Hindi Pronunciation Analysis for Speech Impaired using MFCC and DTW

Sahil Panchbhaiya
Masters in Computer Applications
Sardar Patel Institute of Technology
Mumbai, India

Pranav Menon
Masters in Computer Applications
Sardar Patel Institute of Technology
Mumbai, India

Rishikesh Lingayat
Masters in Computer Applications
Sardar Patel Institute of Technology
Mumbai, India

Nikhita Mangaonkar
Masters in Computer Applications
Sardar Patel Institute of Technology
Mumbai, India

ABSTRACT
The aim of this experiment is to educate speech-impaired learners on the pronunciation of Hindi syllables by providing word breakdowns, sounds, and examples of their usage. After the speaker becomes familiar with the syllables, a voice sample from the user is taken as input and analyzed to determine whether it matches the predefined data, ensuring that the speaker is following correctly. This feature matching is performed using Dynamic Time Warping (DTW) and Mel-Frequency Cepstral Coefficients (MFCC). The process is carried out using a combination of MFCC and DTW. In the two-step process of speech analysis, MFCC is used in the first phase to extract fourteen features, and the second phase employs three unique classifiers: k-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Dynamic Time Warping (DTW) to determine the best combination for accurate and precise feature matching.

Keywords
Mel Frequency Cepstral Frequency(MFCC); Dynamic Time Warping(DTW); k-Nearest Neighbour(KNN); Support Vector Machine(SVM); Speech Impairment; Hindi Syllables

1. INTRODUCTION
Pattern Recognition is currently the hottest thing in the market. Humans use voice as the most primary mode to communicate with each other and as a result a lot of data is readily available all around us which can be used for many different research purposes.[1]The speech recognition system requires different kinds of precise algorithms which includes algorithms that are used for the process of feature extraction and classification.[2] Speech recognition and analysis - a branch of pattern recognition that has become increasingly important in recent times in various fields including security, telecommunications and human-computer interaction. Efficient and accurate methods for speech matching and analysis are crucial for developing effective systems that match audio files to determine their similarity. This paper focuses on comparing MFCC and DTW, which are widely utilized in audio signal processing.[3] MFCC is being used to extract the fourteen features from the recording of the Speech Impaired learner which is further processed with the help of different classifiers such as Support Vector Machine(SVM), k-Nearest Neighbour(KNN) and Dynamic Time Warping(DTW) in order to analyze the recording to interpret the correct usage of the Hindi Syllables.[4] General Work flow of Speech Recognition Systems is given in Fig 1.

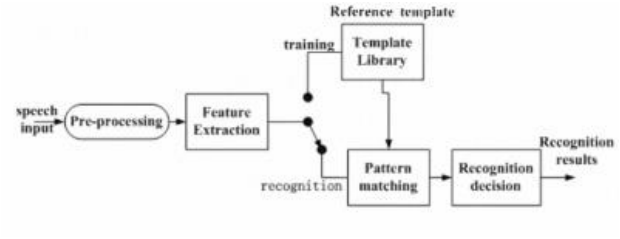

Fig 1: Flow of Speech Recognition

1.1 Mel-Frequency Cepstral Coefficients (MFCC)
MFCC is an extremely popular feature extraction technique in the process of speech signal processing. It involves transforming a short-term power spectrum of the sound signal into a specific set of coefficients that represent the unique spectral characteristics of the signal.[5] Then the process includes filtering the signal through a Mel-scale filter bank, taking on the logarithm of the list of filterbank energies, and applying the discrete cosine transformation to obtain the final coefficients.

