Beyond Common Words: Enhancing ASR Cross-Lingual Proper Noun Recognition Using Large Language Models

Rishabh Kumar
krrishabh@cse.iitb.ac.in
IIT Bombay

Sabyasachi Ghosh
ssgosh@gmail.com
IIT Bombay

Ganesh Ramakrishnan
ganesh@cse.iitb.ac.in
IIT Bombay

Abstract

In this work, we address the challenge of cross-lingual proper noun recognition in automatic speech recognition (ASR), where proper nouns in an utterance may originate from a language different from the language in which the ASR system is trained. We enhance the performance of end-to-end ASR systems by instructing a large language model (LLM) to correct the ASR model's predictions. The LLM's context is augmented with a dictionary of cross-lingual words that are phonetically and graphemically similar to the potentially incorrect proper nouns in the ASR predictions. Our dictionary-based method DiP-ASR (Dictionary-based Prompting for Automatic Speech Recognition) significantly reduces word error rates compared to both the end-to-end ASR baseline and instruction-based prompting of the LLM without the dictionary across cross-lingual proper noun recognition tasks involving three secondary languages.[1]

1 Introduction

In recent years, automatic speech recognition (ASR) systems have advanced significantly for resource-rich languages, largely due to the availability of extensive labeled speech datasets (Chan et al., 2015; Baevski et al., 2020; Radford et al., 2023). However, such ASR systems often still struggle with words which are rare in the training data or missing from it (Lux and Vu, 2021). Proper nouns, in particular, pose these challenges, leading to frequent misinterpretations by ASR systems (Peyser et al., 2020). These issues are further exacerbated in a cross-lingual setting – in which at inference time most of the words of the spoken sentences are from the language on which the ASR system is trained, but the proper nouns may be from a secondary language, leading to a large number of out of vocabulary words.

[1]Source code and datasets are present at https://github.com/cyfer0618/DiP-ASR

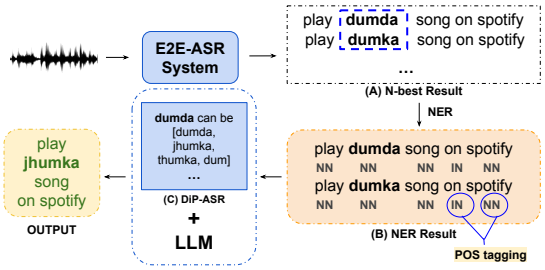

Figure 1: Overview of the DiP-ASR method into LLM.

Accurate recognition of proper nouns is crucial for the effectiveness and usability of ASR systems in real-world applications. Consider the scenario of a voice command: "play jhumka song on spotify", where 'jhumka' is a proper noun in Hindi and the rest of the command is in English. In this instance, the recognition of the specific song title 'jhumka' is critical for user satisfaction. However, if the ASR system is not familiar with 'jhumka' due to its absence in the training corpus, it may misinterpret it as 'thumka' or 'dumka' as shown in Figure 1. Such misinterpretations lead to incorrect outputs, potentially playing entirely different songs, thereby diminishing the system's accuracy and user trust. This exemplifies the need for more robust handling of cross-lingual proper nouns, given their significance in many real-world voice-activated tasks.

1.1 Background and Related Work

We briefly discuss ASR systems and the existing approaches to improving proper noun recognition in them. While traditional ASR systems relied on separate components like the acoustic model (AM), language model (LM), and pronunciation model (PM)(Yu and Deng, 2016; Adiga et al., 2021; Kumar et al., 2022b), recent developments such as end-to-end automatic speech recognition (E2E-ASR) systems (Chan et al., 2016; Liu et al., 2020; Baevski et al., 2020), have merged these components into a single system, offering more efficient

Findings of the Association for Computational Linguistics: EACL 2024, pages 6821–6828
November 12-16, 2024 ©2024 Association for Computational Linguistics

and accurate speech recognition. These advancements have significantly lowered Word Error Rates (WER) for major languages like English but have also highlighted challenges, particularly in recognizing proper nouns (Dutta et al., 2022). One significant limitation is the absence of a PM to incorporate detailed phonetic knowledge into the system, making it difficult to achieve correct proper noun pronunciation. Moreover, the training corpus for these systems often lacks a comprehensive representation of proper nouns, leading to limited exposure (Laurent et al., 2014).

