A Novel Approach for Speech Recognition of Malayalam Using Enhanced MFCC Features

Muhammed Shafi M[1*] D. Muhammad Noorul Mubarak[2*] S.A. Shanvas[3*]

[1*]Research Scholar, Department of Linguistics, University of Kerala
[2*]Associate Professor and Head, Department of Computer Science, University of Kerala
[3*]Professor and Head, Department of Linguistics, University of Kerala

Abstract: The study Investigates into the realm of speech recognition, particularly focusing on the evaluation of Hidden Markov Models (HMM), leveraging the MalayalamVoice dataset. The research scrutinizes the performance of HMM concerning word error rate (WER) and accuracy across varied word lengths, revealing a consistent trend of increased WER and decreased accuracy with longer utterances. This underscores the challenges inherent in accurately transcribing extended speech segments, accentuating the necessity for algorithmic enhancements. Moreover, analyses across diverse datasets and noisy environments underscore the criticality of comprehending dataset characteristics for refining recognition algorithms. Additionally, comparisons of different feature extraction methods elucidate the efficacy of Enhanced MFCC, particularly for shorter word lengths. However, as the word length extends, the distinctions between extraction methods diminish, highlighting the multifaceted nature of speech recognition. Overall, this study underscores the intricacies involved in speech recognition and the imperative of algorithmic refinements for augmenting accuracy, especially in practical scenarios.

Key words: Speech recognition, HMM, MFCC, Signal processing, Speech to text

Introduction

The recognition of speech samples from visually impaired individuals is crucial for the operation of a smart cane designed for their assistance. By utilizing speech commands, visually impaired users can efficiently navigate their surroundings with the smart cane. This technology allows them to communicate their intentions and receive feedback or guidance accordingly, enhancing their independence and mobility. Integration of GPS technology into the system further enhances its functionality. With GPS, the smart cane can provide real-time location information to the user, aiding in navigation and ensuring they can reach their desired destinations safely. This combination of speech recognition and GPS technology creates a powerful tool for visually impaired individuals, offering them greater autonomy and confidence in their daily travels.
Overall, the incorporation of speech recognition and GPS technology into the smart cane system represents a significant advancement in assistive technology for the visually impaired, empowering them to navigate the world with greater ease and independence.

Architecture of the Speech Recognition Module

The Speech recognition system in this study for visually impaired individuals toperates through the following steps:

- **Audio Input**: The user speaks commands or instructions into a microphone integrated into the smart cane.
- **Signal Processing**: The audio signal captured by the microphone is processed to enhance its quality and extract relevant features. This processing may involve noise reduction, filtering, and other techniques to improve the accuracy of speech recognition.
- **Feature Extraction**: From the processed audio signal, key features of the speech, such as frequency components and timing patterns, are extracted. These features are then used as input for the speech recognition algorithm.
- **Speech Recognition Algorithm**: The extracted features are analyzed by a speech recognition algorithm, which matches them against a predefined set of speech commands or vocabulary. This algorithm may use various techniques such as Hidden Markov Models (HMMs).
- **Command Interpretation**: Once the speech is recognized, the system interprets the command or instruction conveyed by the user. This interpretation involves mapping the recognized speech to specific actions or functionalities within the smart cane system.
- **Response Generation**: Based on the interpreted command, the smart cane generates an appropriate response. This response may involve providing auditory feedback to the user, initiating navigation instructions, or activating other features of the smart cane.

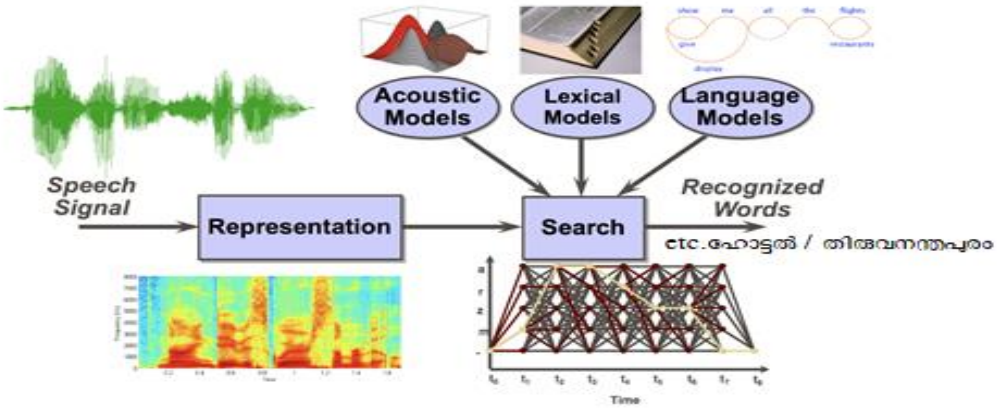


Fig 1: Architecture of the Speech Recognition Module

Methodology

In the initial phase of speech recognition, the raw audio waves are fed into a computer, where they are represented as simple 1D signals. Since raw audio is essentially an analog signal, it needs to be transformed into a digital form that the computer can comprehend. This transformation is achieved through a process called sampling, wherein the height of the sound wave is recorded at equally spaced intervals. For speech recognition purposes, a typical sampling rate of 8KHz or 16KHz is considered sufficient to capture the frequency range of human speech. Each sample is then quantized, typically into 8-bit or 16-bit values. By ensuring that the signal is sampled at least twice the highest frequency to be recorded, the original sound wave can be faithfully reconstructed from the spaced-out samples. Thus, a 1-second audio clip can be dissected into individual samples, denoted as a1, a2, ..., an, facilitating subsequent processing and analysis for speech recognition tasks.
samples $a_1, a_2, \ a_n$. This can be represented as1D vector=A=[$a_1, a_2, \ a_n$]