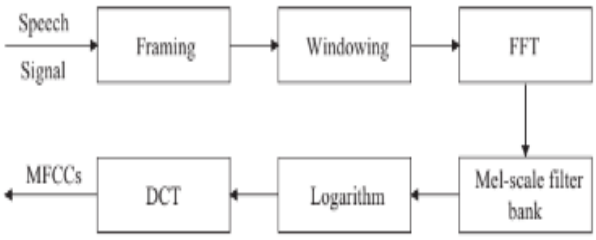

Fig 2: MFCC Feature Extraction Process

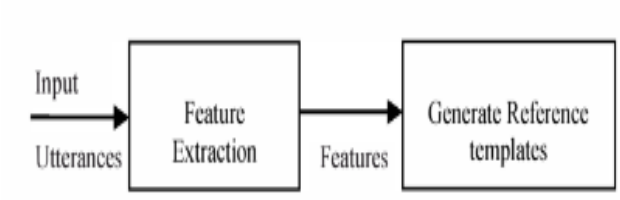

Fig 3: Feature Extraction Phase

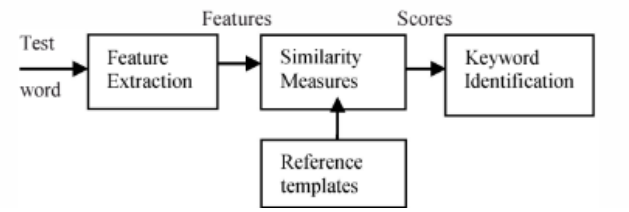

Fig 4: Feature Matching Phase

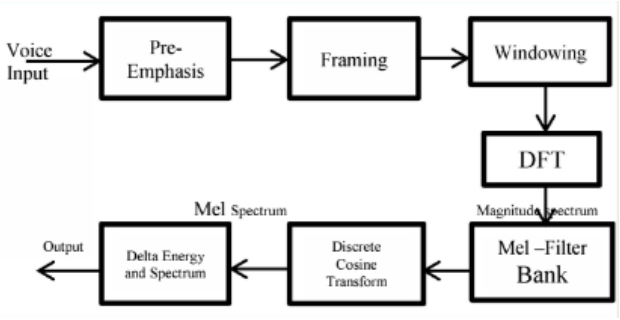

Fig 5: MFCC Block Diagram

1.2 Dynamic Time Warping (DTW)

DTW is a dynamically programmed algorithm used in aligning two time-series with different lengths. In the context of speech processing, DTW can be applied to compare two speech signals by finding the optimal alignment between their frames. DTW considers local temporal distortions, making it robust to variations in speech rate and duration.[6]

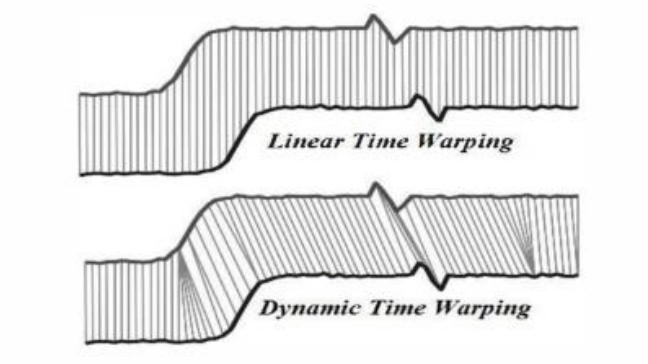

Fig 6: DTW Alignment of Two Series

Linear Time Warping (LTW) and Dynamic Time Warping (DTW) are the two methods for aligning time series. DTW accommodates non-linear distortions, making it versatile but computationally intensive.[7] LTW, assuming linear distortions, is simpler and computationally efficient, and one is chosen based on data characteristics and analysis requirements.

