# 6. Performance Analysis, Results, Conclusion and Future Scope for Extension of Research Work

## 6.1 Introduction

The research work titled “Development of a Model to Analyse and Interpret Vernacular Voice Recognition of Gujarati Dialects” addresses the growing need for region-specific and dialect-sensitive voice recognition systems. While significant progress has been made in the development of speech and voice recognition technologies globally, most existing systems lack the capability to recognize and process vernacular dialects effectively. Gujarati, a widely spoken language in India and across the world, comprises rich linguistic diversity with distinct dialects, including Kathiyawadi, Standard Gujarati, Surti, and Kutchhi. However, there are few researches and tools to work with the distinguishing peculiarities of these dialect variations, firstly, between pronouncing, intonating, and vocabulary differences. Recognizing this gap, the researcher aimed to develop a robust voice recognition model that could accurately identify speakers, dialects, and speech inputs from audio samples specific to Gujarati dialects. The reason for the development of this system was to accommodate the local linguistic needs and make a positive contribution to the development of vernacular speech processing. The thesis concludes with a summary of the main findings, performance evaluation and the entire impact of the research with future scope.

## 6.2 Performance Analysis of Vernacular Voice Recognition Model for Gujarati Dialects

The performance analysis of the proposed Voice Recognition Model for Gujarati Dialects is conducted to evaluate its accuracy, efficiency, and robustness in recognizing speakers and dialects under various conditions. This section analyses the performance of the model in various evaluation parameters, test scenarios and comparative analysis. Specifically, we focus on the areas of speaker identification accuracy, dialect specific recognizers, gender-based performance and prediction with audio length.

The researcher has identified and focused on four prominent regional dialects of the Gujarati language for this study. To develop and evaluate the voice recognition model, a comprehensive dataset of voice samples was collected from a diverse group of speakers representing these dialects. More specifically, the dataset consists of approximately 17000 voice samples from 79 different speakers, which gives us enough speakers of varying dialectal variations and gender as well as speaker particular characteristics.

Many voice data was thoroughly collected from different public resources like online audio libraries, openly available datasets, and multimedia streaming services. The diverse sources were useful because they helped ensure a well-rounded dataset that reflects actual speech variances.

From the total samples, 30% was reserved for testing. To evaluate the performance of the proposed Voice Recognition Model with MVR Tool, this subset has been taken as input. To run a robust learning process in the speaker and dialect recognition, we used the remaining 70% of the dataset for training and validation. By applying this structured, the research assures a correct and systematic evaluation of the model’s accuracy, speed and effectiveness as a whole.

### 6.2.1 Speaker Wise Result Analysis for Voice Recognition

The result analysis of voice recognition for each speaker involves evaluating the performance of the proposed model across all speakers from the four identified Gujarati dialects: Kathiyawadi, Standard Gujarati, Surti, and Kutchhi. Researcher has analysed the model's predictions to how well the model could identify a person speaking based on their voice sample. Measurements of standard performance metrics for each speaker like correctly identified, incorrectly identified, and not identified samples was done for each speaker shows in Table 6-1. Finally, our results show that the model was extremely accurate for speakers with longer and clearer audio inputs. However, for speakers of short audio samples, or with speech from which noise or overlapping speech patterns can be derived, there were slight variations. The analysis concludes that overall, the model is robust as recognition rates remain high with the majority of speakers across the majority of the speakers.