Denoising the Signal

In the context of audio signal denoising, a multi-step process is employed. Initially, the noisy signal undergoes transformation from the time domain to the frequency domain via the Fourier Transform, such as the Fast Fourier Transform (FFT), to acquire its spectrum. Subsequently, the noise spectrum is estimated, often derived from silent segments within the signal or a separate recording of noise. This estimation typically involves computing the average or median of multiple frames' spectra. Following the noise spectrum estimation, spectral subtraction is applied, involving the subtraction of the estimated noise spectrum from the spectrum of the noisy signal. This subtraction operation aims to mitigate the presence of noise components within the signal. Finally, the denoised spectrum is converted back to the time domain through the Inverse Fourier Transform, yielding the denoised audio signal. This comprehensive approach to denoising effectively reduces noise interference, enhancing the overall quality and clarity of the audio output.
$Y(f)=\max(X(f)-\alpha N(f),0)$
Where:
$Y(f)$ is the denoised spectrum.
$X(f)$ is the noisy signal spectrum.
$N(f)$ is the estimated noise spectrum.
α is a scaling factor or smoothing parameter (typically between 0 and 1), which controls the amount of noise subtraction.

Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) have emerged as a pivotal feature in speech and voice recognition, with their inception credited to Davis and Murmelstein in 1980.

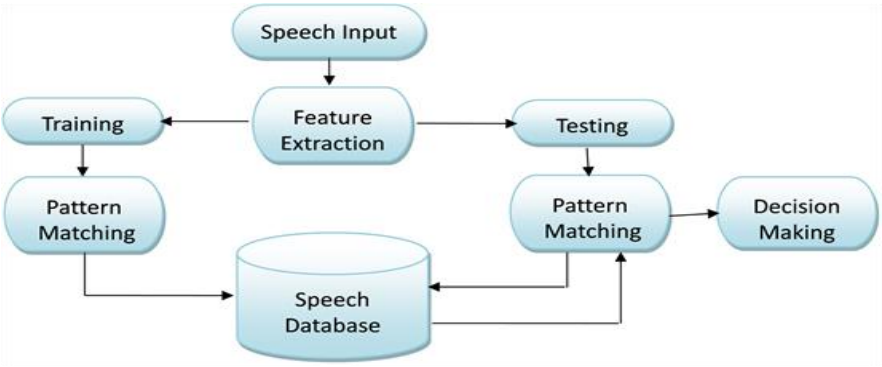


Fig 2: Block Diagram of the Speech Recognition Module

Initially conceived to emulate the perceptual characteristics of the human ear, MFCCs have found application beyond speech processing, notably in the analysis of ship-radiated noise. By mimicking auditory perception, MFCCs excel in compressing spectral information, reducing dimensionality, and enhancing recognition accuracy. Central to the computation of MFCCs is the conversion between linear frequency and Mel frequency, encapsulated in the following formula:

$$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right)$$

where Mel(*f*) denotes the Mel frequency and *f* denotes the Fourier frequency.

Mel Frequency Cepstral Coefficients (MFCC) serve as a technique for extracting spectral characteristics from sequences of frames. The process involves transforming the signal into the frequency domain using Fast Fourier Transform (FFT), as described by Equation (2.1). Subsequently, following pre-emphasis, framing, and windowing of the input signal, FFT is applied to the speech frames, resulting in the derivation of certain parameters based on 256-point power spectrum. This power spectrum is then converted into a Mel-frequency spectrum utilizing Equations (2.2) and (2.3). Finally, the logarithm of this spectrum is computed, and its inverse Fourier transform is performed, as illustrated in Figure 2. This pivotal transformation, facilitated by Mel-scale filter banks, imbues the spectral representation with a characteristic shape mirroring the nonlinearities inherent in human auditory perception. The transformation equation, encapsulating this pivotal mapping, stands as follows:

$$X(k) = \sum_{N=0}^{N-1} x[n]\mathrm{e}^{-j2\pi nk/N},\ 0 \le k < N$$

$$F_{mel} = 2595 \cdot \log_{10}(1 + \frac{f_{Hz}}{700})$$

$$F_{Hz} = 700 \cdot (10^{\frac{F_{mel}}{2595}} - 1)$$

This equation epitomizes the essence of MFCC extraction, wherein the spectral information undergoes a profound metamorphosis, aligning it intricately with the perceptual intricacies of the human auditory system.

