Automated Recognition of Hindi Word Audio Clips for Indian Children using Clustering-based Filters and Binary Classifier

Anuj Gopal
Pratham Education Foundation
anuj.gopal@pratham.org

Abstract

Speech recognition systems have made remarkable progress in the last few decades but most of the work has been done for adult speech. The rise of online learning during Covid-19 pandemic highlights the need for voice-enabled assistants for children so that they can navigate the menus and interfaces seamlessly. Speech recognition for children will also be very useful to develop automated reading assessment tools. However, such technology for children is challenging for a country like India where differences in accents, diction and enunciation is significant but available children speech data is limited. Through this paper, I tried various approaches to recognize hindi word audios. Commercially available Google Speech-to-Text performs poorly with only 49.7% accuracy at recall of 0.24 while recognising audio samples containing hindi words spoken by children. Using the same dataset, I experimented with clustering algorithm and logistic regression and found that the accuracy improves upto 81% with logistic regression. The paper also highlights the importance of data preprocessing by performing noise reduction using Butterworth low pass filters.

Keywords: *EdTech, Reading Skills, Assessment, Speech processing,Voice Command*

1 Introduction

According to a study by KPMG (2017), India's online education is set to grow to 9.6 million users by 2021. Online device-enabled education has been on a growth ever since the pandemic hit and it is estimated to rise even more in the coming years, with colleges and universities modifying their systems in accordance with the technological surge and Edtech companies investing heavily on creating platforms for the education of the future generation. According to Research and Markets (2019), the online education market will reach $350 Billion by 2025. This highlights the need for designing systems for children to have frictionless interactions with technology. Voice-enabled systems can play a crucial role in accessibility of education systems at an early age but recognizing words spoken by children is still a challenge. There has been limited progress in speech technology for children due to limited data.

Moreover, Speech recognition technology can play an important role in automating reading level assessments of children. With voice-enabled devices, the complete education system can be better controlled and automated, leading to high quality teaching and learning(Motiwalla, 2009; Azeta et al., 2010), especially for children as they are at a crucial stage of development.

This paper presents my experience with assessment of word reading audios by developing an automated assessment system for one of the low resource languages - Hindi which is the medium of instruction in government schools in many states of India. The children were of the age group 8-14 years from rural areas of India. The proposed Binary Classifier model uses Logistic Regression as the algorithm, along with preprocessing techniques using K-means clustering and Butterworth Low Pass filters to obtain an overall validation accuracy of 81.53%The main contributions of this paper are:

- A robust speech recognition methodology based on Binary Classifier catering to the needs of Indian regional languages like Hindi, Marathi etc.

- A preprocessing based audio classification methodology specifically designed for children which can be incorporated into existing education system to enhance reading level of children

2 Related Works

A multitude of similar works have been done in the past. In this section, I summarize previous work related to speech recognition for children and/or for Indian regional languages:

2.1 Highly accurate children's speech recognition for interactive reading tutors using subword units(Hagen et al., 2007)

This paper presents methodologies for advancement in speech recognition in the context of an interactive literacy tutor for children that aims to improve accuracy and modelling capabilities. To improve oral reading recognition, a more focused approach towards a novel set of speech recognition techniques is presented, where they have also shown that error rates for interactive read aloud can be reduced by more than 50% through a combination of advances in both statistical language and acoustic modeling.The efficacy of the approach is demonstrated using data collected from children in grades 3–5, namely 34.6% of partial words with reasonable evidence in the speech signal are detected at a low false alarm rate of 0.5%

2.2 Automatic Speech Recognition Systems for Regional Languages in India(Bachate and Sharma, 2019)

This paper discusses various aspects of building an automated speech recognition system, the parameters affecting the development of speech recognition systems, tools and techniques used and also research done on regional languages. The paper concludes that the Deep Neural network provides a better and a more accurate way of recognising speech.

2.3 ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages(Singh et al., 2020)

This paper provides a systematic survey of the existing literature related to automatic speech recognition (i.e. speech to text) for Indian languages. The survey analyses the possible opportunities, challenges, techniques, methods and to locate, appraise and synthesize the evidence from studies to provide empirical answers to the scientific questions. The survey was conducted based on the relevant research articles published from 2000 to 2018. The purpose of this systematic survey is to sum up the best available research on automatic speech recognition of Indian languages that is done by synthesizing the results of several studies.

2.4 Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation(Kumar and Aggarwal, 2020)

In this work, they have selected the Time-delay Neural Network (TDNN) based acoustic modeling with i-vector adaptation for limited resource Hindi ASR. The TDNN can capture the extended temporal context of acoustic events. To reduce the training time, they used sub-sampling based TDNN architecture in this work. Further, data augmentation techniques have been applied to extend the size of training data developed by TIFR, Mumbai. The results show that data augmentation significantly improves the performance of the Hindi ASR. Further, 4% average improvement has been recorded by applying i-vector adaptation in this work. They found the best system accuracy of 89.9% with TDNN based acoustic modeling with i-vector adaptation.

