Exploring the Effect of Dialect Mismatched Language Models in Telugu Automatic Speech Recognition

Aditya Yadavalli **Ganesh S Mirishkar** **Anil Kumar Vuppala**

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad
Gachibowli, Hyderabad, Telangana, 500032
{aditya.yadavalli, mirishkar.ganesh}@research.iiit.ac.in
{anil.vuppala}@iiit.ac.in

Abstract

Previous research has found that Acoustic Models (AM) of an Automatic Speech Recognition (ASR) system are susceptible to dialect variations within a language, thereby adversely affecting the ASR. To counter this, researchers have proposed to build a dialect-specific AM while keeping the Language Model (LM) constant for all the dialects. This study explores the effect of dialect mismatched LM by considering three different Telugu regional dialects: Telangana, Coastal Andhra, and Rayalaseema. We show that dialect variations that surface in the form of a different lexicon, grammar, and occasionally semantics can significantly degrade the performance of the LM under mismatched conditions. Therefore, this degradation has an adverse effect on the ASR even when dialect-specific AM is used. We show a degradation of up to 13.13 perplexity points when LM is used under mismatched conditions. Furthermore, we show a degradation of over 9% and over 15% in Character Error Rate (CER) and Word Error Rate (WER), respectively, in the ASR systems when using mismatched LMs over matched LMs.

1 Introduction

Automatic Speech Recognition (ASR) systems are rapidly becoming part of our everyday lives through voice assistants such as Siri, Alexa, and Google Assistant. Since these voice assistants can now perform various day-to-day tasks exceedingly well, they have now become an integral part of many devices such as phones, televisions, music players, and smartwatches.

Accurate and reliable ASR systems for Indian languages would have a significant impact due to two reasons: Firstly, India is home to many languages and dialects. Many of these languages and dialects do not have a written form. Secondly, a considerable amount of the population in India cannot read or write, as evidenced by the low literacy rates.[1] This leaves such people with only one mode of communication – the spoken form.

Despite the advances made by spoken technology research in recent years, dialect or accent variation proves to be a huge challenge.[2] Huang et al. (2001) show that accent variation contributes most to speech variability after gender. Biadsy et al. (2012); Elfeky et al. (2018) show the amount of degradation in ASR performance when it is not trained on dialect-specific data. Therefore, currently, state-of-the-art systems, including those of Google and Microsoft, use dialect-specific ASR systems.

However, multi-dialect ASR is an attractive solution in scenarios where sufficient dialect-specific data or information is not available. Therefore, Liu and Fung (2006); Rao and Sak (2017); Jain et al. (2018); Yang et al. (2018); Fukuda et al. (2018); Jain et al. (2019); Viglino et al. (2019); Li et al. (2018); Deng et al. (2021) attempt to improve multi-dialect ASR systems.

Liu and Fung (2006) use *auxiliary accent trees* to model Chinese accent variation. These are decision trees that model accent-specific triphone units and have a similar function as the decision trees that are used for state-tying of standard triphones. Rao and Sak (2017) show that grapheme-based Recurrent Neural Network-Connectionist Temporal Classification (RNN-CTC) ASR models outperform their phoneme-based counterparts when trained and used in multi-dialect English conditions. Furthermore, they study modelling phoneme recognition as an auxiliary task to im-

[1] https://censusindia.gov.in/2011-prov-results/data_files/mp/07Literacy.pdf

[2] In this paper, we use the words, 'dialect' and 'accent' interchangeably. However, we make one important distinction between dialect and accent: accent differences are largely constrained to the spoken form while dialect differences are not.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 292 - 301

Dialect	Sentence
Coastal Andhra	ప్రతి పౌరుడు ఓటు తప్పక వేయాలండి
Rayalaseema	మాకు మా పల్లెటూరు అంటే చానా ఇష్టము
Telangana	గా ఫుట్బాల్ గురించి అయితే నాకు మస్త్ గా తెలుసు రా బ్రె