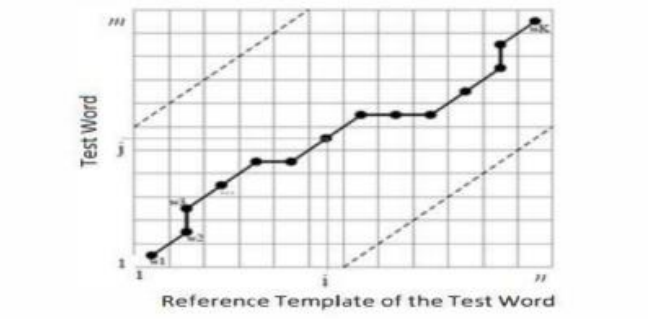

Fig 7: Warping between Two Time Series

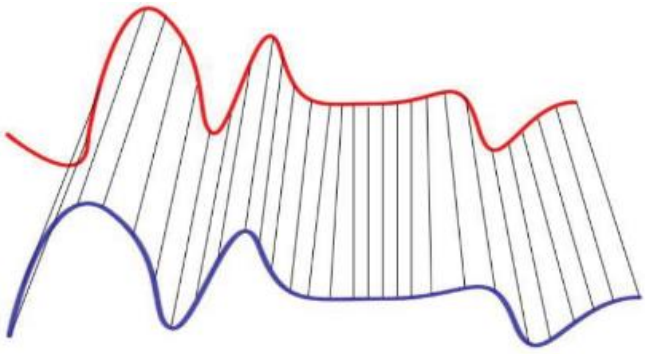

Fig 8: Dynamic Time Warping Approach

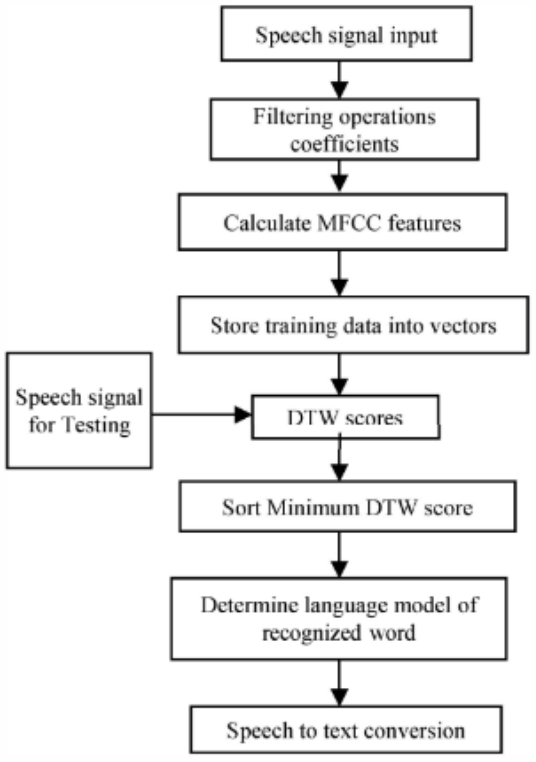

Fig 9: Flowchart for Isolated Word Recognizer using DTW

1.3 Support Vector Machines (SVM)

SVM is adept at phonetic matching by learning distinctive features associated with phonemes or pronunciation patterns.[8] It excels in speaker-dependent pronunciation matching, considering unique pronunciation characteristics of individual speakers. SVM is useful for accent-based matching, identifying and matching pronunciation patterns associated with specific accents or regional variations. Its strength lies in pattern recognition, distinguishing subtle differences in pronunciation and providing matching scores or labels.

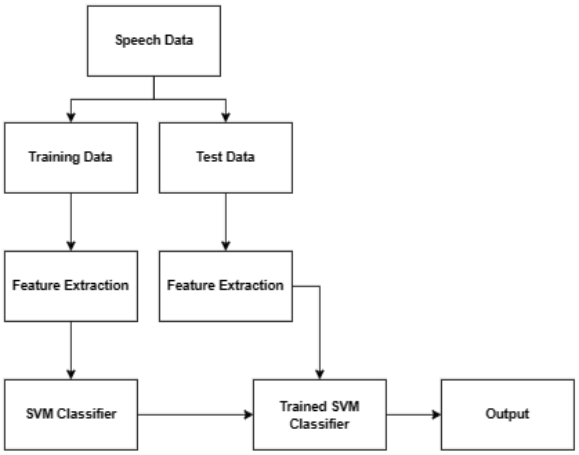

Fig 10: Flowchart of working of SVM

1.4 K-Nearest Neighbors (kNN)

KNN performs similarity-based matching, identifying similar pronunciation patterns by considering neighboring instances in the feature space. It is well-suited for dynamic pronunciation matching, adapting to variations in pronunciation over time. Local context matching is a feature of kNN, capturing subtle variations within specific contexts for more accurate matching. kNN exhibits noise robustness, making it suitable for pronunciation matching in scenarios with background noise or signal variations. Its non-parametric nature allows adaptation to variations in pronunciation without making strong assumptions about the underlying distribution.[9]