*Table 6- 1 Result Analysis of each Spekaer of Kathiyawadi Dialect*

| Speaker ID | Total Test Samples | Correctly Identified | Incorrectly Identified | Not Identified |
|---|---|---|---|---|
| speaker021 | 42 | 34 | 0 | 8 |
| speaker008 | 40 | 37 | 0 | 3 |
| speaker004 | 95 | 85 | 0 | 10 |
| speaker003 | 43 | 39 | 0 | 4 |
| speaker018 | 40 | 31 | 0 | 9 |
| speaker017 | 54 | 31 | 1 | 22 |
| speaker027 | 87 | 85 | 0 | 2 |
| speaker024 | 91 | 84 | 0 | 7 |
| speaker025 | 19 | 14 | 1 | 4 |
| speaker005 | 48 | 37 | 1 | 10 |
| speaker022 | 52 | 37 | 0 | 15 |
| speaker002 | 25 | 11 | 2 | 12 |
| speaker023 | 15 | 6 | 0 | 9 |
| speaker007 | 32 | 21 | 0 | 11 |
| speaker015 | 41 | 26 | 1 | 14 |
| speaker001 | 15 | 8 | 0 | 7 |
| speaker026 | 79 | 65 | 0 | 14 |
| speaker014 | 39 | 31 | 0 | 8 |
| speaker012 | 41 | 31 | 0 | 10 |
| speaker020 | 30 | 26 | 0 | 4 |
| speaker019 | 91 | 72 | 1 | 18 |
| speaker010 | 76 | 56 | 0 | 20 |
| speaker011 | 44 | 34 | 0 | 10 |
| speaker013 | 50 | 37 | 0 | 13 |
| speaker009 | 27 | 20 | 0 | 7 |
| speaker006 | 33 | 22 | 0 | 11 |
| speaker016 | 21 | 16 | 0 | 5 |

The performance analysis for speakers belonging to the Kathiyawadi dialect shows in Figure 6-1 reveals varying levels of recognition accuracy across individual speakers. Twenty-six speakers were evaluated with test sample size from 15 to 95 audio inputs. speaker027 and speaker004 showed that they are robust, with high recognition accuracy on their inputs, of 85 out of 87 and 85 out of 95 samples respectively. Similarly, speakers speaker008 and speaker014 got a very good performance with the minimal amount of unknown samples, with a accuracy of 92,5% and 79,5% respectively. However, various speakers struggled during the recognition part. For instance, speaker002 and speaker017 recorded comparatively lower correct identifications with 11 out of 25 and 31 out of 54 samples, respectively, while also exhibiting higher "not identified" rates. Further, the analysis suggests that speakers with lesser audio input, such as speaker001 and speaker023, yielded changing results because their datasets are smaller.

Overall, it is shown that accuracy decreases with noisy data or fewer test samples, while speakers with larger and clearer datasets obtain quite remarkable recognition performance. This underscores the importance of data quality and sample size in ensuring the model's efficiency for speaker-specific recognition in the Kathiyawadi dialect.

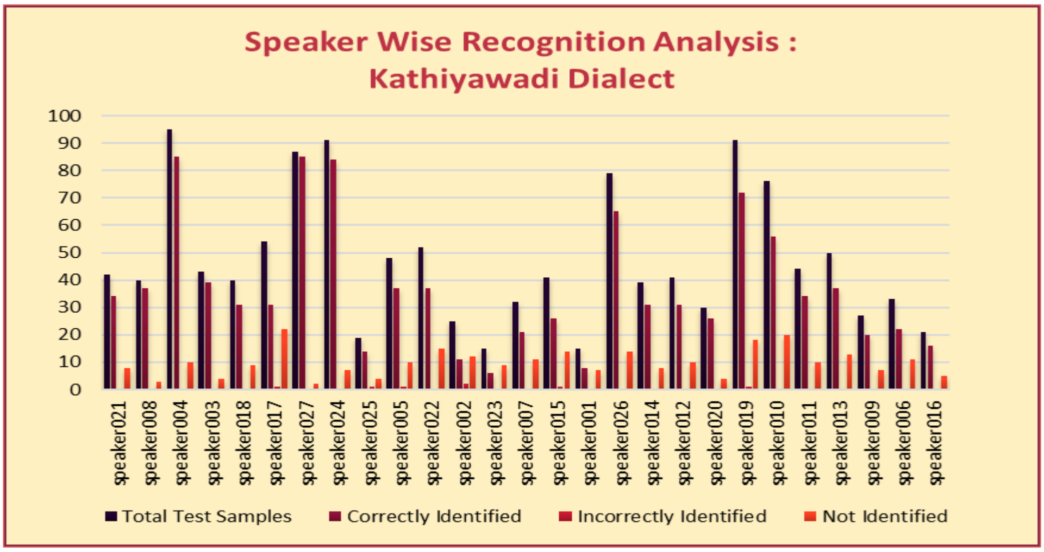

*Figure 6- 2 Speaker Wise Recognition Analysis of Kathiyawadi Dialects*

The table 6-2 below shows the result analysis of each speaker belongs to Kutchhi dialects.