Table 1: Sentences of Different Dialects Taken from the Dataset

prove grapheme recognition and show improved performance when tested on multiple English dialects. Yang et al. (2018); Jain et al. (2018) explore the benefits of learning an accent classifier and multi-accent acoustic model under a multi-task learning framework. Viglino et al. (2019) explore incorporating various accent embeddings into a multi-accent End-to-End ASR model. All of these multi-accent studies report significant relative Word Error Rate improvements in their ASR models on various English accents. Li et al. (2018) incorporate dialect-specific information at the acoustic feature and textual level into multi-dialect End-to-End ASR and report that such a model outperforms dialect-specific End-to-End ASR systems. Zhang et al. (2021) propose a Transformer-based (Vaswani et al., 2017) encoder to simultaneously detect the dialect and transcribe an audio sample. More recently, with increased interest in self-supervised learning, Deng et al. (2021) explored self-supervised learning techniques to predict the accent from speech and use the predicted information to train an accent-specific self-supervised ASR. They report that such a model significantly outperforms an accent-independent ASR system.

Many researchers have previously studied the effects of dialect mismatched acoustic models in ASR systems. However, to the best of our knowledge, we are the first to explore the effects of a dialect mismatched Language Model (LM) in ASR systems.

Our language of interest in this paper is Telugu. Telugu is a South Central Dravidian language primarily spoken in two states of India: Telangana, and Andhra Pradesh. As previously mentioned, low literacy states in these states has motivated researchers to build Telugu ASR systems (Srivastava et al., 2018; Diwan and Jyothi, 2020; Bhanuprasad and Svenson, 2008; Vegesna et al., 2017; Diwan et al., 2021). However, they largely concentrate on building ASR systems for "standardised" Telugu. While Mirishkar et al. (2021b) collect dialect-specific Telugu data, they do not conduct any ASR experiments on individual Telugu dialects. We conduct our experiments on three regional Telugu dialects, i.e., Telangana, Rayalaseema, and Coastal Andhra. A considerable portion of dialect variation in Telugu can be seen in the lexicon, grammar, and occasionally semantics. Additionally, since Indian languages are considered to be low-resource in nature, adding external text to the LM is a solution that has gained interest (Pham et al., 2020; Karpov et al., 2021; Mirishkar et al., 2021a; Klejch et al., 2021). While such a method has shown significant benefits in their ASR systems, we argue that if proper care is not taken in matching the dialect of the external text with that of the ASR, it could lead to degradation in performance. These are the primary motivations for us to conduct this study. To this effect, the following are the major contributions of the paper:

- We show significant degradation of the perplexity scores of the LMs when tested on a different Telugu dialect.
- We use these LMs in a dialect mismatched ASR and report degradation of over 15% WER in such a setting compared to matched setting.

The rest of the paper is organised as follows. In Section 2, a brief description of the three dialects used in this study is given. In Section 3, we describe the dataset used in the study. In Section 4 and Section 5, we describe our experimental setup and discuss results under matched and mismatched settings. We conclude the paper with Section 6 and discuss possible future directions.

2 Telugu Dialects

All Telugu dialects can be broadly classified into three regional dialects: Telangana, Coastal Andhra, and Rayalaseema. The formation of these dialects is primarily due to the influence of neighboring states, and the regional culture (Mannepalli

et al., 2016). The Nizams ruled the Telangana region, whose official languages were Persian and Urdu. Thus, one can see the influence of Urdu with many nativised Urdu words present in Telangana (Ithagani, 2014). Here are some such examples: కాకా, జాగా, దవాఖానా .[3] There is also some influence of the neighboring states' languages like Kannada on Telangana. The Coastal Andhra dialect is largely influenced by Sanskrit as well as Tamil due to historical and geographical reasons (Shivaprasad and Sadanandam, 2020). Finally, the Rayalaseema dialect is influenced by neighboring states' languages, i.e., Tamil and Kannada (Shivaprasad and Sadanandam, 2020). Interested readers are referred to Table 1 to see a few sample sentences of each dialect from the corpus. We also discuss these sentences in detail in Appendix B.