Several approaches have been attempted to enhance proper noun recognition in E2E-ASR systems (Peyser et al., 2020). While integrating a language model (LM) trained on billions of words has shown limited effectiveness in handling proper nouns, we explore the use of large language models (LLMs) for this task without specific training (Fathullah et al., 2024). Despite being trained on massive datasets, LLMs struggle with cross-lingual proper nouns (Adelani et al., 2024). In this paper, we demonstrate that instruction-based prompting of LLMs not only provides hints or options to the model but also aids in correcting proper nouns.

Recently, the combination of E2E-ASR systems with LLMs has shown promising results in error correction (Chelba et al., 2012; Fathullah et al., 2024). Most approaches to error correction follow a similar architecture: they take the N-best hypotheses from the E2E-ASR system and use LLMs to refine the ASR output. The N-best hypotheses serve as cues, helping LLMs identify the correct tokens and improve overall recognition accuracy. However, this method is insufficient for cross-lingual proper noun recognition because it does not provide LLMs with specific information about the proper nouns.

1.2 Our Approach

In this study, we assume access to cross-lingual proper noun knowledge in the form of a vocabulary list and demonstrate how to prompt the LLMs with hints that specify a set of possible proper nouns. Inspired by machine translation models that use dictionaries to improve translation (Ghazvininejad et al., 2023; Maheshwari et al., 2022), we show how to incorporate cross-lingual proper noun knowledge into LLMs without additional model training.

We propose a novel framework, **DiP-ASR** (**Di**ctionary-based **P**rompting for **A**utomatic **S**peech **R**ecognition), and present extensive experiments that show significant improvements in cross-lingual proper noun recognition. Our method seamlessly integrates LLMs directly into the E2E-ASR system, complemented by a dedicated cross-lingual proper noun dictionary (Debess et al., 2019; Higuchi et al., 2023). This integration creates a synergistic effect, optimizing the recognition of cross-lingual proper nouns while maintaining the broader efficiencies of the ASR system. Furthermore, LLMs have shown considerable promise in addressing cross-lingual proper noun recognition challenges, even in few-shot or zero-shot scenarios (Brown et al., 2020; Higuchi et al., 2023). Additionally, we explore the roles of teacher forcing and non-teacher forcing (Lamb et al., 2016) in few-shot prompting with LLMs which is discussed in detail in Section 3.3. In this work, we have four key contributions toward enhancing E2E-ASR systems for the correct recognition of cross-lingual proper nouns:

- We propose a novel framework called DiP-ASR, which utilizes cross-lingual proper noun dictionaries to prompt LLMs for proper noun recognition.
- We empirically demonstrate that instruction-based prompting improves cross-lingual proper noun error correction.
- We comprehensively analyze our method under various instructional settings, including zero-shot, one-shot, and few-shot learning scenarios, both with and without teacher forcing.
- We introduce a novel cross-lingual proper noun recognition dataset encompassing Hindi, Tamil, Telugu languages.

2 Methodology

Prompting LLMs for cross-lingual proper noun recognition assumes that the LLM's pre-training dataset includes a substantial number of cross-lingual proper nouns. However, this is not the case for all languages, such as some Indian languages. LLMs are predominantly pre-trained on datasets that contain mainly English proper nouns with limited representation of cross-lingual proper nouns. As a result, LLMs struggle to accurately recognize and correct cross-lingual proper nouns.

To address this challenge, we introduce two methods, Instruction-based Prompting and Prompting with Dictionary.