*Table 6- 2 Result Analysis of each Spekaer of Kutchhi Dialect*

| Speaker | Total Test Samples | Correctly Identified | Incorrectly Identified | Not Identified |
|---|---|---|---|---|
| speaker032 | 97 | 89 | 0 | 8 |
| speaker028 | 13 | 9 | 0 | 4 |
| speaker034 | 17 | 11 | 2 | 4 |
| speaker030 | 28 | 20 | 0 | 8 |
| speaker036 | 76 | 74 | 0 | 2 |
| speaker029 | 20 | 14 | 0 | 6 |
| speaker037 | 30 | 19 | 0 | 11 |
| speaker035 | 121 | 114 | 0 | 7 |

| | | | | |
|---|---|---|---|---|
| speaker033 | 46 | 38 | 0 | 8 |
| speaker031 | 33 | 21 | 0 | 12 |

The voice recognition model performance analysis for speakers of Kutchhi Dialect which can contain their total test samples, correctly identified samples, incorrect identified samples, samples not identified shows in figure 6-2. Levels of accuracies across the dataset have different being highlighted in the analysis. speaker036 and speaker035 were perfect speakers, correctly identifying 76 and 121 out of 76 and 121 samples, respectively, with high accuracy and a relatively low error rate. Similarly, speaker032 was very successful in classifying 89 samples out of 97 with no more than one discrepancy. However, speakers like speaker037 and speaker031 located lower recognition accuracy, with 19 of 30 samples identified correctly but 11 samples left unrecognized, and speakers like speaker031 left 12 samples unrecognized. Occasional misclassifications on two incorrectly identified samples were recorded in Speaker034.

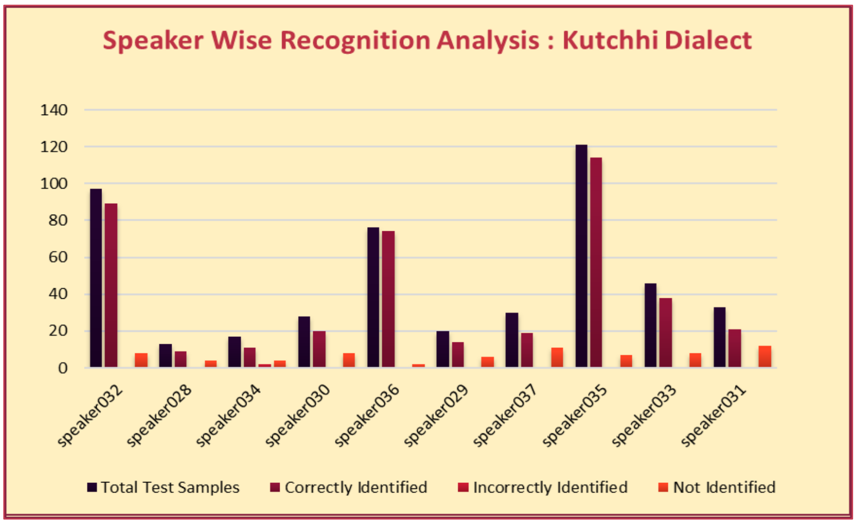

*Figure 6- 3  Speaker Wise Recognition Analysis of Kutchhi Dialects*

A performance analysis of the proposed voice recognition model for selected speakers is performed and is given in Table 6-3. It describes total test samples, how many were correctly classified, incorrectly classified and how many were not classified. Results indicate that speakers having more test samples performed better at recognition. As an example, speaker067 and speaker040 were able to distinguish 142 out of 148 and 132 out of 145 samples respectively with great accuracy. Similarly, speaker048 and speaker054 performed well, correctly identifying 119 out of 135 and 100 out of 110 samples, respectively, with minimal "not identified" cases. Interestingly, speakers with fewer test samples, such as speakers 057 (12 samples) and 039 (18 samples) had more unclassified cases (7 and 8 samples) than recognized cases. This difference suggests that smaller datasets might not realize consistent predictions. Some speakers also showed noticeable difficulties in identification such as speaker058 who correctly identified only 27 of 47 samples, and speaker073, who identified correctly 33 of 48 samples with 15 not identified. Theoretical analysis demonstrates that the model is capable of reliable performance for the majority of speakers and correctly classifies a significant fraction of the test samples.