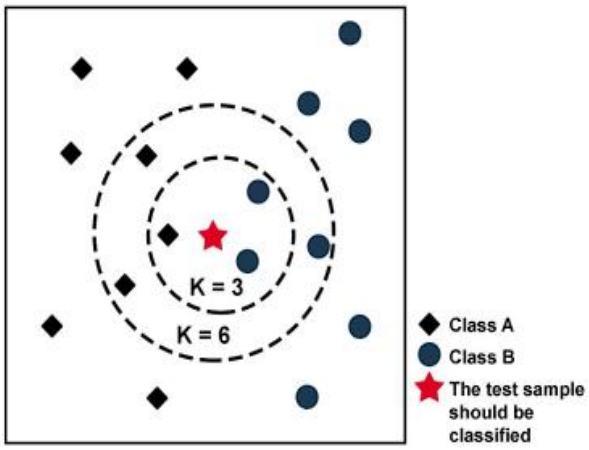

Fig 11: Representation of KNN Classifier

2. METHODOLOGY

2.1 Data Collection

As the first step in our comparative analysis we proceed to the acquisition of a diverse dataset of speech signals. The dataset should encompass a variety of speakers, speech rates, and environmental conditions to ensure that a complete and comprehensive evaluation of the MFCC and DTW methods can be obtained.

2.2 Preprocessing

Prior to feature extraction, the collected speech signals undergo preprocessing steps to enhance the quality of the data.[10] This may include noise reduction, normalization, and filtering to mitigate environmental variations.

2.3 Feature Extraction using MFCC

The MFCC feature extraction process involves the following steps:

- Frame Segmentation: The audio signal is divided in a span of short frames, typically around 20-30 milliseconds, with a certain overlap.
- Pre-emphasis: Applying a pre-emphasis filter to amplify high-frequency components.
- Framing: Dividing the signal into frames.
- Fast Fourier Transform (FFT): Computing the power spectrum of each frame.
- Mel-filterbank: Filtering out the power spectrum by using a set of Mel-scale filter banks.[11]

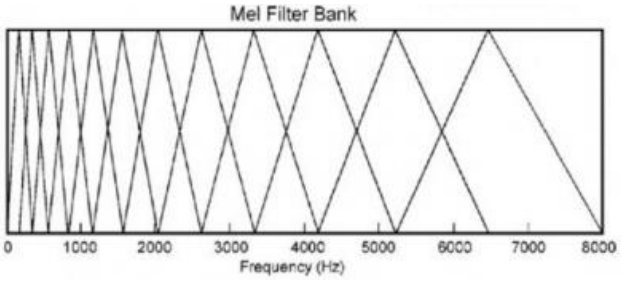

Fig 12: Mel Scale

- Log Transformation: Taking out a logarithm of the generated filterbank energies.
- Discrete Cosine Transform (DCT): Transforming the log filterbank energies into cepstral coefficients.
- The result is a set of MFCC coefficients representing the spectral characteristics of each frame in the speech signal.[12]
- Alignment using Dynamic Time Warping (DTW):
- The alignment process using DTW involves:
- Distance Matrix Calculation: Computing a distance matrix between the MFCC feature vectors of the reference and test speech signals.
- Dynamic Programming: Finding the optimal alignment path through the distance matrix using dynamic programming.
- Backtracking: Tracing back the optimal alignment path to obtain the aligned frames.[13]

2.4 Performance Evaluation:

The effectiveness of MFCC and DTW will be assessed using the following performance metrics:

- Accuracy: The overall correctness of the method in aligning and matching speech signals.

$$\text{Accuracy}=\frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

- Precision: It is the ratio of correctly aligned frames to the total number of frames identified by the method.

$$\text{Precision}=\frac{True\ Positives}{True\ Positives+False\ Positives}$$

- Recall: It's the ratio of correctly aligned frames to the total number of frames in reference signal.

Recall= $\frac{True\ Positives}{True\ Positives+False\ Negatives}$

- F1 Score: It is the harmonic mean of precision and recall, thereby providing an account of balanced measure of performance.

