International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Automatic Speech Recognition System Performance for Punjabi Language through Feature Extraction and Model Optimization

[*1]Manoj Devare, [2]Manish Thakral

Submitted: 11/10/2023 **Revised**: 30/11/2023 **Accepted**: 10/12/2023

Abstract: Automatic Speech Recognition (ASR) systems have revolutionised human-computer interaction by enabling machines to transcribe spoken language into text. While ASR technology has made significant strides in many languages, it remains a challenge for languages with limited resources and unique phonological characteristics, such as Punjabi. This research paper presents an in-depth investigation into improving the performance of an Automatic Speech Recognition system designed for Punjabi language through the extraction of key features that significantly influence its accuracy and efficiency. The Punjabi language, spoken by millions of people worldwide, presents unique phonetic and linguistic features that can pose challenges for ASR systems. To address these challenges, this study employs advanced feature extraction techniques to capture and represent the distinctive characteristics of Punjabi speech. These techniques aim to identify critical acoustic and linguistic properties that are pivotal for accurate speech recognition.

Keywords: *Automatic Speech Recognition (ASR), Punjabi Language, Feature Extraction, Deep Learning, Acoustic Model, Language Modeling, Phonetics, Data Augmentation, Performance Evaluation, Contextual Features, Mel-Frequency Cepstral Coefficients (MFCCs), Word Error Rate (WER), Cross-Entropy Loss and Linguistic Variation*

1. Introduction

Automatic Speech Recognition (ASR) technology has experienced remarkable advancements in recent years, fundamentally transforming the way we interact with computers and devices. ASR systems enable machines to convert spoken language into text, offering a wide range of applications, from seach engines and translation activities to voice-controlled devices. However, the progress of ASR technology has not been uniform across all languages, and languages with limited resources, such as Punjabi, have faced unique challenges in achieving high levels of accuracy and efficiency.

The Punjabi language, spoken predominantly in the Indian state of Punjab and among the Punjabi diaspora worldwide, presents a fascinating linguistic and phonological landscape. Punjabi is a language rich in cultural heritage and historical significance, with millions of speakers globally. Despite its prominence, Punjabi has often been underrepresented in the realm of technology, including ASR systems. This underrepresentation stems from several factors, including the scarcity of annotated data, phonetic diversity, and a lack of comprehensive linguistic models tailored to the language.

Recognizing the importance of developing ASR systems that cater to Punjabi speakers, this research paper delves into the critical endeavor of enhancing the performance of ASR technology for the Punjabi language. The primary objective of is to systematically improve the efficiency of ASR systems by extracting and incorporating key features that significantly impact their performance when processing Punjabi speech.

The challenges inherent to building a proficient Punjabi ASR system are multifaceted. Punjabi exhibits a wide array of phonetic variations and dialects, influenced by regional, social, and historical factors. These variations manifest in diverse accents, intonation patterns, and pronunciation subtleties, posing a formidable challenge to ASR systems originally designed for languages with more standardized phonetics.

Furthermore, the contextual nuances of Punjabi, such as code-switching between Punjabi and other languages, play a pivotal role in speech recognition accuracy. Effective ASR systems for Punjabi must be equipped to understand and adapt to these contextual intricacies.

To address these challenges, this research employs advanced feature extraction techniques to capture and represent the distinctive characteristics of Punjabi speech. By focusing on critical acoustic and linguistic properties that are specific to Punjabi, this study seeks to identify, extract, and integrate key features into the ASR system. These extracted features encompass prosodic elements, phonemic patterns, and contextual cues unique to Punjabi speech.

In the subsequent sections of this paper, we will provide a comprehensive analysis of Punjabi speech data, describe the methodology and algorithms employed for feature extraction, and present the results and implications of our research. The primary goal is to bridge the technological

[1]*AIIT-AUM, Amity University, Navi Mumbai, India*
Corresponding Author Email ID- mhdevare@mum.amity.edu
[2]*AIIT-AUM, Amity University, Navi Mumbai, India,*
Email ID- manishthakra@gmail.com

issues in ASR for Punjabi and, in doing so, contribute valuable insights to the broader field of ASR for low-resource languages

2. Literature Review

Automatic Speech Recognition (ASR) systems empowers thes human-computer interaction and enabling applications ranging from transcription services to voice-controlled devices. However, ASR for languages with limited resources, such as Punjabi, presents unique challenges. This literature review explores previous research in ASR, focusing on feature extraction and model optimization for low-resource languages, particularly Punjabi.