3 Dataset

We conduct our experiments on a corpus of three Telugu dialects collected by Mirishkar et al. (2021b). It is a crowd-sourced read speech corpus collected from the native speakers of the regional dialects of Telugu. In Table 2, we present dataset statistics we use in this study.[4]

Dialect	Train	Test	Vocabulary
Coastal Andhra	70.90K	1.99K	91737
Telangana	84.88K	2K	115505
Rayalaseema	65.32K	1.99K	90093

Table 2: Number of utterances in training and test set in each dialect (K for thousand)

All audio used in this study is mono channel, sampled at 16KHz with 16-bit encoding. The prompt given to the speakers is hand-curated. Therefore, we were able to ensure that the datasets across dialects have no domain mismatch. This allows us to study dialect mismatch better, which is our primary interest in this study.

3.1 Analysis

Since the dataset used in this paper is crowd-sourced read speech, we found a number of speakers not speaking in their native regional accent but in the "standardised" Telugu accent. However, the prompt given to the speakers is hand-curated which reflects the variations exhibited by the three dialects of interest. Additionally, we focus on dialect mismatched LMs in this paper. These reasons motivated us to limit ourselves to a textual analysis.

To analyse the three dialects, we choose to fine-tune *IndicBERT* (Kakwani et al., 2020) on a dialect classification task. *IndicBERT* is an ALBERT (Lan et al., 2020) based pre-trained multilingual model. It achieves state-of-the-art results on many Indic benchmarks and is trained on *IndicCorp* (Kakwani et al., 2020), one the largest publicly available Indian corpora.

To fine-tune *IndicBERT*, we use the same transcripts provided to the ASR models for training. We tokenise the input sequence using *IndicBERT*'s pre-trained tokeniser. We add a classification head to the pre-trained model. We use an initial learning rate of 1×10^{-5} with an Adam optimiser (Kingma and Ba, 2015). We train this model for 10 epochs. To get t-SNE representations, we take the sentence representations of the fine-tuned model and use *sklearn*'s implementation with default parameters.[5]

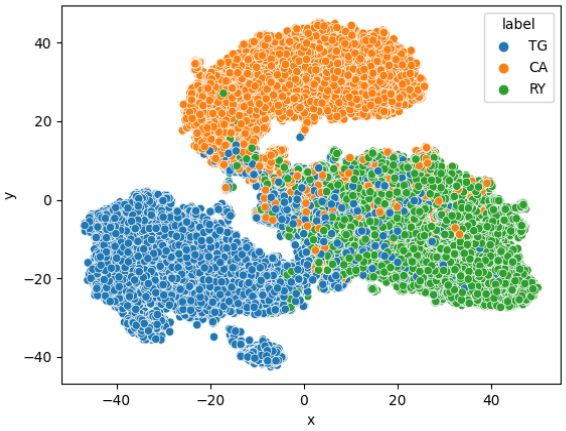

Figure 1: t-SNE plot of IndicBERT sentence representations of the three Telugu dialects. In this plot, TG is Telangana, CA is Coastal Andhra, and RY is Rayalaseema.

Figure 1 shows the t-SNE (van der Maaten and Hinton, 2008) plot of the sentence representations of *IndicBERT*. It can be observed that each of the dialects form its own cluster with some overlap with the other dialects. Out of the three, Ray-

[3] transliteration of the words using the WX notation (Gupta et al., 2010) are as follows: kAka, jAgA, xavAKAnA

[4] A more detailed analysis of the data used in this paper has been provided by Mirishkar et al. (2021b). We refer interested readers to their paper.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

alaseema cluster overlaps most with both Coastal Andhra and Telangana dialects, which shows that Rayalaseema dialect has a lot of similarities with both Coastal Andhra and Telangana dialect.