2.5 Syllable based Hindi speech recognition(Bhatt et al. ,2021)

In this paper, one of the acoustic units of speech, the syllable, is used for the development of a continuous Hindi speech recognition system. Earlier research works related to Hindi speech recognition were performed using the word, phoneme, and context-dependent models. The authors proposed a syllable based Hindi speech recognition system in this study due to different advantages of syllable units such as longer acoustic units, fast decoding, reducing contextual effects, and reduction of irregularities due to phonemes. The continuous Hindi speech recognition system was developed utilizing syllable based acoustic units. The research outcomes reveal that by using syllables, the performance of the system was increased by 27% than phoneme and 20% than triphones.

3 Dataset

The dataset used was collected using a novel data collection methodology (Agarwal et al.,2020) where an Android app was designed to collect human evaluated training data to collect data from three states of India - Maharashtra, Uttar Pradesh and Rajasthan. While conducting the test, the audio clips were recorded for each section and the assessor marked the correctness of every question. The details of the test were further recorded in a json file. Although dataset contains letters, paragraphs and stories too, I only used hindi words for my analysis. The details of dataset is shown in the following table:

No. of audio files	1071
Unique word count	37
Total Duration	51.56 mins
Avg Duration	3.13s
Total FileSize	415.29MB
Avg FileSize	430.43KB
Avg Unique Speakers	30

Table 1: Statistics of dataset used

I have also tried running the model using few audio files from other word samples to be used as 'False' so that there is an increase in the training dataset. But the accuracy obtained is lower than the one without using those audio files. The possible reason for this might be the difference in pattern learning for different words, which primarily depends on their frequency response and temporal variations. Finally, the dataset was divided into 70% Training and 30% Testing data.

4 Methodology

The proposed methodology utilised three different interconnected algorithmic approaches to obtain optimum results for the problem proposed:

4.1 Audio Quality based data segregation using Clustering algorithms

The clustering of audio files is important for understanding the features of the signals, as the audio signals are nothing but high-dimensional data. As presented in EURASIP research(Lil et al., 2017), distance calculation failures, inefficient index trees, and cluster overlaps, derived from the equidistance, redundant attribute, and sparsity, respectively, seriously affect the clustering performance.

Mel-frequency cepstrum coefficient(MFCC) is one of the most important features of any audio signal, which is extracted from the files using a multistep process starting with signal windowing followed by Direct Fourier Transformation. The next steps include taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse Discrete Cosine Transform.

I propose a K-means clustering for segregation of inaudible/low quality audio clips using the means of MFCC values for each of the sound signals.The MFCC features of a single audio file were clustered to observe differences in patterns within the audio, to detect human voice and silences, and to observe other variations. This can not only remove noise and silences as a preprocessing part, but can also provide valuable patterns for the whole audio file.There are differences in features for the two files as expected from audios as well.

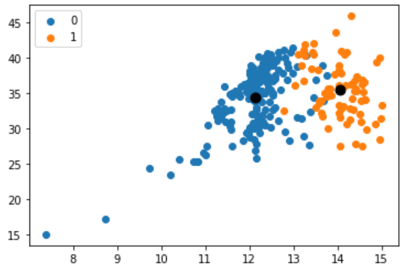

Figure 1:Plotting MFCC using 2 coefficients

The average value of mean MFCC for audio files that were Google STT transcripted is 1.23 while that for empty audio files is 0.72.

The K-means clustering method with reduced MFCC values [taking means] provided us with the sample set that can be directly fed to models. And, since plots of using MFCC mean values show patterns as shown, I used 264 audio files with label 0 as shown here, to infer that the accuracies are positively enhanced as the clusters are formed based on energy/frequency ranges that segregated bad quality audio files to separate clusters.

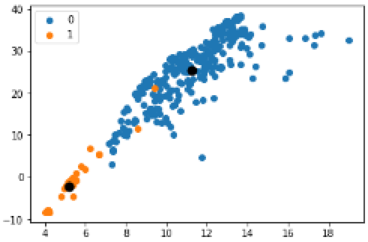

Figure 2: K-means clustering plot

4.2 Noise Reduction using Butterworth Low Pass Filters

The Butterworth filters are designed for signal processing which have flat frequency response in the pass-band and in the stop-band they have a zero roll off response. They are widely used filters in video motion analysis and in audio signals.The filter-frequency adjustability allows the user to remove noise without rolling off the signal. Variable low-pass filtering deals with aliasing by passing no more than 1% of the high-frequency content that causes aliases. This ensures that only the aliased frequencies are removed, not the signal of interest.Tunable digital filters are widely used in medical electronics, digital audio instrumentation, telecommunications and control systems. It is often the essential requirement for removal of noise from the audio signal.