Comparison of each speaker’s performance of recognition analysis has shown in Figure 6-3 below:

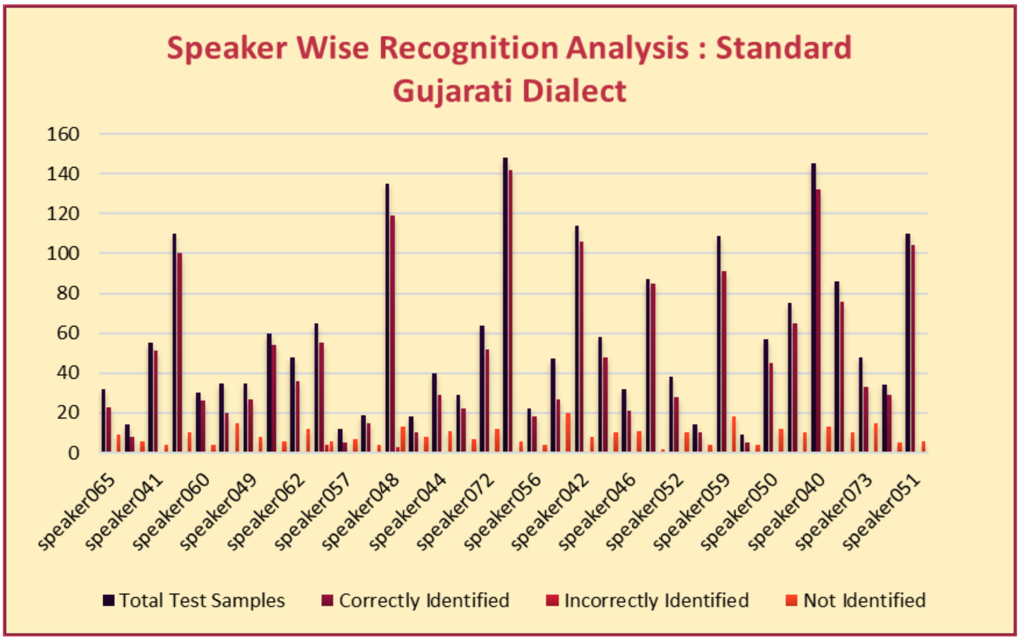

*Figure 6- 6  Speaker Wise Recognition Analysis of Standard Gujarati Dialects*

*Table 6- 3 Result Analysis of each Spekaer of Standard Gujarati Dialect*

| Speaker | Total Test Samples | Correctly Identified | Incorrectly Identified | Not Identified |
|---|---|---|---|---|
| speaker065 | 32 | 23 | 0 | 9 |
| speaker070 | 14 | 8 | 0 | 6 |
| speaker041 | 55 | 51 | 0 | 4 |
| speaker054 | 110 | 100 | 0 | 10 |
| speaker060 | 30 | 26 | 0 | 4 |
| speaker064 | 35 | 20 | 0 | 15 |
| speaker049 | 35 | 27 | 0 | 8 |
| speaker066 | 60 | 54 | 0 | 6 |
| speaker062 | 48 | 36 | 0 | 12 |
| speaker047 | 65 | 55 | 4 | 6 |
| speaker057 | 12 | 5 | 0 | 7 |
| speaker045 | 19 | 15 | 0 | 4 |
| speaker048 | 135 | 119 | 3 | 13 |
| speaker039 | 18 | 10 | 0 | 8 |
| speaker044 | 40 | 29 | 0 | 11 |
| speaker061 | 29 | 22 | 0 | 7 |
| speaker072 | 64 | 52 | 0 | 12 |
| speaker067 | 148 | 142 | 0 | 6 |
| speaker056 | 22 | 18 | 0 | 4 |
| speaker058 | 47 | 27 | 0 | 20 |
| speaker042 | 114 | 106 | 0 | 8 |
| speaker071 | 58 | 48 | 0 | 10 |
| speaker046 | 32 | 21 | 0 | 11 |
| speaker068 | 87 | 85 | 0 | 2 |
| speaker052 | 38 | 28 | 0 | 10 |
| speaker038 | 14 | 10 | 0 | 4 |
| speaker059 | 109 | 91 | 0 | 18 |
| speaker043 | 9 | 5 | 0 | 4 |
| speaker050 | 57 | 45 | 0 | 12 |
| speaker053 | 75 | 65 | 0 | 10 |
| speaker040 | 145 | 132 | 0 | 13 |
| speaker069 | 86 | 76 | 0 | 10 |
| speaker073 | 48 | 33 | 0 | 15 |
| speaker055 | 34 | 29 | 0 | 5 |
| speaker051 | 110 | 104 | 0 | 6 |