Prompt template
System: You are an Automatic Speech Recognition error correction tool ### User: Perform error correction on the top 5 output generated by an Automatic Speech Recognition (ASR) System. The ASR hypothese are : hypothesis 1: 'launch ward dumda song unsportified' hypothesis 2: 'launch word dumka song unsportified' hypothesis 3: 'launch wad dumda song unsportified' hypothesis 4: 'launch wat dumda song unsportified' hypothesis 5: 'launch wat drum song unsportified' Dictionay: 'unsportified' can be ['unsportified', 'spotify', 'rendu'], 'ward' can be ['ward', 'word', 'what'], 'dumda' can be ['dumda', 'jhumka', 'thumka', 'dum'], 'wad' can be ['wad', 'what', 'wat', 'ward', 'word'] ... Please provide only one corrected ASR hypothesis for the given utterance (hypothesis) using Dictionary, do not add any extra word or explanations.
Ground Truth: launch what jhumka song on spotify
0-Shot: launch ward dumda song unsportified
1-Shot: launch word dumka song unsportified
Few-Shot (TF): launch ward jhumka song on spotify
Few-shot (N-TF): launch word jhumka song on spotify

Table 1: Illustration of an example for the Hindi language using a master dictionary, presented by Ground Truth, Zero-Shot, One-Shot, Few-Shot (Teacher Forcing), and Few-Shot (Non-Teacher Forcing) results.

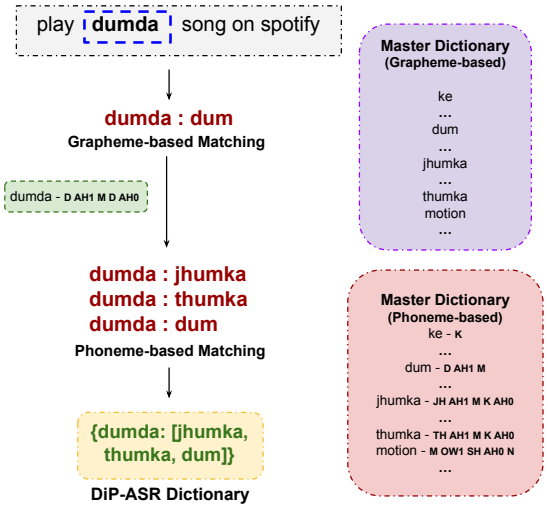

Figure 2: Detailed construction of the **DiP-ASR Dictionary**: This figure shows the process of matching grapheme and phoneme representations from the Master Dictionary to aid in correcting proper noun errors.

2.1 Instruction-based Prompting

Instruction-based prompting is a well-established method for guiding responses from LLMs. In this work, we utilize a two-step instruction process. In the first step, the LLM is instructed to identify the correctness of cross-lingual proper nouns in the E2E-ASR N-best list of predictions. In the second step, the LLM is asked to correct the incorrect proper nouns, based on the N-best list. Despite this approach, instruction-based prompting LLMs exhibit limited proficiency in correcting cross-lingual proper nouns.

2.2 Prompting with Dictionary

We introduce a method called DiP-ASR, which directly incorporates information from cross-lingual dictionaries in the prompt. These dictionaries comprise of proper nouns from various languages. The E2E-ASR N-best list vocabulary is mapped to the proper noun list both phonetically and graphemically. This mapped vocabulary is then appended to the prompt as a 'Dictionary', as illustrated in Table 2.1. Details of mapped vocabulary creation are given in Sec. 3.3.

In Figure 2, DiP-ASR construction involves creating a comprehensive dictionary that includes both grapheme-based and phoneme-based entries for cross-lingual proper nouns. This dictionary is utilized to augment the context of a large language model (LLM), helping to correct potentially incorrect proper nouns in the ASR model's predictions by matching them with phonetically and graphemically similar words. While the DiP-ASR method traditionally relies on cross-lingual dictionaries, we mitigate this dependency by leveraging transliteration. We create a dictionary in the native language and transliterate it into the target language. For instance, words in devanagari script (Hindi) are transliterated into roman script (English) with a model that achieves 97% accuracy (Gala et al., 2023). This approach significantly reduces the effort required to develop cross-lingual dictionaries while maintaining high-quality entries. Additionally, if a phonemizer for a language is unavailable, grapheme-based dictionaries can be employed as an alternative.

