Sentence Recognition using KNN method

Priyanka Wani[1] and Mukesh Ghogare[2]

[1]Assistant Professor, Department of Electronics and Tele- Communication, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune

[2] Department of Instrumentation and Control Engineering, COEP Technological University, Pune,

Abstract: Speech is the most common expressed form of human speech. Talking is simple, hands-free, and quick, and requires no technological expertise. Speech recognition software makes it easy and comfortable for people to communicate with computers. A speech recognition system enabled this. The accessible linguistic and acoustic models for this system are primarily in English. There are a lot of people in India that are illiterate in English. For these persons, the English voice recognition technology is therefore useless. Sentence recognition systems' primary objective is to comprehend voice input from a device and translate it into text so that necessary tasks can be completed.

Keywords: VQ, recognition

1. Introduction

Text is created by automatically converting a spoken word sequence into an audio file. Speaking to a computer via speech is a more straightforward and pleasant method of communication for humans than using a keyboard and mouse, which need more dexterity and hand-eye coordination [1]. People with physical disabilities or those who are blind find using computers challenging. All of these problems are resolved by speech recognition.

A voice recognition system has two primary modes: training mode and testing mode for recognition. During training mode, it listens to all of the speaker's utterances and uses different technique for identifying characteristics [2,3,4] to determine the feature vectors that correlate to each utterance. In this manner the voice signal of a training vector is formed.

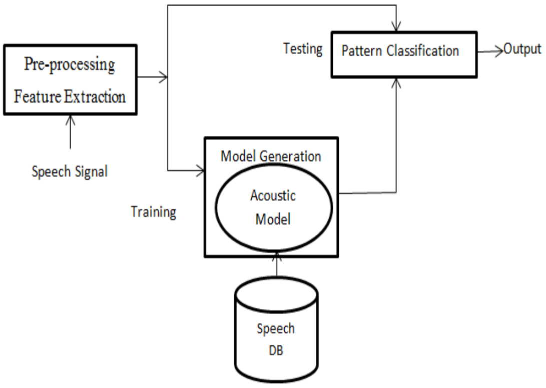


Fig. 1: Speech Recognition Model

2. History

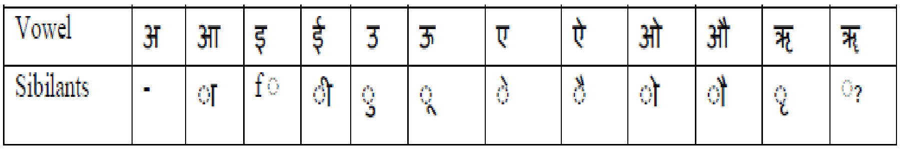
In Hindi, Tarun Pruti[1] created a speaker-dependent isolated word identification system. HMM is employed at the back end for recognition, and front end for feature collecting technique. Two male speakers were intended for the system. Hindi digits (0, pronounced as "shoonya") through (9, pronounced as "nau") make up the recognition vocabulary.

Using a limited vocabulary, Kuldeep Kumar [5,6] created an isolated Hindi speech recognition system that performed well, 94.63% of the time. There is no speaker dependency with this system. Three females and five males are employed for training. The system has a vocabulary of thirty terms. HMM is employed for recognition at the back end, and a informative data collecting technique at the front end. 2011 saw the proposal of the Hindi ASR system for connected digit.

3. **Vowel and Consonant**

The Sanskrit writing system, Devanagari, was the most frequently used language in our country.With 258–422 million speakers, Hindi is regarded as the national language of India. In contrast to other languages, vast progress is ging on into putting Hindi speech recognition techniques into practice. There are two categories in Hindi language: vowels and consonants. Pronouncing vowels opens up the vocal tract considerably. The twelve vowels in Hindi are referred to as Barakhadi. Chart 1 displays Hindi Vowels