The detailed analysis of speaker belongs to Surti Dialects shows in Table 6-4 below:

*Table 6- 4 Result Analysis of each Spekaer of Surti Dialect*

| Speaker | Total Test Samples | Correctly Identified | Incorrectly Identified | Not Identified |
|---|---|---|---|---|
| speaker075 | 66 | 49 | 5 | 12 |
| speaker079 | 12 | 6 | 2 | 4 |
| speaker076 | 34 | 27 | 0 | 7 |
| speaker077 | 48 | 43 | 1 | 4 |
| speaker074 | 98 | 94 | 0 | 4 |
| speaker078 | 12 | 6 | 0 | 6 |

The graph (Figure 6-4) displays the Speaker-Wise Recognition Analysis for the Surti dialect, showing performance across six speakers. High accuracy was demonstrated by Speaker074, the best performing speaker, since it only misidentified only 4 out of 98 samples. Speaker 077 also did well with 43 correct identifications out of 48 samples with little errors. Speaker075 and Speaker076 exhibited moderate performance, with 49/66 and 27/34 samples correctly recognized, but they had notable "Not Identified" counts of 12 and 7, respectively. For Speaker078 and Speaker079, recognition accuracy was lower on 12-sample datasets for Speaker078 with 6 unrecognized samples and for Speaker079, 4 unrecognized, 2 misclassified samples on 6-sample datasets. The analysis suggests that larger datasets, such as those for Speaker074, yield higher recognition accuracy, while smaller datasets result in a higher percentage of "Not Identified" outcomes, emphasizing the importance of sample size for robust speaker recognition.

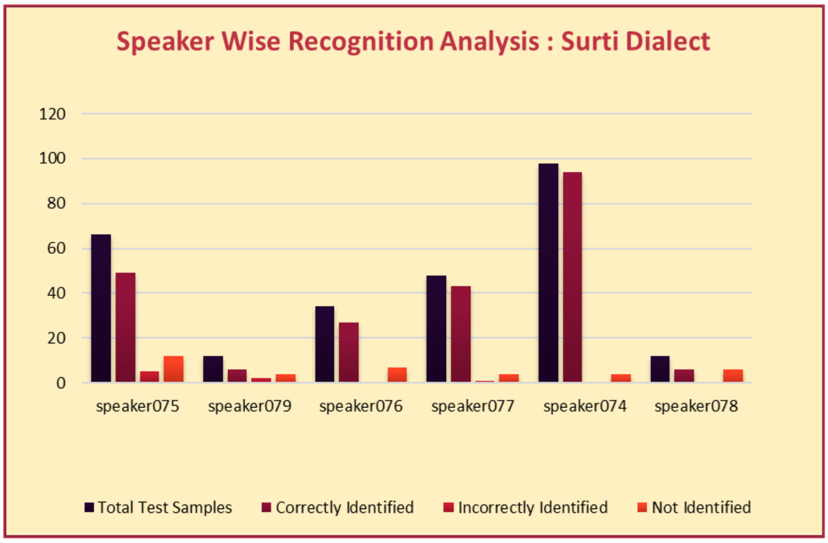

*Figure 6- 7 Speaker Wise Recognition Analysis of Surti Dialects*

### 6.2.2 Gender Wise Recognition Analysis

For performance in the dataset, the proposed voice recognition model is evaluated and Gender Wise Recognition Analysis is performed. Results show the model achieved higher recognition accuracy for male speakers, fewer misclassifications and less unrecognized inputs. The performance was better in audio inputs from male speakers, which tended to give clearer and more consistent audio. In contrast, the analysis for female speakers shows relatively lower recognition rates, with a noticeable increase in "Not Identified" samples. This is because female voice recordings come in different pitches, tones and quality and this made it difficult for the model during recognition.

Overall, it appears that the model performs much better gender-wise, in terms of handling male speaker data well, but has an area of improvement in optimization for female speakers. Taken together, these insights call to suitably balance gender representation and improve model robustness against voice characteristics variation.