4 Experimental Setup

All of the ASR experiments are conducted using ESPnet (Watanabe et al., 2018). The input acoustic features are 80-dimensional log mel features extracted on the fly. We choose to use the Conformer model (Gulati et al., 2020) as it was able to achieve state-of-the-art performance on many standard datasets. The encoder of the ASR uses 12 Conformer (Gulati et al., 2020) blocks with 8 attention heads while the decoder uses 6 Transformer (Vaswani et al., 2017) blocks with 4 attention heads. We train both the encoder and decoder with a dropout rate of 0.1. All the models are trained based on the Hybrid CTC/Attention architecture (Kim et al., 2017; Watanabe et al., 2017). The training is done within the Multi-Objective Learning (MOL) framework. The CTC loss term helps the Attention model converge faster. The training objective (L_{MOL}) is as follows:

$$L_{MOL} = \lambda \log p_{ctc}(c|x) + (1-\lambda) \log p^{*}_{att}(c|x)$$

Here, λ is the multitask coefficient which should satisfy the following condition: $0 \leq \lambda \leq 1$. We found λ set to 0.3 while training and 0.4 while decoding gave us the best results for our datasets. c is the output unit. This could be characters, subword units, or words. Using words as output units could lead to two major issues: Out of Vocabulary (OOV%) cannot be handled well. The number of output units could be very high, especially in an agglutinative language like Telugu, which could lead to data sparsity. Chiu et al. (2018) show that using subwords over characters leads to better performance of End-to-End ASR systems. Thus we opted to use subwords as the output units. We used SentencePiece (Kudo and Richardson, 2018) to tokenise the words into subwords.[6] We found a vocabulary of around 500 tokens to give us the best performance on all the three datasets. We refer readers interested in how vocabulary size affects the performance of ASR of different Telugu dialects to Appendix A. Finally, x, in the above equation, is the input acoustic features.

[6]We used no external text to train the tokeniser.

We take mucs21_subtask1[7] recipe in ESPnet since it is tuned to perform well on a similar sized Indian dataset and make the following modifications: Change the initial learning rate to 5×10^{-4}, and use early stopping with a criterion to stop training the model if its performance does not improve for 5 consecutive epochs on the validation set.

We train an independent 16 block Transformer LM with an embedding size of 128 and a hidden encoder size of 512 for a maximum of 25 epochs. Finally, the decoder uses an LM weight of 0.6 to predict a sequence of subwords.[8] This method of integrating LM into the End-to-End ASR is known as Shallow Fusion (Kannan et al., 2018) and it is shown to give better results than other forms of integrating LM into the End-to-End ASR (Toshniwal et al., 2018). To decode, we use beam search of size 10 to predict the sequence.

5 Results & Discussion

In this section, the results of the experiments conducted are reported, and briefly analysed.

Biadsy et al. (2012) experiment the effectiveness of cross-dialect ASR in Arabic by experimenting with cross-dialect Acoustic Model (AM) and training the LM on target dialect data. In this paper, we take the exact opposite approach, i.e., train the AM (in this case, End-to-End ASR before the independent LM is fused) on the target dialect data and experiment by using a cross-dialect LM. We do this to test the effectiveness of the LM and thereby the ASR in cross-dialect conditions.[9] No external text was used to train LMs as it is difficult to obtain dialect information of external text.[10]

We report the performance of the LM both in terms of extrinsic metric, i.e., CER and WER of the ASR which uses the LM in question as well as an intrinsic metric, i.e., perplexity. Table 3 shows the performance of ASR systems in terms of CER and WER in both dialect matched and mismatched settings.[11]

[7]https://github.com/espnet/espnet/tree/master/egs2/mucs21_subtask1

[8]The rest of the weight is given to the CTC/Attention Hybrid Model.

[9]This is only possible because all dialects we experiment with share a common orthography

[10]For the rest of the paper, when we refer to a setting as mismatched consider only the LM to be mismatched.

[11]Even though WER is the most widely used metric, we report CER as we find WER to be not as reliable for agglutinative languages like Telugu as it is for analytic languages like English. However, in this paper, both the metrics are largely in agreement with each other.