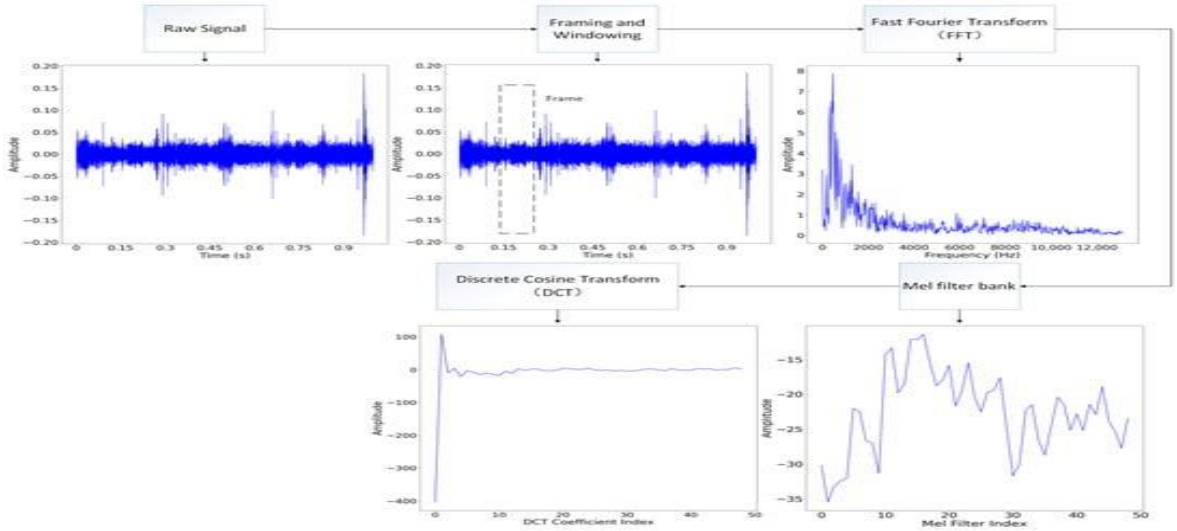


Fig 4: The extraction process of MFCC.

3Pre–emphasis

Pre-emphasis is the first step in MFCC which will boost the amount of energy of signal at higher frequencies. This step processes the passing of signal through a filter which emphasizes higher frequency in the band of frequencies the magnitude of some higher frequencies with respect to magnitude of other lower frequencies in order to improve the overall SNR. This process will increase the energy of signal at higher frequency.

Pre-emphasis employs a filter to amplify higher frequencies. The following illustrates the signal before and after processing, demonstrating how the high-frequency components are enhanced.

$$x'[t_d] = x[t_d] - \alpha x[t_d - 1] \qquad 0.95 < \alpha < 0.99$$

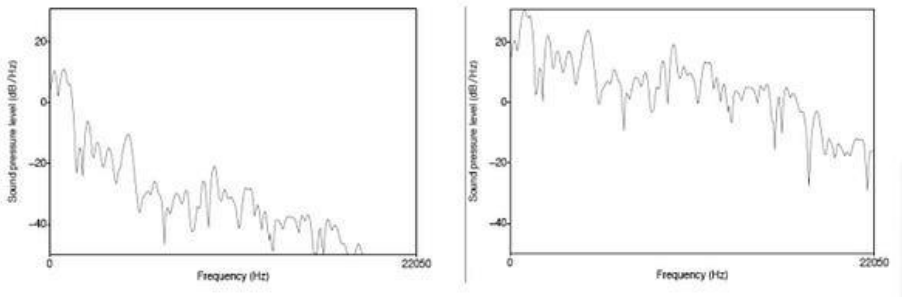


Fig 5: Frequency variations after preephasising Framing

The process of segmenting the sampled speech samples into a small frames. The speech signal is divided into frames of N samples. Adjacent frames are being separated by M(M < N). Typical values used are M = 100 and N = 256 (which is equivalent to ˜ 30 m sec windowing

The feature extraction process is implemented using Mel Frequency Cepstral Coefficients (MFCC) in which speech features are extracted for all the speech samples. Then all these features are given to pattern trainer for training and are trained by HMM to create HMM model for each word. Then Viterbi decoding will be used to select the one with maximum likelihood which is nothing but recognized word.

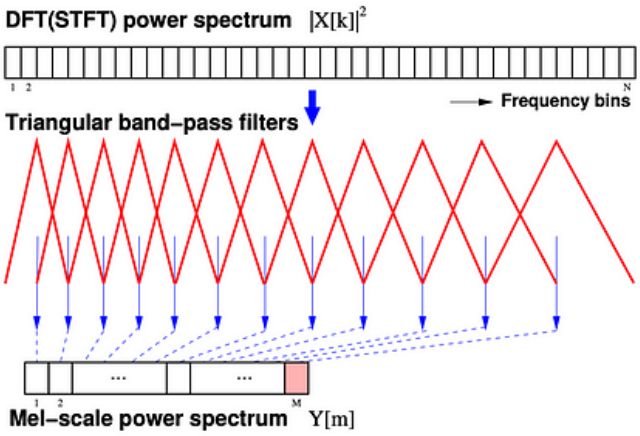


Fig 6: Generating Speech Frames

Hamming Windowing

Each individual frame is windowed so as to minimize the signal discontinuities at the beginning and end of each frame. Hamming window is used as window and it integrates all the closest frequency lines. The Hamming window equation is given as: If the window is defined as

W (n),0 n N 1

where N = number of samples in each frame

Y [n] = Output signal

X (n) = input signal.