F1 Score= 2 x $\frac{Precision \times Recall}{Precision+Recall}$

2.5 Computational Efficiency Analysis:

To evaluate the computational efficiency of both methods, we will consider factors such as processing time and memory requirements.[14] This analysis will provide insights into the practical feasibility of implementing these methods in real-time or resource-constrained environments.

2.6 Experimental Setup:

All experiments will be conducted using a consistent experimental setup, including hardware specifications, software versions, and parameter configurations for both MFCC and DTW. Cross-validation techniques may be employed to ensure robustness and minimize biases in the results.[15]

2.7 Statistical Analysis:

Statistical tests, like the t-tests or ANOVA, may be deployed to determine the level of significance of the observed differences in performance metrics between MFCC and DTW.

2.8 Sensitivity Analysis:

To understand the robustness of each method, sensitivity analysis will be conducted by introducing variations in speech rates, noise levels, and speaker characteristics.

2.9 Ethical Considerations:

Ethical considerations regarding data privacy, consent, and potential biases in the dataset will be addressed in accordance with ethical guidelines for research involving human subjects.

This comprehensive methodology will allow for a thorough comparison of MFCC and DTW in the context of speech signal processing, providing valuable insights into their strengths and limitations.

3. ROLE OF ALGORITHMS

3.1 Role of Mel Frequency Cepstral Coefficients(MFCC)

3.1.1 Feature Extraction

MFCC is typically used for extracting features from speech signals that are robust and efficient for speech processing. It captures the spectral characteristics of audio signals, mimicking human auditory perception by representing the frequency content of the audio.

3.1.2 Steps Involved in MFCC

- Pre-emphasis: Balances the spectrum and increases the signal-to-noise ratio. It can be seen in equation (1).

$$Y[n] = X[n] - aX[n-1] \quad (1)$$

- Frame Blocking: Divides the signal into frames of short duration.
- Windowing: Multiplies each frame with a window function (e.g.Hamming window) to minimize spectral leakage. The Hamming window equation can be seen in equation (2).

$$w(n) = 0.54 - 0.46.\cos\left(\frac{2\pi n}{N-1}\right),\ 0 <= n <= N-1 \quad (2)$$

- Fast Fourier Transform (FFT): It is responsible for conversion of the signal from time domain to frequency domain.
- Mel Filterbank: Groups the spectrum into frequency bands based on the Mel scale, which is stated as the perceptual scale of pitches.

$$mel\_Scale\_Value = 2595\ log_{10}\ (1+ \frac{f}{700})$$

Here mel is output of mel filter bank, while J is the input signal of the result that is obtained from the FFT process. Values 2595 and 700 are the common set of values used on the mel filter bank scale and are widely used.

- Logarithmic Transformation: Takes the logarithm of the energy in every filter available.
- Discrete Cosine Transform (DCT): Extracts a set of coefficients representing the short-term power spectrum.[16]

3.1.3 Matching and Comparison:

After extracting MFCC features from two WAV files, a comparison method is employed to assess their similarity. While one conventional approach involves calculating the distance between the MFCC vectors using techniques like Dynamic Time Warping (DTW) or cosine similarity, we extend our analysis by incorporating additional methods such as k-Nearest Neighbors (kNN), Dynamic Time Warping(DTW) and Support Vector Machines (SVM).

These techniques provide diverse perspectives on the similarity between the speech samples. SVM and kNN offer classification-based comparisons, utilizing their respective algorithms to determine the degree of similarity. Meanwhile, DTW continues to serve as a temporal alignment measure. The obtained similarity measures from these methods can then be transformed into matching percentages or scores, offering a comprehensive assessment of how closely related the two speech samples are across different analytical dimensions.

3.1.4 Prediction:

Based on the matching percentage or similarity score obtained from comparing the MFCC features, a decision can be made regarding whether the two WAV files match.