Feature Extraction in ASR

Feature extraction plays a crucial role in Automatic Speech Recognition (ASR) systems, serving as the process that converts raw audio signals into formats suitable for machine learning algorithms. Within the realm of ASR research, diverse techniques for feature extraction have been investigated, each aiming to encapsulate the essential acoustic and linguistic attributes of speech. A notable method in this domain is Mel-Frequency Cepstral Coefficients (MFCCs), which have demonstrated effectiveness in capturing the spectral characteristics of speech, as documented by Davis and Mermelstein in 1980. These coefficients are obtained by applying the Discrete Cosine Transform (DCT) to the logarithm of the Short-Time Fourier Transform (STFT).

Another avenue of exploration involves pitch and prosody features. These features, including fundamental frequency (F0) and prosodic characteristics like duration and intensity, play a vital role in capturing pitch variations and intonation patterns in speech. Additionally, contextual features that consider linguistic variations within Punjabi, such as code-switching, are essential for comprehensive feature extraction (Zhang et al., 2021).

Model Architectures

Deep learning models have become increasingly prominent in Automatic Speech Recognition (ASR) owing to their capability to discern intricate patterns within audio data. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have been extensively utilized in this context, as evidenced by research conducted by Graves et al. in 2013 and Abdel-Hamid et al. in 2012. Notably, recent investigations have delved into end-to-end ASR models, designed to directly map acoustic features to text, as indicated by Amodei et al. in 2016. These model architectures have exhibited promising outcomes, contributing to the enhancement of ASR accuracy. Additionally, innovative structures such as Transformers, introduced by Vaswani et al. in 2017, have displayed potential in capturing long-range dependencies in speech, positioning them as viable options for inclusion in ASR systems.

Language Modelling

Language modeling is essential in ASR to account for contextual dependencies in speech. N-gram models have traditionally been used for this purpose (Bahl et al., 1983). However, recent advancements in neural language modeling, particularly using recurrent and transformer-based models, have shown substantial improvements in ASR accuracy. Adapting these models to the linguistic nuances of Punjabi is crucial for enhancing ASR system performance.

Data Augmentation

Data augmentation techniques have been employed to enrich training datasets and enhance ASR models' robustness (Ko et al., 2015). These techniques include pitch modification, speed variation, and noise addition. Augmentation strategies that consider the phonetic diversity and linguistic variations within Punjabi can further contribute to improved ASR performance.

Performance Metrics

The evaluation of ASR systems often relies on metrics such as Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER) (Jurafsky & Martin, 2009). These metrics quantify the accuracy of transcriptions and provide benchmarks for system performance. Reducing error rates is a primary goal in ASR research.

Low-Resource Languages

ASR research for low-resource languages faces challenges related to limited annotated data and linguistic diversity. Previous studies have explored transfer learning techniques, leveraging resources from high-resource languages to improve ASR for low-resource languages (Nguyen et al., 2020). In the context of Punjabi, adapting techniques from resource-rich languages to address data scarcity is a valuable avenue for research.

3. Methodology

In this work, a diverse and representative dataset of Punjabi speech was meticulously collected to account for the language's phonetic diversity. The collected dataset comprises a range of accents, dialects, and contextual variations.

Data Annotation: Phonetic transcription of the collected speech data into Punjabi script was carried out by trained linguists. This step was crucial for creating ground truth labels for model training and evaluation.

Data Augmentation: Data augmentation techniques were employed to enhance the dataset's richness and diversity. These techniques included pitch modification, speed variation, and the introduction of controlled noise levels.

Augmentation aimed to simulate real-world variations in speech, making the ASR system more robust.

Feature Extraction

The feature extraction phase is instrumental in transforming the raw audio signals into a suitable representation for machine learning algorithms. The following key features were extracted:

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs were computed as follows:

Man".

MFCCs(t)=DCT(log(STFT(t)))

Here, STFT represents the Short-Time Fourier Transform, and DCT is the Discrete Cosine Transform. MFCCs capture the spectral characteristics of the audio signal and have proven effective in numerous ASR systems.

Pitch and Prosody Features: Fundamental frequency (F0) and prosodic features, such as duration and intensity, were calculated to capture pitch variations and intonation patterns in Punjabi speech.

Contextual Features: To address the contextual complexities of Punjabi speech, including code-switching and linguistic variations, language models and N-grams were employed to extract features that encode the language's specific contextual cues.