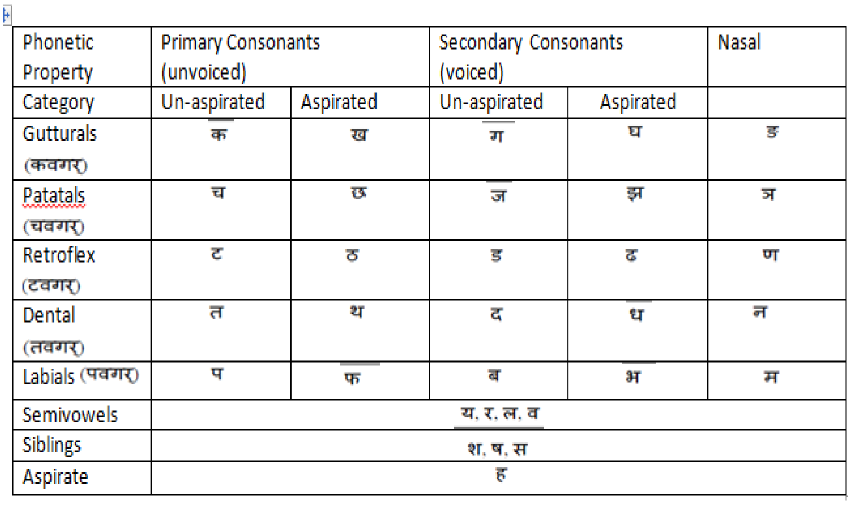
Chart- I



Vowel	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	ऋ	ऋ
Sibilants	·	◌ा	f◌	◌ी	◌ु	◌ू	◌े	◌ै	◌ो	◌ौ	◌ृ	◌?

The vowel, 'ॐ' also known as a grapheme, is included in this set. It corresponds to a phoneme sound that has two or more concatenated phonemes. Vowel matras are used to represent this vowel sound different than the implied one. Because of the restriction in the vocal tract shape during consonant pronunciation, consonants are categorized based on the location and manner of articulation (POA and MOA). It has five vowels and nine non-vowels [7-10]. There are five consonants in each varg; the nasal consonant is the last one. Every variant consists of a primary and secondary pair made up of the first four consonants. While secondary consonants are voiced, primary consonants are unvoiced. each other's secondary consonants" also known as a grapheme, is included in this set. It corresponds to a phoneme sound that has two or more concatenated phonemes. Vowel matras, commonly known as vowel signs, are used to represent this vowel sound different than the implied one. The consonant set is displayed in Chart 2.

Chart- II



Phonetic Property	Primary Consonants (unvoiced)		Secondary Consonants (voiced)		Nasal
Category	Un-aspirated	Aspirated	Un-aspirated	Aspirated	
Gutturals (कवगर)	क	ख	ग	घ	ङ
Patatals (चवगर)	च	छ	ज	झ	ञ
Retroflex (टवगर)	ट	ठ	ड	ढ	ण
Dental (तवगर)	त	थ	द	ध	न
Labials (पवगर)	प	फ	ब	भ	म
Semivowels	य. र. ल. व				
Siblings	श. ष. स				
Aspirate	ह				

4. Technique used

An overview of a Word Recognition (WR) system with three main steps: pre-processing, extracting MFCC feature vectors, and matching these vectors against a database using a pattern recognition algorithm. This is a common approach in speech processing systems. Let's break down each step:

The block diagram (Fig. 2) likely illustrates the flow of information between the pre-processing, feature extraction, and pattern recognition components.

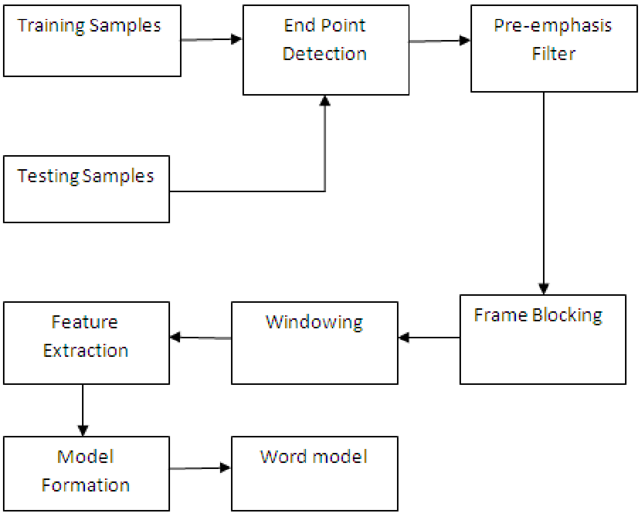


Fig.2: Block Diagram of WR

i. Pre-processing:

This step involves cleaning and enhancing the raw speech signal obtained from the microphone. Common pre-processing techniques include noise reduction, filtering, and normalization to improve the quality of the signal. The goal is to prepare the signal for further analysis and feature extraction.

ii. Feature Extraction:

The speech signal is divided into short frames, and for each frame, the MFCCs are computed to represent the spectral characteristics. These coefficients capture essential features of the speech signal, especially focusing on the frequency components that are most relevant to human hearing.

iii. Pattern Recognition:

The extracted coefficient from the speech signal are compared or matched against a database of known patterns. The aim of pattern recognition is to identify the sentence spoken based on the similarity between the extracted features and the stored patterns in the database.