Dialect/LM	None	Coastal Andhra	Telangana	Rayalaseema	All Dialects
Coastal Andhra	11.6/36.4	11.6/34.3	14.0/38.7	14.8/39.4	**11.2/34.3**
Telangana	7.6/27.9	16.7/40.1	7.6/25.4	17.0/40.9	**7.5/24.7**
Rayalaseema	8.6/26.5	8.2/25.4	7.9/25.1	**7.7**/24.2	9.0/**23.0**

Table 3: CER/WER(%) with Dialect Matched & Mismatched Language Models

As expected, ASR performs best when the LM is trained on all dialects outperforming ASR systems under matched conditions by approximately a WER of 1%. Since the dialect-specific text in our setup is not heavily skewed towards one dialect, the ASR performs well on all dialects. However, text collected from most external sources are heavily skewed towards the "standardised" Telugu dialect. Therefore, in the remaining part of the section, we focus on ASR systems where its LM is trained on a single dialect.

From Table 3, it can be observed that the average WER of the ASR in matched conditions is 27.96% and average CER is 8.96%. On the other hand, the average WER of the ASR in mismatched conditions is 34.93% and the average CER is 13.1%. This absolute difference of 6.97% in WER and 4.14% in CER of the ASR shows that having a dialect-specific LM can lead to the superior performance of an ASR. Moreover, our experiments with having no LM in ASR show that such a system can outperform ASR in mismatched conditions by upto 13% absolute WER. This shows that when the LM of the ASR is trained on text from a different dialect, it can *actively* hinder the performance of the ASR.

From Table 3, we can also observe that there is dissimilar amount degradation across all the three dialect ASR under mismatched settings. Telangana-specific ASR under mismatched conditions leads to over 15% WER drop compared to matched conditions. This is primarily due to the data imbalance in the dataset we used. Telangana dialect has most amount of data which leads to a superior performance when the LM is trained on it. However, when it is trained on other dialects, it is not only of different dialect but also trained on significantly lesser amount of data, which leads to an inferior model. On the other hand, Rayalaseema-specific ASR is robust to dialect mismatch with only slightly above 1% drop in performance compared to matched conditions. This is because Rayalaseema has a significant overlap with both Telangana and Coastal Andhra as shown in Figure 1. Since it has similarities with both Coastal Andhra and Telangana dialect, it performs relatively well even under dialect mismatched conditions.

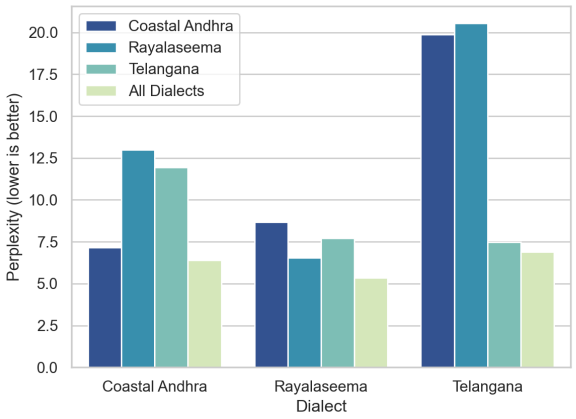

Figure 2: Perplexity in Cross-Dialect Conditions

Figure 2 shows the perplexity scores of LMs in dialect matched and mismatched settings. One can draw similar inferences from the perplexity scores as from the WERs of the ASR systems under different conditions presented in Table 3.

As expected, the perplexity of the LM trained on all the dialects is the least. LM's perplexity under matched settings much better in all the three dialects compared to mismatched conditions. As discussed before, Telangana LM is highly sensitive to dialect mismatch, with perplexity increasing by over 13 points. LM trained on Coastal Andhra and Telangana and tested on Rayalaseema leads to the highest increase in perplexity, i.e., 5.82 and 13.13 points, respectively. On the other hand, Rayalaseema LM is most robust to any dialect mismatch.