4.3 Logistic Regression based Binary Classifiers

Normal logistic regression maximizes the following log-likelihood function:

$$\ell(\beta) = \Sigma[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)]$$

Since smaller datasets are more prone to overfitting and have outliers which can cause significant vulnerability in the model, regularization techniques such as Ridge regression and Lasso regression are used along with tuned Logistic Regression to add penalties which create a balance between producing accurate predictions and producing smaller coefficients. The objective of any model is to learn patterns from the available training dataset and generalise relationships between the dependent and the independent variables. At the same time, the model should be biased towards the training dataset and therefore a general pattern recognition approach is preferred over more specifically accurate models compatible only with the dataset provided to train. Complex models perform worse since they have a higher possibility of overfitting according to the dataset.

5 Experimental Results

The summarised experimental results are shown below for 4 samples of different words. The table contains word samples, their total instances used for the analysis and accuracy obtained using various machine learning models:

Word	Total	SVC	Logistic
लाल	66	0.70	0.85
ताला	57	0.78	0.89
नाक	42	0.60	0.80
आग	57	0.78	0.89

Table 2: Accuracy of ML models

The Logistic Regression with aforementioned preprocessing techniques provides the best results. The overall accuracy of the proposed method is 81.53% with a precision of 0.84 and a recall of 0.94.

6 Error Analysis

The small dataset poses a variety of problems in the modeling.The first is lack of available information, which produces less accurate models. Secondly, the small dataset does not reflect the population's distribution in an accurate way, thus creating a necessary condition to preprocess datasets and work sensitively to remove any kind of noise in the data. With audio data, the problem becomes more as the pattern recognition techniques available for audio datasets are relatively new and most of them use an indirect methodology of learning, converting files to required formats to extract necessary information. Getting closer to true population parameters, thus, becomes extremely important. The fluctuations in accuracy measures as the dataset reduces in number becomes extremely vulnerable to variations below 200 observations(Canario, 2020).

Using regularization techniques to penalise predictions leading to overfitting, I generalised Logistic Binary Classification to optimum requirements. The regularization techniques used was Ridge Regression which works well to reduce overfitting, optimizing the desired performance(Salehi et al., 2019), to finally develop a parsimonious model, creating a less complex, yet a more accurate model. Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (lbfgs) was used as the solver for Logistic Regression as it performs best for small datasets, using low memory and utilising Quasi-Newton methodology(Dennis and Moree, 1977).

7 Conclusion

In this paper, I have presented Audio Quality based data segregation using Clustering algorithms, followed by Noise Reduction using Butterworth Low Pass Filters and finally Logistic Regression based Binary Classifiers to achieve optimum speech recognition results for dataset containing audio clips for Indian Children reading Hindi words as part of their reading level assessment. While Google Speech-to-Text could transcribe 49.7% of the audio with a recall of 0.24, our method of Logistic Regression based Binary Classifier with preprocessing resulted in an overall validation accuracy of 81.53%. This can be used to develop advanced learning tools for children in the future and to also significantly enhance Voice command based devices specific to regional languages.

8 References

Ambrose Azeta, Charles Ayo, Prof. Aderemi Atayero, and Nicholas Omoregbe. 2010. Intelligent Voice-based E-education system: A framework and evaluation. *International Journal of Computing*. 9. 327-334.

Amitoj Singh, Virender Kadyan, Munish Kumar, and Nancy Bassan. 2020. ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. *Artificial Intelligence Review*, 53(5), pp.3673-3704.

Andreas Hagen, Bryan Pellom, and Ronald Cole. 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, 49(12), pp.861-873.

Ankit Kumar and Rajesh Kumar Aggarwal. 2020. Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation. *International Journal of Speech Technology*, pp.1-12.

Dolly Agarwal, Jayant Gupchup, and Nishant Baghel. 2020. A Dataset for measuring reading levels in India at scale. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 9210-9214). IEEE.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. 2019. The impact of regularization on high-dimensional logistic regression. arXiv preprint arXiv:1906.03761.

John E. Dennis, Jr. and Jorge J. Moree. 1977. Quasi-Newton methods, motivation and theory. SIAM review, 19(1), pp.46-89.

KPMG in India and Google, 2017. *Online Education in India: 2021*

Luvai F. Motiwalla. 2009. A Voice-Enabled E-Learning Service (VòIS) Platform. In *Fifth International Conference on Networking and Services 2009*, (pp. 597-602). IEEE.

Ravindra Parshuram Bachate and Ashok Sharma. 2019. Automatic speech recognition systems for regional languages in India. *International Journal of Recent Technology and Engineering*, 8(2).

Remy Canario, 2020. *The Best Classifier for Small Datasets: Log-F(m,m) Logit*

Research and Markets, 2019. *Online Education Market Global Forecast, by End User, Learning Mode (Self-Paced, Instructor Led), Technology, Country, Company*. ID: 4876815.

Shobha Bhatt, Anurag Jain, and Amita Dev. 2021. Syllable based Hindi speech recognition. *Journal of Information and Optimization Sciences* 41, no. 6 (2020): 1333-1351.

Wenfa Li1, Gongming Wang, and Ke Li. 2017. Clustering algorithm for audio signals based on the sequential Psim matrix and Tabu Search. *EURASIP Journal on Audio, Speech, and Music Processing 2017*, pp.1-9.