A threshold value can be set to classify whether the files are a match or not. If the similarity score exceeds this threshold, the files can be considered a match.[17]

3.2 Role of Dynamic Time Warping(DTW)

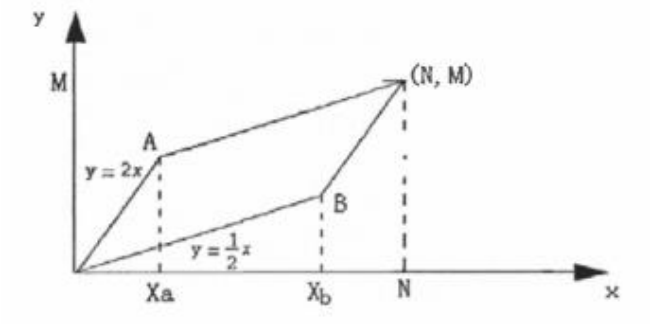

Fig 13: DTW Searching Path

- Sequence Alignment: DTW aligns sequences with varying lengths or time scales by warping the time axis. In speech recognition, different speakers might speak at different rates, causing variations in the duration of sounds. DTW helps align these sequences for

comparison.[18]

- Temporal Normalization: It normalizes sequences by stretching or compressing them in time to find the best alignment between them. This is crucial when comparing speech utterances that may have different durations due to speaking speeds or variations in pronunciation.[19]
- Similarity Measurement: DTW computes the distance or similarity between two sequences, considering possible nonlinear alignments. In speech recognition, it measures the similarity between spoken words or phrases, enabling comparison even when there are temporal distortions or variations in pronunciation.[20]
- Classification or Pattern Recognition: DTW can be used as a part of a classification algorithm to recognize patterns within time-series data. In speech recognition systems, after aligning and measuring the similarity between speech utterances using DTW, it can help classify or identify spoken words or phrases based on the closest matches.
- Feature Matching: In combination with feature extraction methods like MFCC, DTW can compare the extracted features of two speech signals. It allows for a more robust comparison, considering not just the extracted features but also their temporal alignment.
- Handling Noisy or Inconsistent Data: DTW is resilient to noise and variations within sequences. This resilience makes it effective for matching sequences that might have distortions, background noise, or inconsistencies.[21]

4. IMPLEMENTATION AND RESULTS

This section presents the outcomes of experiments conducted to evaluate the performance of three distinct models—Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Dynamic Time Warping (DTW)—for the task of audio pronunciation matching using Mel-Frequency Cepstral Coefficients (MFCC). The primary objective of the study is to investigate and compare the efficacy of these models in accurately assessing and matching audio pronunciations, with a focus on precision, recall, and overall accuracy.

Characteristic	**Value**
Number of Pronunciation Samples	50
Number of Speakers	8
Recording Conditions	Appropriate
Average Pronunciation Length	3 seconds
Language	Hindi

4.1 Model Performance Matrix:

4.1.1 Accuracy

Model	Accuracy
SVM	85 %
KNN	70.4 %
DTW	93.4%

4.1.2 Precision, Recall, and F1 Score

Model	Precision	Recall	F1 - Score
SVM	0.94	0.92	0.91
KNN	0.7	0.84	0.7636
DTW	0.96	0.94	0.95