6 Conclusion & Future Work

This paper studies how LMs perform under dialect mismatched conditions. Our experiments reveal that LMs perform poorly, with the perplexity score increasing sharply in dialect mismatched

conditions. We use the mismatched LMs in ASR systems to study how they are affected. Similar to what we have observed with perplexity scores of the LM, we notice a significant degradation in the performance of the ASR with over 15% difference in WER in dialect mismatched conditions when compared to its matched counterpart. Furthermore, through our study, we show that mismatched LMs can *actively* hinder the performance of ASR by comparing it to ASR systems with no LM. These findings show the importance of careful curation of external text when training a dialect-specific ASR system.

These experiments have also led to an interesting finding: Rayalaseema dialect is more robust under dialect mismatched conditions as it shares a lot of similarities with both Coastal Andhra and Telangana.

In the future, we plan to improve the LM and thereby the ASR in dialect mismatched conditions using various adaptation techniques available in the literature. We hope that our future work would lead to LMs that are more robust to dialect mismatched conditions, thereby leading to improved ASR systems.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. We would like to acknowledge Technology Development for Indian Languages (TDIL), Ministry of Electronics and Information Technology (MeitY), Government of Republic of India for supporting us for the pilot project on "Crowd Sourced Large Speech Data Sets To Enable Indian Language Speech - Speech Solutions".

References

Kamadev Bhanuprasad and Mats Svenson. 2008. Errgrams a way to improving asr for highly inflected dravidian languages. In *IJCNLP*.

Fadi Biadsy, Pedro J. Moreno, and Martin Jansche. 2012. Google's cross-dialect arabic voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4441–4444.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Z. Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art speech recognition with sequence-to-sequence models. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778.

Keqi Deng, Songjun Cao, and Long Ma. 2021. Improving accent identification and accented speech recognition under a framework of self-supervised learning.

Anuj Diwan and Preethi Jyothi. 2020. Reduce and reconstruct: Improving low-resource end-to-end asr via reconstruction using reduced vocabularies.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*.

Mohamed G. Elfeky, Pedro Moreno, and Victor Soto. 2018. Multi-dialectical languages effect on speech recognition: Too much choice can hurt. *Procedia Computer Science*, 128:1–8. 1st International Conference on Natural Language and Speech Processing.

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata. 2018. Data augmentation improves recognition of foreign accented speech. In *INTERSPEECH*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Rohit Gupta, Pulkit Goyal, and Sapan Diwakar. 2010. Transliteration among indian languages using wx notation. In *KONVENS*.

Chao Huang, Tao Chen, S. Li, Eric Chang, and Jian-Lai Zhou. 2001. Analysis of speaker variability. In *INTERSPEECH*.

Venkanna Ithagani. 2014. Linguistic convergence and divergence in telugu-urdu contact situation: A study with special reference to telangana dialect.

Abhinav Jain, Vishwanath Pratap Singh, and Shakti Prasad Rath. 2019. A multi-accent acoustic model using mixture of experts for speech recognition. In *INTERSPEECH*.

Abhinav Jain, Minali Upreti, and Preethi Jyothi. 2018. Improved accented speech recognition using accent embeddings and multi-task learning. In *INTERSPEECH*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Anjuli Kannan, Yonnghui Wu, Patrick Nguyen, Tara N. Sainath, Zhifeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. *ICASSP*.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. 2021. Golos: Russian dataset for speech research. *Interspeech 2021*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ondřej Klejch, Electra Wallington, and Peter Bell. 2021. The cstr system for multilingual and code-switching asr challenges for low resource indian languages. In *Proceedings of Interspeech 2021*, pages 2881–2885. International Speech Communication Association. Interspeech 2021 : The 22nd Annual Conference of the International Speech Communication Association ; Conference date: 30-08-2021 Through 03-09-2021.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Z. Chen, Yan-Qing Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753.

Yi Y. Liu and Pascale Fung. 2006. Multi-accent chinese speech recognition. In *INTERSPEECH*.

Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. 2016. Mfcc-gmm based accent recognition system for telugu speech signals. *International Journal of Speech Technology*, 19(1):87–93.