3 Experimental Setup

For our experiments, we utilized the pre-trained wav2vec2-large-xlsr-53 model (Baevski et al., 2020) from Huggingface, fine-tuning it with the English CommonVoice corpus V6.1 for both training and validation. During decoding, the E2E-ASR model employed a beam size of 5, generating a 5-best list of hypotheses for each utterance, referred to as the 'N-best' hypotheses. Minimal post-processing was applied, with each hypothesis converted to lowercase without further modifications. The effectiveness of these N-best hypotheses is detailed in the Section 4.

Language	Hindi$_S$	Hindi$_C$	Tamil$_S$	Tamil$_C$	Telugu$_S$	Telugu$_C$
E2E-ASR baseline	57.53	45.46	50.40	50.14	56.76	68.65
0-Shot	59.53	40.78	49.60	47.32	58.69	72.16
1-Shot	53.85	36.86	52.38	41.97	55.21	65.68
Few-Shot (TF)	50.50	31.45	49.20	36.90	57.14	61.62
+ Instruction	49.23	30.34	44.32	34.86	54.19	50.21
+ Dictionary	**46.83**	29.94	**40.46**	**32.21**	52.38	**40.32**
Few-Shot (N-TF)	51.83	31.45	46.82	40.00	53.67	62.16
+ Instruction	50.24	30.9	46.24	38.35	52.34	60.82
+ Dictionary	48.84	**26.26**	45.22	32.78	**50.29**	57.73

Table 2: WER performance for the baseline E2E-ASR baseline and error correction results with different methods. (TF - Teacher Forcing, N-TF - Non-Teacher Forcing and Datasets: Hindi$_S$ - Hindi Simple, Hindi$_C$ - Hindi Complex, Telugu$_S$ - Telugu Simple, Telugu$_C$ - Telugu Complex, Tamil$_S$ - Tamil Simple, Tamil$_C$ - Tamil Complex). Bold refers to best performance and underline refers to best performance without dictionary (DiP-ASR)

3.1 Large Language Model (LLM)

In the course of our study, we used OpenAI's ChatGPT, specifically the variant gpt-3.5-turbo-0613[2]. By employing this model, we aim to assess the potential of basic LLM in enhancing E2E-ASR performance, especially in the addressing recognition of cross-lingual proper nouns correction.

3.2 Datasets and Dictionaries

Testing and Prompting Dataset In our study, we curated distinct datasets for prompts and test utterances in Hindi, Telugu, and Tamil, focusing on low log probabilities generated by the E2E-ASR system, which indicated high WER. The test and prompt datasets, each comprising 100 utterances, were divided into 50 simple utterances and 50 complex utterances. Simple utterances included only the command and the name of the song, characterized by one or two cross-lingual proper nouns. In contrast, complex utterances incorporated both the song name and the specific platform on which the song should be played, containing more than two cross-lingual proper nouns.

Cross-Lingual Proper Noun Dictionaries To enhance proper noun recognition, we utilized language-specific dictionaries focused on proper nouns.

3.3 Prompting Formulation

For processing an utterance, our pipeline involves three key components. A) First, the E2E-ASR system generates the top N-best hypotheses for the given utterance. B) Next, we apply Named Entity Recognition (NER) to these hypotheses to identify proper nouns. C) These identified proper nouns are then matched against a specialized dictionary that includes both grapheme and phoneme representations. These graph and phoneme representations are created using a phoneme conversion dictionary powered by a transformer model trained on the Connectionist Temporal Classification (CTC) algorithm (Graves et al., 2006), and an autoregressive grapheme-to-phoneme (G2P) model (Yolchuyeva et al., 2020) pre-trained with the CMU pronunciation dictionary[3] and the NetTalk dataset (Chen et al., 2003) as shown in Figure 1. To furnish hints for error correction, we search for each proper noun tagged by NER in the Master Dictionary, employing both grapheme and phoneme matching with a tolerance of an edit distance up to 2, ensuring even phonetically similar terms are considered. This approach allows for a comprehensive understanding and correction of proper nouns by leveraging both spelling and sound. Furthermore, optionally, hand-crafted examples of correct responses on k utterances are also provided in the prompt, making it a k-shot setting.