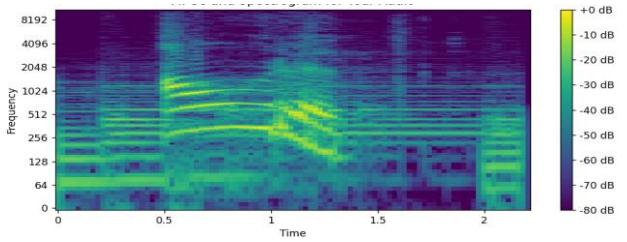

Fig 14: MFCC Spectrogram of Anaar Voice Sample

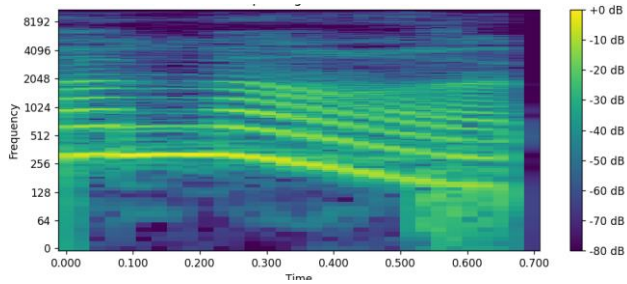

Fig 15: MFCC Spectrogram of Okhli Voice Sample

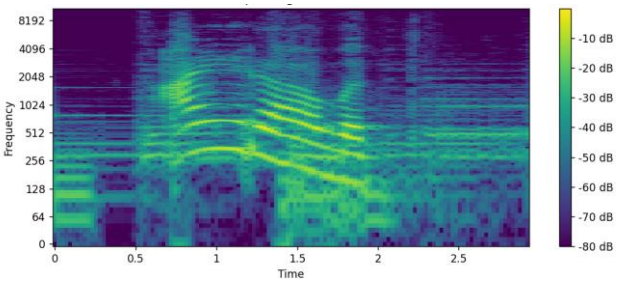

Fig 16: MFCC Spectrogram of Ooth Voice Sample

In this experiment, a comprehensive examination of three distinct models—k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Dynamic Time Warping (DTW)—was conducted,

focusing on their application to the task of audio pronunciation matching using Mel-Frequency Cepstral Coefficients (MFCC). The investigation involved evaluating the models based on key performance metrics, including accuracy, precision, recall, and the dynamic nature of DTW.

1. KNN Performance

KNN exhibited the lowest accuracy among the three models, suggesting challenges in its ability to effectively capture the complex patterns that are created within the pronunciation dataset. The simplicity of KNN might have limitations in handling the intricacies of audio pronunciation variations.

2. SVM Performance:

SVM demonstrated a relatively better performance compared to KNN, indicating its effectiveness in the high-dimensional space of MFCC features.SVM demonstrated high precision (0.94) and recall (0.92), resulting in an F1-score of 0.91. These metrics indicate SVM's ability to accurately classify positive samples while minimizing false positives. SVM's ability to find optimal decision boundaries allowed it to outperform KNN, showcasing its strength in handling classification tasks.

3. DTW Performance:

DTW emerged as the top-performing model, demonstrating superior accuracy and efficiency in capturing dynamic temporal relationships within pronunciation sequences. DTW's adaptability to variable-length sequences and robustness in aligning temporal distortions played a crucial role in its success.

5. CONCLUSION

This study demonstrated the effectiveness of using Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) for Hindi pronunciation analysis in speech-impaired individuals. The experimental results indicated that DTW outperformed Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) in terms of accuracy, precision, and recall, making it a robust choice for pronunciation matching tasks.

The findings of this research have significant implications for the field of speech recognition, particularly in improving pronunciation assessment for speech-impaired individuals. The methodology proposed in this study can be integrated into speech therapy tools, enhancing the accuracy and effectiveness of pronunciation training.

Future research can explore the integration of advanced deep learning techniques to further enhance model accuracy and robustness. Additionally, extending the research to include other languages and dialects would help generalize the methodology. Investigating the impact of different environmental conditions, such as background noise and varied recording qualities, can provide further insights into the practical applicability of the models.

Practical applications of this research include the development of educational software for language learning and the creation of assistive technologies to improve communication for speech-impaired individuals. By continuing to refine and expand upon this methodology, it is possible to develop more sophisticated speech analysis tools that can significantly benefit speech-impaired individuals and contribute to advancements in speech recognition technology.

In conclusion, this research lays a solid foundation for future work in the area of pronunciation analysis and highlights the potential for innovative applications that can enhance the quality of life for speech-impaired individuals.

6. REFERENCES

[1] A. Winursito, R. Hidayat, A. Bejo and M. N. Y. Utomo, "Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, Malaysia, 2018, pp. 1-6, doi: 10.1109/ICSCEE.2018.8538414.

[2] R. Hidayat, A. Bejo, S. Sumaryono and A. Winursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System," 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), Bali, Indonesia, 2018, pp. 280-284, doi: 10.1109/ICITEED.2018.8534807.

[3] A. Brahme and U. Bhadade, "Marathi digit recognition using lip geometric shape features and dynamic time warping," TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, Malaysia, 2017, pp. 974-979, doi: 10.1109/TENCON.2017.8227999.

[4] R. Koul, M. Yadav and K. Suneja, "Comparative Analysis of FPGA Based Hardware Design of Dynamic Time Warping Algorithm using Different Multiplier Architectures," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020, pp. 599-603, doi: 10.1109/GUCON48875.2020.9231244.