The table 6-5 shows the statistic related Gender wise analysis:

*Table 6- 5 Gender Wise Recognition Analysis*

| Gender | Total Test Samples | Correctly Identified | Incorrectly Identified | Not Identified | Accuracy |
|---|---|---|---|---|---|
| Male | 1949 | 1630 | 6 | 313 | 96.00% |
| Female | 2106 | 1717 | 18 | 371 | 94.87% |

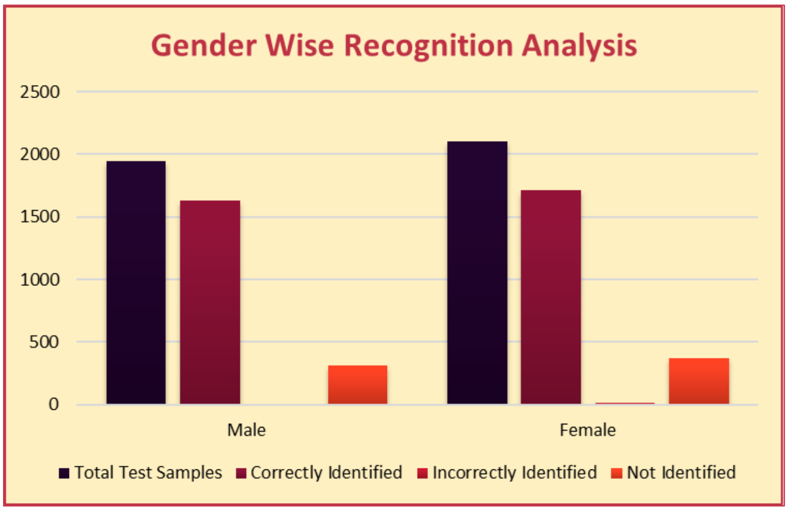

*Figure 6- 8 Gender Wise Recognition Analysis*

The statistics show that the model correctly identified 96.00% of male test samples and 94.87% of female test samples shows in Figure 6-6. While the male category has a slightly higher accuracy, both genders show strong performance with minor misclassifications (6 incorrect male and 18 incorrect female samples) and a small number of test samples not identified at all (313 male and 371 female samples) shows in Figure 6-5.

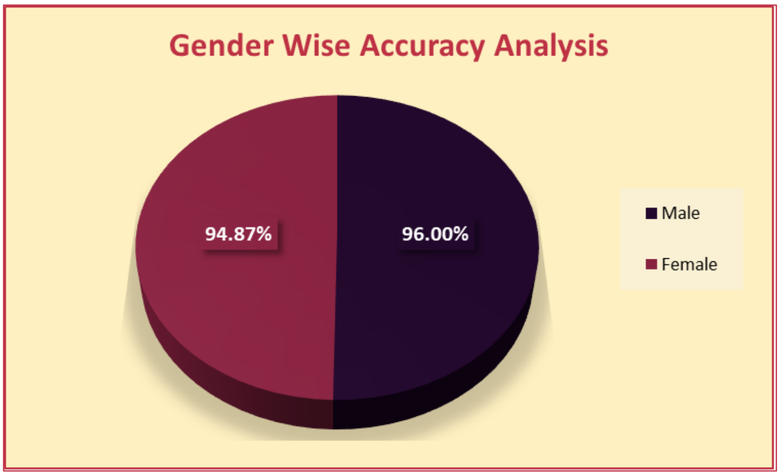

*Figure 6- 11 Gender Wise Accuracy Analysis*

### 6.2.3 Dialect Wise Recognition Analysis

The performance analysis statistics shows in Table 6-6 of the regional dialect classification model shows the differences of the accuracy of the four dialects. Results showed that for 73.3% of Kathiyawadi samples, they were correctly identified while 26.2% were not identified this suggest that there is scope for improvement in identification. Standard Gujarati yielded the best results with 87.1% correct identifications with a low value of 12.3% for samples the model was unable to correctly identify. Likewise, Surti performed 87.8%, and only 11.6% of the samples were not recognized again indicating good model performance shows in Figure 6-7.

*Table 6- 6 Dialect Wise Recognition Analysis*

| Dialect | Total Test Samples | Correctly Identifie d | Incorrectly Identified | Not Identified |
|---|---|---|---|---|
| Kathiyawadi | 1065 | 781 | 5 | 279 |
| Standard Gujarati | 2514 | 2189 | 17 | 308 |
| Surti | 311 | 273 | 2 | 36 |
| Kutchhi | 165 | 104 | 0 | 61 |

However, there were some problems; 63.0% of samples were correctly classified and 37.0% were not but there were no misclassifications. This mean that even though the model performs good every time it has correctly classified it lacks recognition in Kutchhi. Therefore, the model shows maximum accuracy in identifying Standard Gujarati and Surti, and only needs tuning and enhanced data set for Kathiyawadi and Kutchhi for minimizing unidentified samples and enhancing the assessment percentile.