$W(n)$ = Hamming window , then the result of windowing signal is shown below:

$$Y(n) = X(n) * W(n) \qquad (3.2)$$

$$W(n) = 0.54 - (-0.46)Cos(2n/N - 1); 0 < n < N - 1 \quad (3.3)$$

Mel Filter Bank Processing.

The frequency range covered by the spectrum obtained through Fast Fourier Transform (FFT) is extensive,

but the voice signal does not adhere to a linear scale. To better capture the characteristics of the voice signal, a bank of filters is applied according to the Mel scale.

Mel Filter Bank

The triangular filters calculate a weighted sum of spectral components from the filters, aiming to approximate a Mel scale in the output process. Each filter's frequency response magnitude follows a triangular shape, reaching unity at the center frequency and linearly decreasing to zero at the center frequency of two adjacent filters. Subsequently, the output of each filter is the sum of its filtered spectral components. The resulting Mel spectrum comprises the output powers of these filters. Next, the logarithm of the Mel spectrum is computed, resulting in the log-Mel spectrum output.

Discrete Cosine Transform

The next step involves converting the log Mel spectrum back into the time domain using the Discrete Cosine Transform (DCT). This transformation yields the Mel Frequency Cepstrum Coefficients, often referred to as acoustic vectors. Consequently, each input utterance undergoes a transformation into a sequence of these acoustic vectors.

MalayalamVoice" A Comprehensive Speech Dataset

In this Study, introduced the "MalayalamVoice" dataset creation process, intended to serve as a cornerstone in the advancement of research within the realms of speech recognition, natural language processing, and accessibility technologies, with a keen focus on the Malayalam language.

The MalayalamVoice dataset emerges as a seminal contribution to the research fraternity, offering an invaluable repository for the training and evaluation of speech recognition models and accessibility technologies tailored explicitly for the Malayalam language.

Dataset Description	
Language	Malayalam
Audio Format	WAV
Sampling Rate	48 kHz
Encoding	16-bit PCM
Speakers	25 male speakers 30 female speakers
Content Focus	Utterances of words in COCO dataset, Domain-specific sentences for visually impaired navigation assistance
Dataset Statistics	
Total Audio Files	3850
Total Duration	2Hr 45Sec
Average Duration per Audio File	2.3Sec

Table 1: Description of MalayalamVoice Data set

Hidden Markov Modeling

HMM serves the purpose of categorizing the features and making accurate decisions. Widely recognized as a potent statistical tool in speech recognition and speaker identification systems, HMM possesses the capability to model the intricate alignment of speech non-linearly and estimate model parameters. Gaussian Mixtures are also employed to model the emission probability distribution function within each state.

During the training process, various parameters including observation parameters, transition probability matrix, prior probabilities, and Gaussian distribution are re-estimated iteratively to optimize their values, as depicted in Figure 3. Consequently, these updated HMM parameters are utilized to compute likelihood scores, which play a crucial role in determining the best path between frames for recognizing the unknown word.

Evaluation Process

In the evaluation process, involving the observation sequence OO and the model parameters $\lambda\lambda$, the Forward ($\alpha\alpha$) and Backward ($\beta\beta$) algorithms are employed to determine the probability of the observation sequence given the model, denoted as $P(O|\lambda)P(O|\lambda)$ [12]. Illustrated in Figure 4, the forward and backward probabilities are combined to assess the likelihood that any sequence of states has generated the observed sequence of observations.