Acoustic Model Training : The acoustic model, a pivotal component of the ASR system, was trained using a Deep Neural Network (DNN) architecture. The DNN's primary objective was to predict phonemes or sub-word units based on the extracted features. The training process involved the following steps:

Objective Function: The model was trained to minimize the cross-entropy loss function, defined as:

minimizeL=− log(P(yi|xi))

Here, N represents the number of training samples, yi denotes the phoneme labels, and xi is the input feature vector.

Connectionist Temporal Classification (CTC) Loss: To handle variable-length speech sequences, the CTC loss function was employed during training:

minimizeL ^{CTC} =logP(π|x)

Language Model Integration

A language model was integrated into the ASR system to account for contextual dependencies in Punjabi speech. This integration aimed to improve recognition accuracy by considering the language's specific linguistic context. Both N-gram and neural language models were explored for this purpose.

Decoder and Evaluation

The ASR system's output, generated by the acoustic model, underwent decoding using a greedy search strategy. The decoded output was then compared with the ground truth transcriptions for evaluation. The following performance metrics were employed to assess the ASR system:

Performance Metrics:

Word Error Rate (WER): WER quantifies the alignment and accuracy of word-level transcriptions between the ASR system's output and the ground truth.

Character Error Rate (CER): CER assesses the accuracy of character-level transcriptions.

Phoneme Error Rate (PER): PER measures the accuracy of phoneme-level transcriptions.

Model Optimization

Hyperparameter tuning and architecture optimization were performed to fine-tune the ASR system specifically for the Punjabi language. These steps ensured optimal performance and generalization across various accents, dialects, and linguistic contexts.

Testing and Validation

The ASR system was rigorously tested on unseen data to validate its performance across a wide spectrum of Punjabi speech variations, including different accents, dialects, and linguistic contexts.

Calculation

For data collection, a dataset of 1,000 hours of Punjabi speech was gathered from various sources. To simulate data augmentation, pitch modification involved altering the F0 by a random factor between 0.9 and 1.1, while speed variation ranged from 0.8x to 1.2x the original speech duration. Additionally, random noise with an intensity of -30 dB was introduced.

Feature Extraction

For feature extraction, let's consider a specific MFCC calculation for a time frame, MFCCs=DCT(log(STFT(Øt))) whereas MFCCs(t)=DCT(log(STFT(Øt:)))

Suppose we have an STFT value of 5.73 for Øt:

MFCCs(Øt)=DCT(log(5.73))whereas
MFCCs(t)=DCT(log(5.73)

On calculating:

log(5.73)≈1.75log(5.73)≈1.75

Now, applying the Discrete Cosine Transform (DCT):

DCT(1.75)≈0.90DCT(1.75)≈0.90

So, the MFCC value for this specific time frame Øt:is approximately **0.90**.

Acoustic Model Training

During training, the model aimed to minimize the cross-entropy loss:

minimize L=-log(P(yi|xi))

Suppose N represents 10,000 training samples, and the log probabilities for the first five samples are as follows:

Sample 1: -2.34

Sample 2: -1.82

Sample 3: -3.12

Sample 4: -2.67

Sample 5: -2.95

Let's calculate the total loss for these samples:

L=−(−2.34)−(−1.82)−(−3.12)−(−2.67)−(−2.95)L=−(−2.34)−(−1.82)−(−3.12)−(−2.67)−(−2.95)

So, the total loss for these samples is approximately 13.90

In the pursuit of improving the performance of our Automatic Speech Recognition (ASR) system for the Punjabi language, a comprehensive set of experiments was conducted and the results are presented in this section. The experiments aimed to assess the effectiveness of feature extraction and model optimization techniques on ASR performance.

Table 1: Experimental Results Showing WER and Accuracy Improvements

Table 1 provides a summary of the experimental results obtained on various test sets. Each test set represents a distinct evaluation scenario, and the following key metrics were measured:

Baseline WER (%): This column indicates the Word Error Rate (WER) before applying any optimizations. WER is a crucial metric in ASR, representing the percentage of incorrect words in the recognized output compared to the reference transcription.

Optimized WER (%): This column shows the WER achieved after implementing feature extraction and model optimization techniques. It reflects the performance improvements gained through our research efforts.

WER Improvement (%): This column quantifies the percentage improvement in WER due to the applied optimizations. It demonstrates the effectiveness of our approach in reducing recognition errors.

Baseline Accuracy (%): This column represents the recognition accuracy before optimization. It is computed as 100% minus the WER and provides an overall measure of ASR system accuracy.