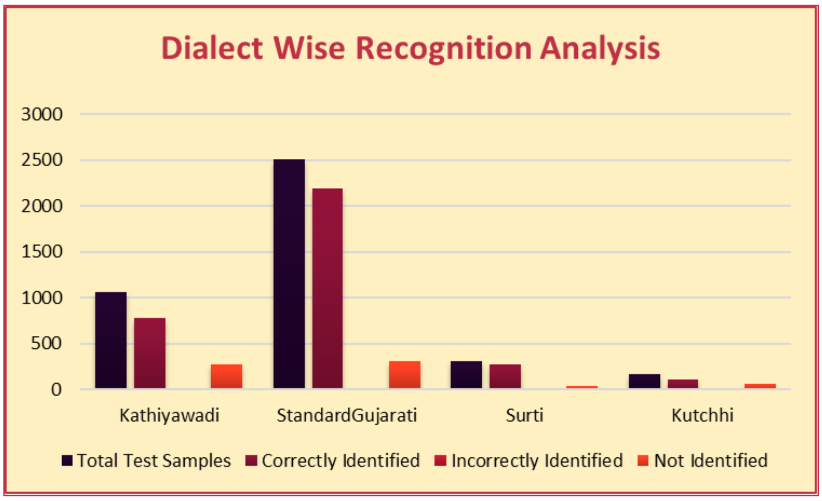

*Figure 6- 14 Dialect Wise Recognition Analysis*

### 6.2.4 Duration Wise Accuracy Analysis

It should be noted that the data in terms of the model’s accuracy by duration demonstrate the dependence on the length of the input. By applying the same model, the accuracy for samples less than 10 seconds duration was 84.75% which means, it works fairly good but still there is room for improvement in accurately identifying these short samples. However, for samples with a duration of more than ten seconds, the increase of accuracy was much higher, precisely 95.41%, which proves that the developed model is much more efficient in case of inputs with high durations. As such, it may be expected that the model uses some additional features or context information that is used in longer samples, and it is also noted that general performance can be improved for samples of lesser duration and strategies such as features extraction or direct modifications of the model’s parameters may be applied.

*Table 6- 7 Duration Wise Accuracy Analysis*

| Duration | Accuracy |
|---|---|
| Duration <10s | 84.75% |
| Duration >10s | 95.41% |

The graph shows in Figure 6-8 the analysis of accuracy for various audio samples of different duration and audio length:

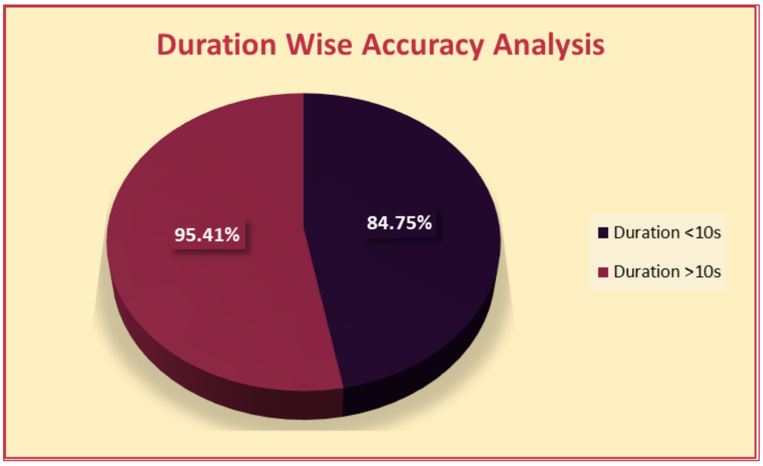

*Figure 6- 17 Duration Wise Accuracy Analysis*

## 6.3 Results & Discussions

Researcher has determined various different objectives for the proposed research work and able to achieve most of the objectives during research work. Result of these objectives are mentioned as follow and detail result analysis of proposed research is as discussed in above section 6.2.