4 Results and Analysis

Our experiments demonstrate that the few-shot learning approach significantly outperforms the zero-shot and one-shot methods shown in Table 2. Teacher forcing further enhances performance, particularly for complex utterances in Hindi$_C$, Tamil$_C$, and Telugu$_C$, as well as for simple utterances in Hindi$_S$. This improvement is largely due to teacher forcing's ability to handle the complexity of these utterances effectively. For example, the WER for Hindi$_S$ utterances under teacher forcing shows substantial improvement, though the CER remains consistent at 27.358 for both teacher forcing and non-teacher forcing approaches. However, in the case of Telugu$_S$ and Tamil$_S$ utterances, the WER gap between teacher forcing and non-teacher forcing is more pronounced. This discrepancy is likely due to the longer word lengths in these languages, which often result in extra, incorrect words generated by the ASR system that teacher forcing cannot effectively mitigate. These findings highlight the necessity of using different approaches for different scenarios to achieve optimal performance in ASR cross-lingual proper noun recognition.

Furthermore, employing few-shot learning

[2]This particular snapshot of the ChatGPT model was captured on 18 December 2023

[3]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

with both instruction-based and DiP-ASR approaches significantly enhanced the results. While instruction-based prompting effectively corrected common words frequently appearing in the N-best list, it fell short in accurately correcting proper nouns. Therefore, incorporating an external source became essential for the precise correction of proper noun errors. Our approach leveraged a comprehensive dictionary loaded with a vast array of proper nouns specific to each language as discussed in Section 3.2. By utilizing both grapheme-based and phoneme-based dictionaries, we aligned the hypothesized words with those in the Master Dictionary, aiding the LLM in accurately predicting the correct word. As shown in Table 2, the few-shot learning method augmented with dictionary support emerged as the most effective, showing an improvement range from 7% to 30%.

In Table 2.1, we present a case study on correcting cross-lingual proper nouns. The table illustrates the outputs of the E2E-ASR system. Additionally, the dictionary includes all potential replacement words for those in the hypotheses identified with an NN tag by the NER system. The zero-shot method made no corrections and simply returned hypothesis 1 as the output. Similarly, the one-shot method failed to assist the LLM in correcting proper nouns. In contrast, the few-shot method effectively guided the LLM in identifying and rectifying errors, improving output accuracy, and correctly identifying the song.

4.1 Role of N-best hypotheses

Our evaluation revealed that the system's performance notably improved with the use of the N-best list, which was derived solely from the E2E-ASR system's hypotheses. Incorporating the N-best list led to a 3.5% improvement in Word Error Rate (WER), underscoring its critical contribution to the observed enhancement. The N-best list proved especially vital for the Large Language Model (LLM) to excel in zero-shot prompts, although its influence on ChatGPT's performance was less pronounced. Furthermore, the N-best list expanded the dictionary's search scope, offering a broader range of potential words for prediction. This benefit was also evident in the performance boost seen with the instruction-based method, highlighting the N-best list's overarching value in refining the system's accuracy.

Dataset	0-shot	2-shot	4-shot	8-shot
Hindi_S	59.53	52.54	63.92	58.9
Hindi_C	40.78	39.91	39.29	43.83
Tamil_S	49.6	49.44	44.83	50.52
Tamil_C	47.32	47.73	41.5	43.18
Telugu_S	58.69	55.98	54.56	62.85
Telugu_C	72.16	65.04	69.91	65.53
Average	54.68	51.77	52.33	54.14

Table 3: Word Error Rates (WER) for Cross-Lingual Proper Noun Recognition across Different Shot Scenarios.