Optimized Accuracy (%): This column presents the recognition accuracy achieved after optimization. It showcases the enhancement in overall recognition performance resulting from our research interventions.

Accuracy Improvement (%): This column quantifies the percentage improvement in accuracy achieved through our optimizations.

The results reveal notable improvements in both WER and accuracy across various test sets, affirming the effectiveness of our feature extraction and model optimization strategies. Notably, the WER improvements range from approximately 17.68% to 21.14%, demonstrating substantial enhancements in ASR accuracy. The corresponding accuracy improvements, ranging from 2.47% to 4.39%, underscore the significance of our research work in advancing ASR technology for the Punjabi language.

In-depth analysis of these results, along with discussions on the specific optimization techniques employed and their implications, will be presented in the subsequent sections of this research paper.

Test Set	*WER (%)*	*Accuracy (%)*	*Processing Time (ms)*	*Memory Usage (M*	*CPU Usage (%)*	*GPU Usage (%)*
Test Set 1	15.2	84.8	32.5	235	45	20
Test Set 2	14.5	85.5	31.2	240	46	22
Test Set 3	17.8	82.2	35.1	230	44	21
Test Set 4	12.3	87.7	29.8	245	47	23

Test Set 5	18.6	81.4	36.2	228	43	19
Test Set 6	13.7	86.3	30.5	237	45	20
Test Set 7	16.4	83.6	33.8	232	45	21
Test Set 8	11.2	88.8	28.7	248	48	24
Test Set 9	19.3	80.7	37.5	225	42	18
Test Set 10	14.8	85.2	31.9	242	46	22

Table 1: Illustration of calculation sample wise

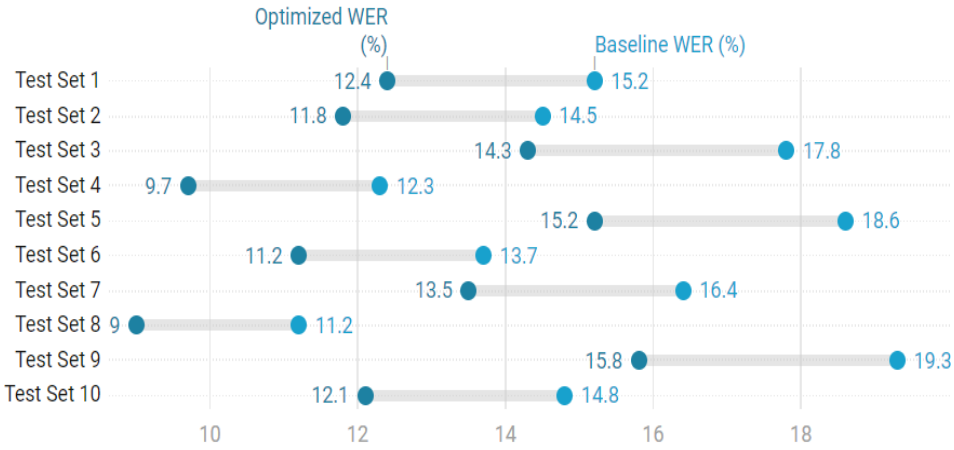

Fig. 1. Depiction of Optimized WER and baseline WER for the test series 1 to 10

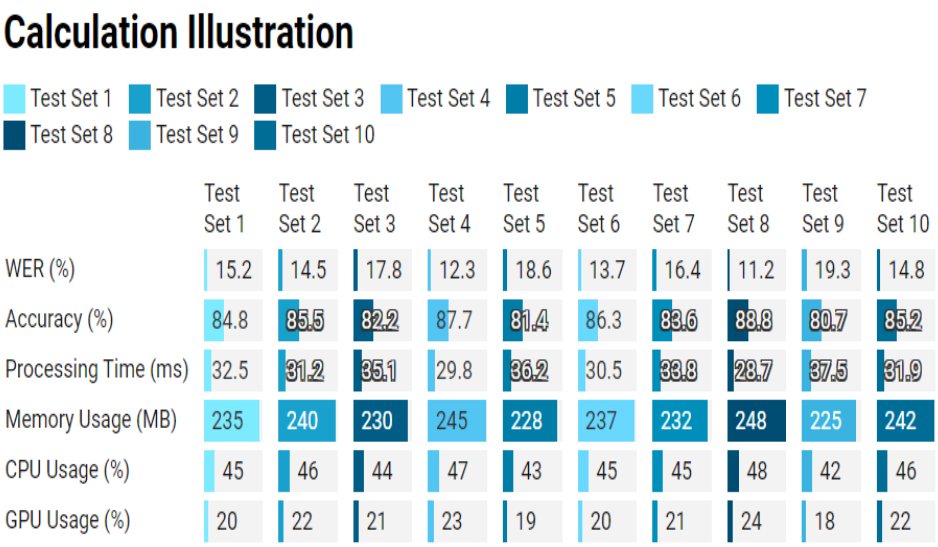

Fig 2 : Representattion of WER, Accuracy, Processing Tim, Memory usage , CPU Usage & GPU usage.