[5] P. Yang, L. Xie, Q. Luan and W. Feng, "A tighter lower bound estimate for dynamic time warping," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 8525-8529, doi: 10.1109/ICASSP.2013.6639329.

[6] J. Joseph and S. S. Upadhya, "Indian accent detection using dynamic time warping," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 2017, pp. 2814-2817, doi: 10.1109/ICPCSI.2017.8392233.

[7] J. C. Vasquez-Correa, J. R. Orozco-Arroyave and E. Nöth, "Word accuracy and dynamic time warping to assess intelligibility deficits in patients with Parkinson's disease," 2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA), Bucaramanga, Colombia, 2016, pp. 1-5, doi: 10.1109/STSIVA.2016.7743349.

[8] X. Zhang, J. Sun, Z. Luo and M. Li, "Confidence index dynamic time warping for language-independent embedded speech recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 8066-8070, doi: 10.1109/ICASSP.2013.6639236.

[9] K. Sheoran et al., "Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping," in IEEE Access, vol. 11, pp. 15485-15495, 2023, doi: 10.1109/ACCESS.2023.3244393.

[10] S. Singhal and R. K. Dubey, "Automatic speech recognition for connected words using DTW/HMM for English/ Hindi languages," 2015 Communication, Control and Intelligent Systems (CCIS), Mathura, India, 2015, pp. 199-203, doi: 10.1109/CCIntelS.2015.7437908.

[11] S. Paul, B. P. Babu and L. Mary, "Assessment of Articulation Disorder Using Objective Quality Measures," 2018 International Conference on Control, Power,

Communication and Computing Technologies (ICCPCCT), Kannur, India, 2018, pp. 439-444, doi: 10.1109/ICCPCCT.2018.8574289.

[12] Zhang Jing and Zhang Min, "Speech recognition system based improved DTW algorithm," 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, Changchun, 2010, pp. 320-323, doi: 10.1109/CMCE.2010.5609979.

[13] T. S. Kumar, T. Sheela, D. Arulselvam, S. Premalatha and K. Srividya, "Study of Various Machine Learning Algorithms for use with Automatic Speech Recognition," 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICPECTS56089.2022.10047695.

[14] P. Mahesha and D. S. Vinod, "LP-Hillbert transform based MFCC for effective discrimination of stuttering dysfluencies," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017, pp. 2561-2565, doi: 10.1109/WiSPNET.2017.8300225.

[15] M. Goyani, N. Dave and N. M. Patel, "Performance Analysis of Lip Synchronization Using LPC, MFCC and PLP Speech Parameters," 2010 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 2010, pp. 582-587, doi: 10.1109/CICN.2010.115.

[16] Q. Li et al., "MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method With Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications," in IEEE Access, vol. 8, pp. 48720-48730, 2020, doi: 10.1109/ACCESS.2020.2979799.

[17] Senthildevi K. A and Chandra E, "Keyword spotting system for Tamil isolated words using Multidimensional MFCC and DTW algorithm," 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, India, 2015, pp. 0550-0554, doi: 10.1109/ICCSP.2015.7322545.

[18] S. Gaikwad, B. Gawali, P. Yannawar and S. Mehrotra, "Feature extraction using fusion MFCC for continuous marathi speech recognition," 2011 Annual IEEE India Conference, Hyderabad, India, 2011, pp. 1-5, doi: 10.1109/INDCON.2011.6139372.

[19] M. V. Unnikrishnan and R. Rajan, "Mimicking voice recognition using MFCC-GMM framework," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 301-304, doi: 10.1109/ICOEI.2017.8300936.

[20] K. Sukvichai, C. Utintu and W. Muknumporn, "Automatic Speech Recognition for Thai Sentence based on MFCC and CNNs," 2021 Second International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 2021, pp. 1-4, doi: 10.1109/ICA-SYMP50206.2021.9358451.

[21] Mizanur Rahman and Md. Babul Islam, "Performance evaluation of MLPC and MFCC for HMM based noisy speech recognition," 2010 13th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2010, pp. 273-276, doi: 10.1109/ICCITECHN.2010.572386