languages, improved decoding methods, and techniques for handling different noise conditions.

8. LIMITATIONS:

The scope of this work is confined to Hindi recordings from the Agriculture and Finance categories of the RESPIN dataset. Since other languages, subject areas, and recording conditions were not examined, the findings should not automatically be extended to those settings.

Each domain-specific model was trained on 20,000 utterances for two epochs. Other choices of training duration, data volume, or adaptation method could lead to different recognition outcomes. In particular, the Agriculture experiment yielded a lower CER but a slightly higher WER, showing that the two evaluation measures did not change in the same direction.

The experiments used only greedy CTC decoding; other decoding options and language-model assistance were not tested. In addition, the evaluation concentrated on transcription accuracy; performance under high user loads and in large-scale deployment was not assessed.

9. FUTURE SCOPE:

Future experiments could test the system with more training data, longer training runs, and alternative methods for handling noisy speech. The effects of different decoding techniques on transcription accuracy could also be measured.

The work could be expanded beyond Hindi to cover other Indian languages and subject areas. Testing with varied speakers, background sounds, and recording conditions would provide further evidence of how reliably the models perform in different settings. The models could also be assessed in practical voice-based tasks related to agriculture and finance.

REFERENCES:

1. Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. https://arxiv.org/abs/2006.11477
2. Javed, T., Doddapaneni, S., Raman, A., Bhogale, K. S., Ramesh, G., Kunchukuttan, A., Kumar, P., and Khapra, M. M. (2022). Towards building ASR systems for the next billion users. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 10813–10821. https://doi.org/10.1609/aaai.v36i10.21327
3. Kumar, S., et al. (2025). RESPIN-S1.0: A read speech corpus of 10000+ hours in dialects of nine Indian languages. *Advances in Neural Information Processing Systems, 38*. https://doi.org/10.52202/085713-4928
4. Graves, A., Fernández, S., Gomez, F. J., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369–376). https://doi.org/10.1145/1143844.1143891
5. AI4Bharat. (n.d.). *IndicWav2Vec*. https://indicnlp.ai4bharat.org/indicwav2vec/