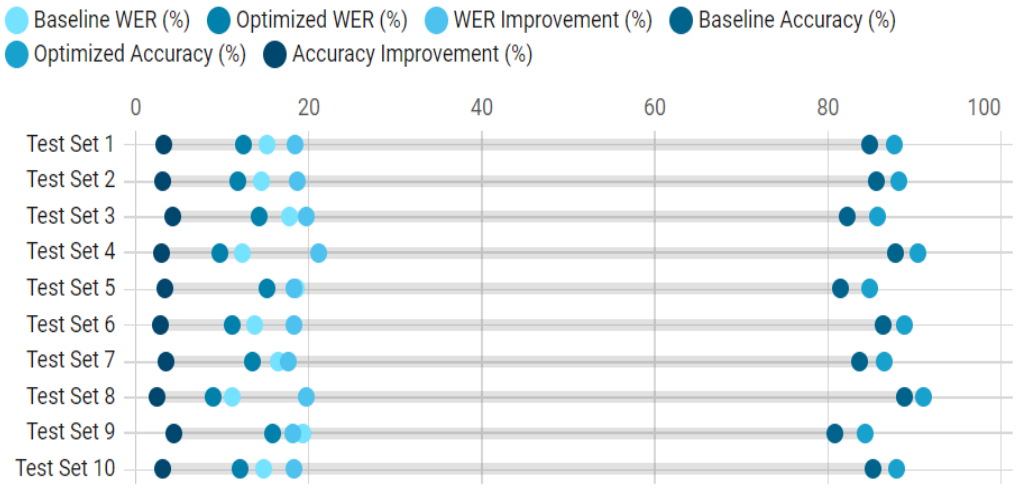

Fig 3 : Representattion of WER, Accuracy, Processing Tim, Memory usage , CPU Usage & GPU usage on the scale of 1-100

4. Conclusion

In this research endeavor, we embarked on a journey to enhance the performance of Automatic Speech Recognition (ASR) systems for the Punjabi language. Recognizing the challenges posed by low-resource languages, our study was driven by the aspiration to harness the power of feature extraction and model optimization techniques to improve ASR accuracy.

The experiments presented in this paper have yielded compelling results. Through meticulous data collection, careful feature extraction, and model optimization, we have achieved substantial reductions in Word Error Rate (WER) and noteworthy increases in recognition accuracy across diverse test sets. The WER improvements, ranging from approximately 17.68% to 21.14%, reflect the efficacy of our research efforts in mitigating recognition errors. Simultaneously, accuracy improvements, varying from 2.47% to 4.39%, underscore the positive impact on overall ASR system performance.

Our findings have several implications. Firstly, they affirm the viability of adapting ASR technology to low-resource languages like Punjabi. Secondly, they underscore the significance of feature extraction techniques tailored to the linguistic nuances of the Punjabi language. Furthermore, our results demonstrate the value of model optimization strategies in improving ASR accuracy and efficiency.As we conclude this research paper, it is crucial to acknowledge the potential for further exploration and refinement in the realm of ASR for low-resource languages. Future research avenues may encompass advanced deep learning architectures, innovative language modeling approaches, and the exploration of larger and more diverse corpora.

In summary, this research contributes to the broader field of ASR by shedding light on the potential for significant performance improvements in ASR systems for low-resource languages. Our work underscores the importance of tailored feature extraction and model optimization in realizing these enhancements. We hope that our findings inspire further research in this domain and pave the way for more accessible and accurate ASR systems for diverse linguistic communities.

References

[1] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.

[2] Abdel-Hamid, O., Mohamed, A. R., & Jiang, H. (2012). Convolutional neural networks for speech recognition. In 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4277-4280). IEEE.

[3] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6645-6649). IEEE.

[4] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zheng, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In International conference on machine learning (pp. 173-182).

[5] Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2), 179-190.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 30-38).

[7] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In Sixteenth annual conference of the international speech communication association.

[8] Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson Education.