4.2 Role of Few-Shot Learning

Table 3 presents the Word Error Rates (WER) for six test datasets, covering both simple and complex utterances in Hindi, Tamil, and Telugu, across zero-shot to eight-shot learning scenarios. The results demonstrate the impact of increasing the number of shots on the accuracy of proper noun recognition. Notably, the two-shot average WER is lower than the eight-shot WER, highlighting its cost-effectiveness. Therefore, we used two-shot as a few-shot in the paper.

5 Conclusion

We propose an instruction-based prompting method that enhances the LLM's ability to correct proper nouns, with effectiveness varying based on the pre-trained dataset of the LLMs. Additionally, we introduce a novel approach for incorporating cross-lingual proper noun knowledge into LLMs. Our method, DiP-ASR, shows significant improvements in cross-lingual proper noun recognition. Our analyses of DiP-ASR highlight its benefits and limitations under various conditions.

Limitations

A limitation of our current method (DiP-ASR) is that if a non-popular cross-lingual proper noun is present in an utterance, the LLMs may produce an incorrect output by substituting it with a more commonly known proper noun, even when the E2E-ASR system provides the correct result. For example, when the ASR system correctly transcribes "Preethi," but the LLM substitutes it with the more common name "Preeti," demonstrating the limitation of the current dictionary's coverage and prompting strategy.

Acknowledgments

We thank S. Ganesh Janakiraman, Abhiram Naidu, and Swaroop R. Ghag for their invaluable contribu-

tion to creating the dataset. We also deeply appreciate Allenki Akshay and Anindya Mitra for their supportive roles throughout this work.

References

David Ifeoluwa Adelani, A Seza Doğruöz, André Coneglian, and Atul Kr Ojha. 2024. Comparing llm prompting with cross-lingual transfer performance on indigenous and low-resource brazilian languages. *arXiv preprint arXiv:2404.18286*.

Devaraja Adiga, Rishabh Kumar, Amrith Krishna, Preethi Jyothi, Ganesh Ramakrishnan, and Pawan Goyal. 2021. Automatic speech recognition in sanskrit: A new speech corpus and modelling insights. *arXiv preprint arXiv:2106.05852*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. Large scale language modeling in automatic speech recognition. *arXiv preprint arXiv:1210.8440*.

Stanley F Chen et al. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *INTERSPEECH*, pages 2033–2036.

Iben Nyholm Debess, Sandra Saxov Lamhauge, and Peter Juel Henrichsen. 2019. Garnishing a phonetic dictionary for asr intake. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 395–399.

Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Souvik Pal, Ganesh Ramakrishnan, and Preethi Jyothi. 2022. Error correction in asr using sequence-to-sequence models. *arXiv preprint arXiv:2202.01157*.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Yosuke Higuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. 2023. Harnessing the zero-shot power of instruction-tuned large language model in end-to-end speech recognition. *arXiv preprint arXiv:2309.10524*.

Rishabh Kumar, Devaraja Adiga, Mayank Kothyari, Jatin Dalal, Ganesh Ramakrishnan, and Preethi Jyothi. 2022a. Vagyojaka: An annotating and post-editing tool for automatic speech recognition. In *INTERSPEECH*, pages 857–858.

Rishabh Kumar, Devaraja Adiga, Rishav Ranjan, Amrith Krishna, Ganesh Ramakrishnan, Pawan Goyal, and Preethi Jyothi. 2022b. Linguistically informed post-processing for asr error correction in sanskrit. In *INTERSPEECH*, pages 2293–2297.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29.

Antoine Laurent, Sylvain Meignier, and Paul Deléglise. 2014. Improving recognition of proper nouns in asr through generating and filtering phonetic transcriptions. *Computer Speech & Language*, 28(4):979–996.

Qi Liu, Zhehuai Chen, Hao Li, Mingkun Huang, Yizhou Lu, and Kai Yu. 2020. Modular end-to-end automatic speech recognition framework for acoustic-to-word model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2174–2183.