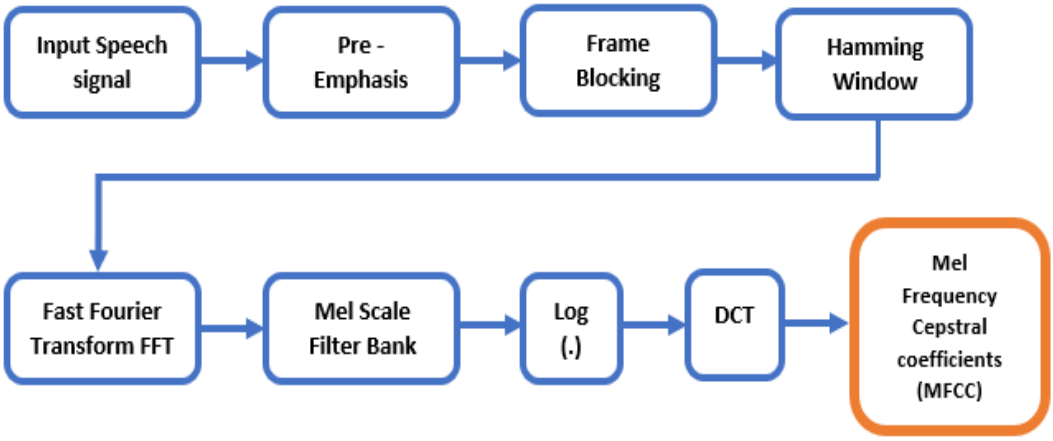


Fig 7: Mel Frequency Cepstral Coefficients (MFCC) block diagram

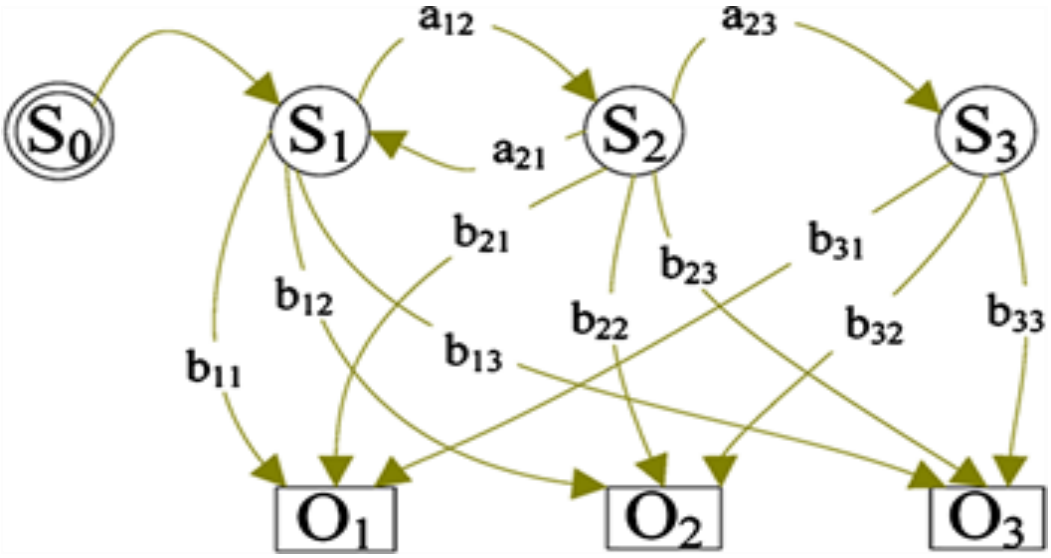


Fig 8:Three states hidden Markov model.

Training Process

With the observation sequence OO and model parameters $\lambda\lambda$ at hand, the Baum-Welch algorithm is employed to readjust and re-estimate the transition probability matrix and Gaussian mixture parameters, including mean and covariance, which optimally characterize the process. Additionally, the Baum-Welch algorithm is utilized to learn and encode the characteristics of the observation sequence, facilitating the recognition of similar observation sequences.

Decoding

In the decoding process, the Viterbi algorithm is employed to compare between the training and testing data, determining the optimal scoring path of the state sequence by selecting the highest probabilities between the model and the testing data. The maximal probability of state sequences is defined using Equation, and the selection of the optimal scoring path of the state sequence is calculated utilizing the following function:

$$\delta t(i) = \max(P(q(1), q(2), \cdots, q(t-1); o(1), o(2), \cdots, o(t)|\lambda))$$

Python Implementation:

Load testing list from .mat file defstart_recognition(testing_list_file,dim):testing_list=sio.loadmat(testing_list_file)

HMM with MalayalamVoice data

This study assesses the effectiveness of HMM through various performance metrics such as word error rate and accuracy, using the MalayalamVoice dataset. MalayalamVoice dataset mainly focused for creating

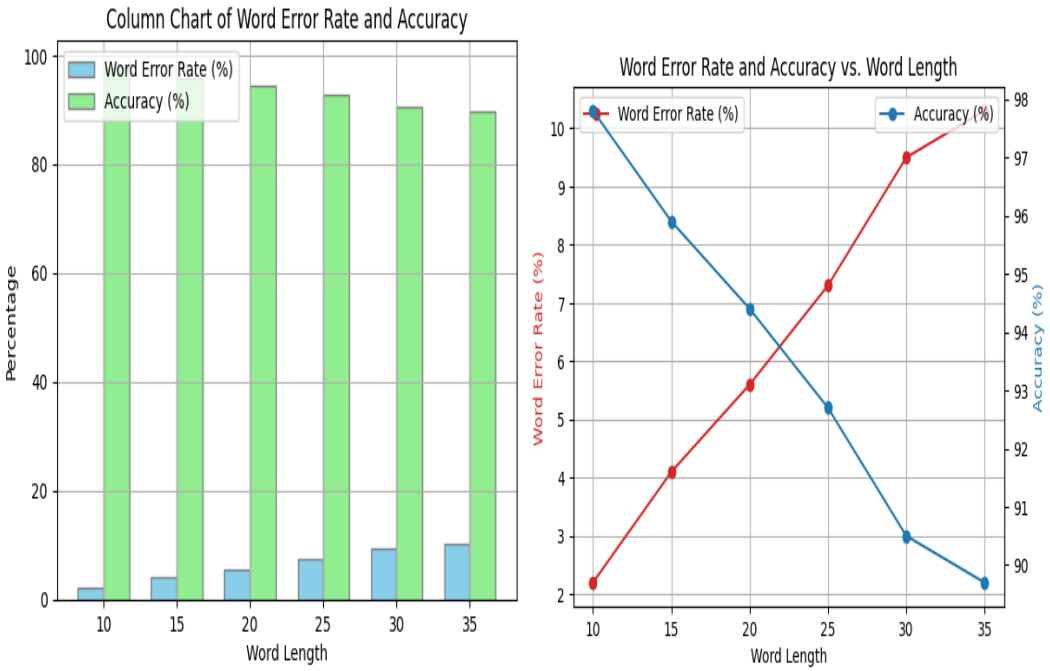
assistive technology like smart cane. The HMM speech recognition algorithm is employed initially on real-time noisy speech and subsequently on noise-filtered speech. The ensuing table, labeled as Table , encapsulates the performance evaluation of HMM.