- Researcher has collected linguistically rich dataset of 4 vernacular Gujarati dialects: Kathiyawadi, Standard Gujarati, Surti and Kutchhi. The dataset consists of large number of audio samples from various public resources like Online Repository, Gujarati Movies, Call/ Voice Recordings from Relatives and News from radio and Televisions. The description of dataset collection has discussed in section 5.2.

- Dataset Collected for proposed work of recognizing speaker's has been organized in structural format according to Global directory structure as discussed in section 5.3.

- For proposed work researcher has studied and further analysed Gujarati dialects and for their structural features. Based on the analysis, researcher is able to extract unique structural feature and other category belongs to speech which is discussed in detail in section 3.2. Section 3.3 and 3.4 shows different Feature extraction techniques and evaluate the best techniques for the model. Section 3.6 discuss the classification techniques of deep learning for the proposed model.

- Researcher has proposed the framework for vernacular voice recognition system and Voice Recognition described in section 4.2 and 4.3 respectively.

- Researcher has developed a Graphical User Interface (GUI) - software tool “Meera’s Voice Recognition Tool (MVR Tool)" that performs voice recognition. To achieve this objective user Interface is designed using Tkinter Library using which user can upload audio files and software tool will provide output provided as a result of processing carried out by Voice Recognition Model. Functionalities and features of this tool is disused in detail in section 5.4. Pseudo code of it is presented in section 5.5. User interface named “MVR” – Meera’s Voice Recognition Tool.

- Researcher has an objective to test proposed dataset of voice recognition in using Voice Recognition Tool user interface. For which all the audio samples were tested and performance analysis generated by Voice Recognition Model is presented in section 6.2.

## 6.4 Conclusion

Voice recognition is widely studied research area since few decades. Due to wide range of applications and challenges involved for recognizing speaker’s voice, researcher has chosen this area for recognizing handwritten character patterns for Gujarati script. Researcher has analysed all Gujarati dialects for their certain structural features for unique classification. Further Voice Recognition Model and Framework is proposed having series of operations to be performed as found generally in any voice recognition system. To implement and test performance of proposed Model scope of research work is determined i.e. researcher has identified 4 Gujarati dialects for recognition. These Model and Framework is implemented using pseudo code presented here in this thesis using various library and for evaluation of it Voice Recognition Tool is designed.

Proposed Voice Recognition Model is found to be satisfactory for recognition of four Gujarati dialects by evaluating its success ratio of correct recognition as presented here in this chapter.

## 6.5 Future Scope

The aim for future scope of the Voice Recognition System for Gujarati Regional Dialects is to improve its limitations by making it robust, adaptable and user friendly. One of the most important things they need to work on is to broaden its coverage of Gujarati dialects. Future development could be to create a larger and more diverse dataset to recapitulate the linguistic nuances and dialects of its different corners. In addition, when applied to an advanced phonetic modelling, the system can potentially better understand and process the special characteristics of each dialect.

Another critical improvement is the integration of a powerful voice verification mechanism. This will allow for secure and trusted user authentication which is integral to many high trust applications. Additionally, the system can be improved further by either, or both, dynamic retraining functionalities that would enable seamless incorporation of new speaker profiles without requiring complete retraining of the system for increased adaptability and user centricity. Future iterations can provide advanced noise cancellation algorithms and deep learning based on speech enhancement models to enhance performance in noisy environments. These features would allow isolation of the speaker’s voice from background noise to allow for recognition in challenging real-world environments. Also, real time processing and scalability is also possible; with cloud-based solutions or edge computing, efficient deployment across numerous devices and platforms can be achieved.

In addition to enabling multimodal interaction frameworks that integrate voice recognition with other modalities, such as text, facial recognition, and gestures, the system's future scope also includes integration into the system itself. It would also broaden its application in Education, healthcare and public services. Furthermore, the system can also be applied to the region of assistive technologies, helping people with disabilities or preserving and spreading Gujarati meanings to speech through transcription and pedagogical applications. Last, another approach is to give the system out as an open platform for customization and community promotion of innovation. The system allows researchers and developers to make it relevant and adaptable to new needs in a timely manner. These developments make Voice Recognition System for Gujarati Regional dialects a potential highly impactful and versatile technology capable of catering demands of various users and applications.

## References

[1] K. S. Bhogale et al., “Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages,” arXiv.org, Aug. 26, 2022. https://arxiv.org/abs/2208.12666 .