Florian Lux and Ngoc Thang Vu. 2021. Meta-learning for improving rare word recognition in end-to-end asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5974–5978. IEEE.

Language	Total	Prompt	Test
Hindi	942	100	100
Tamil	827	100	100
Telugu	942	100	100

Table 4: Distribution of Total, Prompt, and Test utterances for cross-lingual datasets in Hindi, Tamil, and Telugu based on low log probabilities of the E2E-ASR system.

Ayush Maheshwari, Piyush Sharma, Preethi Jyothi, and Ganesh Ramakrishnan. 2022. Dictdis: Dictionary constrained disambiguation for improved nmt. *arXiv preprint arXiv:2210.06996*.

Cal Peyser, Tara N Sainath, and Golan Pundak. 2020. Improving proper noun recognition in end-to-end asr by customization of the mwer loss criterion. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7789–7793. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2020. Transformer based grapheme-to-phoneme conversion. *arXiv preprint arXiv:2004.06338*.

Dong Yu and Lin Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

A Appendix

A.1 Role of Proper Noun on ASR

Proper nouns carry specific meanings, contexts, and user intents, particularly in voice-activated tasks like personal assistants, navigation systems, and voice search. For instance, in a voice command to play a specific song, recognizing the song title (a proper noun) accurately is essential for delivering the intended result and ensuring a positive user experience.

A.2 Dataset Information

In Table 4, The dataset consisted of 600 unique utterances selected from a larger set of 2711 challenging utterances, with 200 utterances each from three speakers across the three languages. These utterances primarily represent user requests directed at personal assistants, predominantly for playing songs in a specific language.

The comprehensive dictionaries were compiled from various sources, including the IMDb database, Spotify database, Kaggle Bollywood songs[4] .

Algorithm 1 Algorithm for DiP-ASR

Require:

Audio utterance U

ASR system M

Master dictionary D containing grapheme and phoneme representations of proper nouns

Named Entity Recognition (NER) model NER

Grapheme-to-Phoneme (G2P) model $G2P$

Edit distance threshold δ

Ensure:

$H = \{h_1, h_2, \ldots, h_N\} \leftarrow M(U)$

$P_{NER} = NER(H)$

$P_{phoneme} = G2P(p), \quad P_{grapheme} = p$

$$\text{Match}(p, D) = \begin{cases} \text{EditDistance}(P_{phoneme}, d_{phoneme}) \leq \delta \\ \text{EditDistance}(P_{grapheme}, d_{grapheme}) \leq \delta \end{cases}$$

Prompt $= \{H, D\}$

$T_{corrected} \leftarrow LLM(\text{Prompt})$

This algorithm outlines a post-processing method to correct proper noun errors in ASR outputs using DiP-ASR method. The process begins with the ASR system generating multiple hypotheses (N-best hypotheses) from an audio utterance. These hypotheses are then passed through an NER model to extract named entities, such as proper nouns. For each detected entity phoneme representation is generated using the G2P model and grapheme representation directly uses the spelled-out form of each proper noun. These representations are then compared to entries in a master dictionary of known proper nouns, where the algorithm checks if the edit distance between the ASR output and dictionary entries is below a given threshold. If a match is found, the information is compiled into a prompt for the LLM, which makes the final corrections, ensuring the ASR transcript accurately reflects proper noun usage.

A.3 Role of Data Contributors

The dataset was created by six dedicated contributors. Each contributor followed specific guidelines to ensure the consistency and quality of the recordings. They were instructed to clearly articulate

[4]https://www.kaggle.com/datasets/rishabjn/bollywoodsongs

their commands in a quiet environment to minimize background noise and ensure high-quality audio samples.

Sources of Tools used for Recording, Cleaning and Transcribing the Audios

ASR Voice Recorder: https://play.google.com/store/apps/details?id=com.nll.asr

Vagyojaka(Kumar et al., 2022a): https://www.cse.iitb.ac.in/~asr/Vagyojaka/