Ganesh Mirishkar, Aditya Yadavalli, and Anil Kumar Vuppala. 2021a. An investigation of hybrid architectures for low resource multilingual speech recognition system in indian context. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 205–210.

Ganesh S Mirishkar, Vishnu Vidyadhara Raju V, Meher Dinesh Naroju, Sudhamay Maity, Prakash Yalla, and Anil Kumar Vuppala. 2021b. Cstd-telugu corpus: Crowd-sourced approach for large-scale speech data collection. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 511–517.

Van Tung Pham, Haihua Xu, Yerbolat Khassanov, Zhiping Zeng, Chng Eng Siong, Chongjia Ni, Bin Ma, and Haizhou Li. 2020. Independent language modeling architecture for end-to-end asr. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7059–7063.

Kanishka Rao and Haim Sak. 2017. Multi-accent speech recognition with hierarchical grapheme based models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4815–4819.

S Shivaprasad and M Sadanandam. 2020. Identification of regional dialects of telugu language using text independent speech processing models. *International Journal of Speech Technology*, 23(2):251–258.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In *SLTU*, pages 11–14.

Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. 2018. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 369–375.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Vishnu Vidyadhara Raju Vegesna, Krishna Gurugubelli, Hari Krishna Vydana, Bhargav Pulugandla, Manish Shrivastava, and Anil Kumar Vuppala. 2017. Dnn-hmm acoustic modeling for large vocabulary telugu speech recognition. In *MIKE*.

Thibault Viglino, Petr Motlícek, and Milos Cernak. 2019. End-to-end accented speech recognition. In *INTERSPEECH*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark A. Hasegawa-Johnson. 2018. Joint modeling of accents and acoustics for multi-accent speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jicheng Zhang, Yizhou Peng, Van Tung Pham, Haihua Xu, Hao Huang, and Chng Eng Siong. 2021. E2e-based multi-task learning approach to joint speech and accent recognition. In *Interspeech*.

Vocab/Dialect	Coastal Andhra	Telangana	Rayalaseema
500	**11.6/34.3**	**7.6/25.4**	**7.7/24.2**
700	11.8/35.0	7.9/25.5	8.2/25.1
1200	13.3/36.5	7.9/25.6	8.7/25.6
2500	15.3/38.1	8.7/25.9	10.0/27.1

Table 4: CER/WER(%) for Different Vocabulary Sizes

A Experiments with Different Vocabulary Sizes

In this paper, we conducted experiments with the following vocabulary sizes: 500, 700, 1200, 2500. Table 4 shows the performance of the ASR under these settings. We found that using 500 tokens results in best performance in all 3 dialect-specific ASR systems. We also conducted preliminary experiments by reducing the vocabulary size beyond 500 tokens but we could not find any noticeable improvement.

B Example Sentences

Table 5 presents the example sentences along with their transliterations using the WX notation (Gupta et al., 2010) and their translations. In Coastal Andhra, we notice the usage of the word "aMdi" frequently. In the example sentence, this word is fused with another word "veyAli" to become "veyAlaMdi". In the example Rayalaseema sentence, we notice the usage of the word "chaana". This is specific to the Rayalaseema dialect. The corresponding equivalent words in Coastal Andhra and Telangana would be "cAnA" and "masw", respectively. In Telangana, we notice the influence of Urdu/Hindi. In the example sentence, the words "masw" and "bE" have its origins in Urdu/Hindi.

Dialect	Sentence with Transliteration and Translation
Coastal Andhra	ప్రతి పౌరుడు ఓటు తప్పక వేయాలండి prawi pOrudu otu wappaka veyAlaMdi every citizen should vote without fail
Rayalaseema	మాకు మా పల్లెటూరు అంటే చానా ఇష్టము mAku mA palleVtUru aMte cAnA iRtamu we like our village very much
Telangana	గా ఫుట్బాల్ గురించి అయితే నాకు మస్త్ గా తెలుసు రా బ్రో gA PutbAl guriMci ayiwe nAku masw gA weVlusu rA bE I know a lot about football

Table 5: Example Sentences of Different Dialects