Word Length	Word Error Rate (%)	Accuracy (%)
10	2.2	97.8
15	4.1	95.9
20	5.6	94.4
25	7.3	92.7
30	9.5	90.5
35	10.3	89.7

Table 2: Performance evaluation of HMM with MalayalamVoice data

The table presents a clear trend in the performance of a speech recognition system concerning word length, as indicated by Word Error Rate (%) and Accuracy (%). With an increase in word length from 10 to 35, there's a consistent rise in the Word Error Rate (%), climbing from 2.2% to 10.3%, indicating a greater likelihood of misidentification with longer words. Conversely, Accuracy (%) demonstrates a corresponding decline from 97.8% to 89.7% as word length extends, reflecting the system's reduced ability to accurately transcribe longer utterances. This inverse relationship underscores the inherent challenges in accurately recognizing and transcribing lengthier words within a speech recognition framework, emphasizing the need for targeted algorithmic enhancements to mitigate error rates and bolster accuracy, particularly in real-world applications where longer phrases or sentences are common.

WER= (NS+NI+ND) / N
Where NS=number of substitutions
NI = number of insertions
ND= number of deletions
N = Total number of words spoken
The accuracy percentage is calculated as: Accuracy = 100 – WER (3.2)



Equations illustrate a direct relationship between the terms WER and accuracy, indicating that changes in one variable correspond to proportional changes in the other.

Comparison with Other Malayalam datasets

Length of the word	MalayalamVoice	Indic Malayalam	Festvox IIITH Malayalam	MSC Reviewed speech
10	2.2	4.5	2.1	2.1
15	4.1	6.8	3.6	3.0
20	5.6	8.9	6.5	5.4
25	7.3	10.9	7.4	7.8
30	9.5	13.7	11.2	9.7
35	10.3	15.7	12.4	12.3

Table 3: Performance evaluationComparison with Other Malayalam datasets

The table presents error rates (%) for different word lengths across four distinct datasets: MalayalamVoice, Indic Malayalam, Festvox IIITH Malayalam, and MSC Reviewed speech. The data reveals a consistent pattern of increasing error rates with longer word lengths across all datasets. This suggests that as the length of words grows, speech recognition systems encounter greater difficulty in accurately transcribing spoken language. Additionally, variations in error rates between datasets highlight potential differences in the performance and characteristics of the respective recognition models or datasets. Understanding these trends is crucial for optimizing speech recognition algorithms and enhancing their effectiveness, particularly when processing longer words, thus contributing to advancements in speech processing technology.

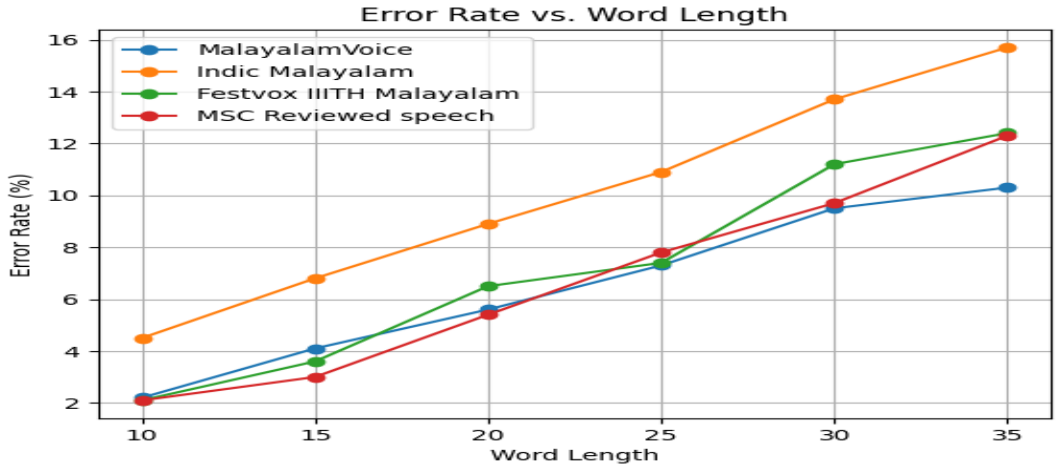


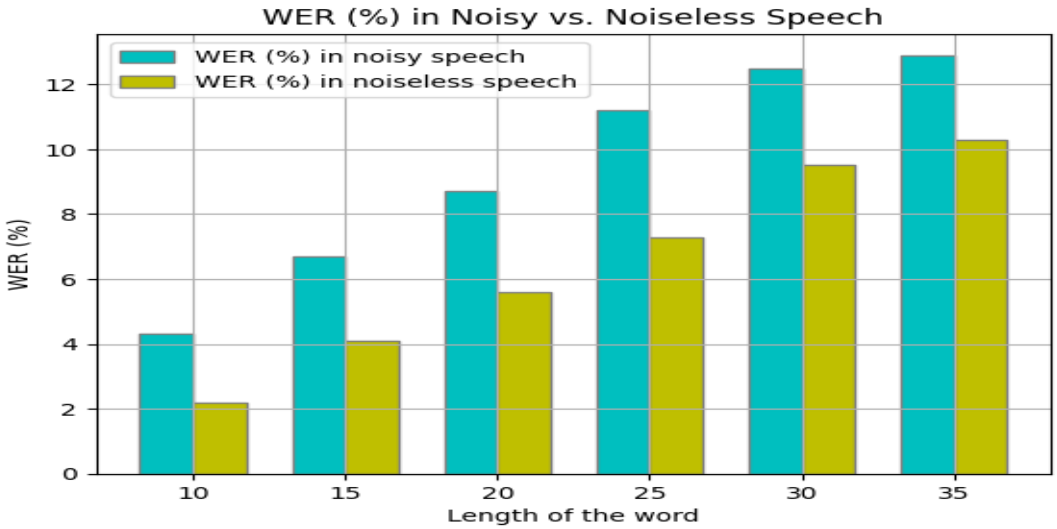
Fig 4: Performance evaluation Comparison with Other Malayalam datasets

Noisy and Noiseless Environment

Length of the word	WER (%) in noisy speech	WER (%) in noiseless speech
10	4.3	2.2
15	6.7	4.1
20	8.7	5.6
25	11.2	7.3
30	12.5	9.5
35	12.9	10.3

Table 4: Pperformance evaluation in noisy Environment

The provided dataset showcases the Word Error Rate (WER) percentages for both noisy and noiseless speech across varying word lengths. As we observe the table, there's a discernible trend indicating an increase in WER with longer word lengths for both noisy and noiseless conditions. For instance, at a word length of 10, the WER is 4.3% in noisy speech and 2.2% in noiseless speech. However, as the word length extends to 35, the WER escalates to 12.9% in noisy speech and 10.3% in noiseless speech. This pattern suggests that longer words pose greater challenges for accurate speech recognition across both noisy and noiseless environments. Furthermore, when comparing the WER between noisy and noiseless conditions, we consistently observe a lower WER in the noiseless speech scenario across all word lengths.

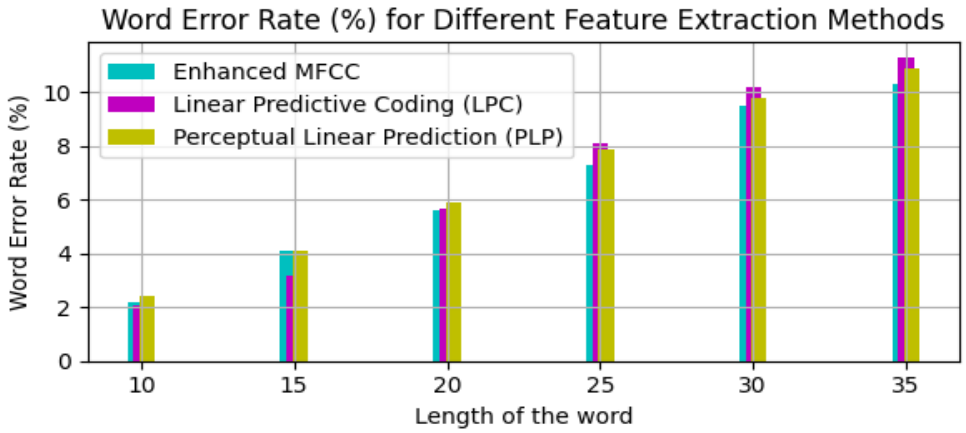


Performance Analysis with Different Feature Extraction Methods

Length of the word	Enhanced MFCC	Linear Predictive Coding (LPC)	Perceptual Linear Prediction (PLP)
10	2.2	2.1	2.4
15	4.1	3.2	4.1
20	5.6	5.7	5.9
25	7.3	8.1	7.9
30	9.5	10.2	9.8
35	10.3	11.3	10.9

Table 5: Performance Analysis with Different Feature Extraction Methods

The table showcases the performance of various feature extraction methods, including Enhanced MFCC, Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP), across different word lengths. Notably, for shorter word lengths such as 10 and 15, the Enhanced MFCC consistently demonstrates the lowest Word Error Rates (WERs) compared to LPC and PLP, with values of 2.2% and 4.1% respectively. However, as the word length increases, the WERs tend to converge, indicating a diminishing gap between the performance of different feature extraction methods. Particularly at longer word lengths of 30 and 35, while Enhanced MFCC still maintains a competitive edge with WERs of 9.5% and 10.3% respectively, the differences between the methods become less pronounced. This suggests that for longer utterances, factors beyond feature extraction may play a more significant role in determining recognition accuracy.



Conclusion

This study focuses into the evaluation of speech recognition algorithms, particularly focusing on Hidden Markov Models (HMM), using the MalayalamVoice dataset. Initially, the performance of HMM is analysed concerning

word error rate (WER) and accuracy across different word lengths, revealing a consistent increase in WER and decrease in accuracy as word length extends. This trend highlights the challenges in accurately transcribing longer utterances, emphasizing the need for algorithmic improvements. Furthermore, error rates across various datasets underscore the importance of understanding dataset characteristics for optimizing recognition algorithms. The chapter also explores the impact of noisy environments on speech recognition, noting an increase in WER with longer word lengths in both noisy and noiseless conditions. Additionally, comparisons between different feature extraction methods demonstrate the superiority of Enhanced MFCC for shorter word lengths, although with diminishing differences as word length increases. Overall, the chapter emphasizes the complexities involved in speech recognition and the significance of algorithmic enhancements for improving accuracy, particularly in real-world applications.

References

1. Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing 22, no. 10 (2014): 1533-1545.
2. Akila, & E. Chandra. (2013). Isolated Tamil Word Speech Recognition System Using HTK.
3. International Journal of Computer Science Research and Application, 3(2), 30-38.
4. Alghamdi, Mansour M., & Yousef Ajami Alotaibi. (2010). HMM Automatic Speech Recognition System of Arabic Alphadigits. Arabian Journal for Science and Engineering, 35(2C), 137-155.
5. Ali, Md. Akkas, Manwar Hossain, & Mohammad Nuruzzaman Bhuiyan. (2013). Automatic Speech Recognition Technique for Bangla Words. International Journal of Advanced Science and Technology, 50, 51-60.
6. Al-Qatab, Bassam A. Q., & Raja N. Ainon. (2010). Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK). In Proceedings of the International Symposium in Information Technology (ITSim) (Vol. 2, pp. 557-562). Kuala Lumpur.
7. Anusuya, M. A., & Katti, S. K. (2009). Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security (IJCSIS), 6(3), 181-205.
8. Dua, Mohit, R. K. Aggarwal, Virender Kadyan, & Shelza Dua. (2012). Punjabi Automatic Speech Recognition Using HTK. International Journal of Computer Science Issues, 9(4), 359-364.
9. Engin Avci & Zuhtu Hakan Akpolat. (2006). Speech recognition using a wavelet packet adaptive network based fuzzy inference system. Expert Systems with Applications, 31(3), 495–503.
10. Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, & Ather Mohsin Zaidi. (2010). Speaker independent Urdu speech recognition using HMM. In Proceedings of the 7th International Conference on Informatics and Systems (INFOS) (pp.1-5). Cairo.
11. Kayte, S. N., Mal, M., & Gujrathi, J. (2015, December). Hidden Markov Model based Speech Synthesis: A Review. International Journal of Computer Applications, 130, 975-8887. doi:10.5120/ijca2015906965
12. Mohit Dua, R. K. Aggarwal, Virender Kadyan, & Shelza Dua. (2012). Punjabi Automatic Speech Recognition Using HTK. International Journal of Computer Science Issues, 9(4), 359-364.
13. Panda, S. P., & Nayak, A. K. (2017). A waveform concatenation technique for text-to-speech synthesis. International Journal of Speech Technology, 20, 959-976.
14. Rabiner, L. (1989, February). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 257-286.
15. Reddy, D. (1966, September). An approach to Computer Speech Recognition by direct analysis of speech wave (Tech. Report No. C549). Computer Science Dept., Stanford University.
16. Redmon, J., & Farhadi, A. (2016). YOLO9000: Better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).
17. Sarada, G. L., Nagarajan, T., & Murthy, H. A. (2004). Multiple Frame Size And Multiple Frame Rate Feature Extraction For Speech Recognition. In International Conference on Signal Processing and Communications, SPCOM '04 (pp. 592-595).
18. Taylor, P., & Black, A. W. (1998). The architecture of the Festival speech synthesis system. In Proceedings of the 4th International Conference on Spoken Language Processing (Vol. 98, pp. 611-614).
19. Tokuda, K., Zen, H., & Black, A. W. (2002). An hmm-based speech synthesis system applied to English. In Sixth International Conference on Spoken Language Processing
20. Zen, H., Tokuda, K., & Black, A. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039–1064.

DOI: https://doi.org/10.15379/ijmst.v10i5.3690