iv. Word Recognition (WR) System:

The overall system involves the integration of these three steps into a coherent process for recognizing spoken words.

v. Procedure to obtain coefficient:

a. Sampling:

The input speech waveform is considered quasi-stationary, so it is divided into short frames (typically 20-30 ms) where the characteristics are assumed to be stationary. This simplifies the analysis, allowing for the consideration of the signal over short durations.

b. Windowing:

Each sample is multiplied by a window function to avoid aliasing effects and minimize discontinuities. Common window options include Rectangular, Hanning, and Hamming. Every frame has the Hamming window function applied to it in order to maintain regularity till endpoints to avoid sudden changes. Hamming window has highest side lobe attenuation and larger transition width of 8π/M where M is filter order. Hamming window is used to avoid Gibbs phenomenon.

The Hamming window is defined as-

$$w(n) = \begin{cases} 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right); & 0 \leq n \leq N-1 \\ 0\ ;\ otherwise \end{cases}$$

For the purpose to maintain continuity between the beginning and ending points of each frame and avoid sudden changes at the end, each framework is multiplied by the hamming window [10].

c. Fourier Transformation:

After windowing, a Discrete Fourier Transform (DFT) is used to each sample to convert the signal domain.

$$\boldsymbol{X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi kn}{N}}} \quad , \qquad \boldsymbol{k = 0, 1, 2, \ldots\ldots, N-1}$$

Since the FFT algorithm is the fastest for calculating DFT, it is occasionally used to faster the counting rate of DFT by a factor of 100 [9]. The vocal tract parameters are represented by a slowly varying envelope, while the fundamental frequency is represented by rapid variations in the FFT log.

d. Logarithmic Spectrum:

The logarithmic spectrum is calculated to show the amplitude of spectrum lines in decibels (dB). This representation captures faster changes corresponding to the normal frequency and slowly changing envelope corresponding to vocal tract parameters.

e. Method of Warping:

Based on how frequencies are perceived by human hearing, the spectrum is scaled using the Mel method. The Mel scale is approximately linear below 1000 Hz and logarithmic above. It is defined by setting 1000 mels equal to 1000 Hz as a reference point.

Formula to calculate Mel Scale is-

$$\boldsymbol{Mel(f) = 2595 * ln\left(1 + \frac{f}{700}\right)}$$

Here Mel (f) denotes perceived frequency and f is original frequency [9].

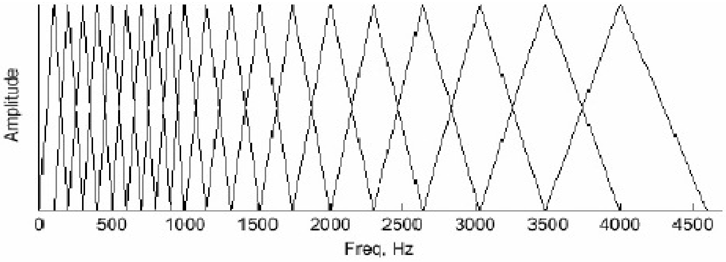


Fig.6: Filter bank of Mel Scale

Mel frequency warping is used to represent the spectrum in a way that is more consistent with human perception. The overall goal of these steps is to transform the input audio into suitable information for capturing the distinctive features of speech, which can then be used for further processing, such as word recognition. The use of MFCCs is a common technique in speech signal processing due to their effectiveness in capturing relevant information for speech analysis

f. Discrete Cosine Transform (DCT):

After obtaining the filtered outputs from the Mel filterbank and evaluating them with a logarithmic process, the next step involves applying the Discrete Cosine Transform (DCT). DCT is applied to the log Mel spectrum to transform it into a set of coefficients, which are the MFCC coefficients. The DCT helps decorrelate the features and represent them in a more compact form.

g. Calculation of Delta and Delta-Delta Coefficients:

The obtained MFCC coefficients capture static information about the speech signal. To incorporate dynamic information and add a temporal dimension, delta and delta-delta coefficients are calculated. The first-order derivative of the cepstral coefficients is known as the delta coefficient, and the second-order derivative is called the delta-delta coefficient. Delta coefficients provide information about the rate of change or speech speed, while delta-delta coefficients provide information similar to acceleration in speech. Calculating these derivatives adds time evolution information to the MFCCs, making them more effective for capturing speech variations over time.

h. MFCC Features:

The final set of features used for word recognition consists of the original MFCC coefficients, along with their delta and delta-delta coefficients. The extracted MFCC features are commonly used as input for pattern recognition algorithms in speech processing systems.

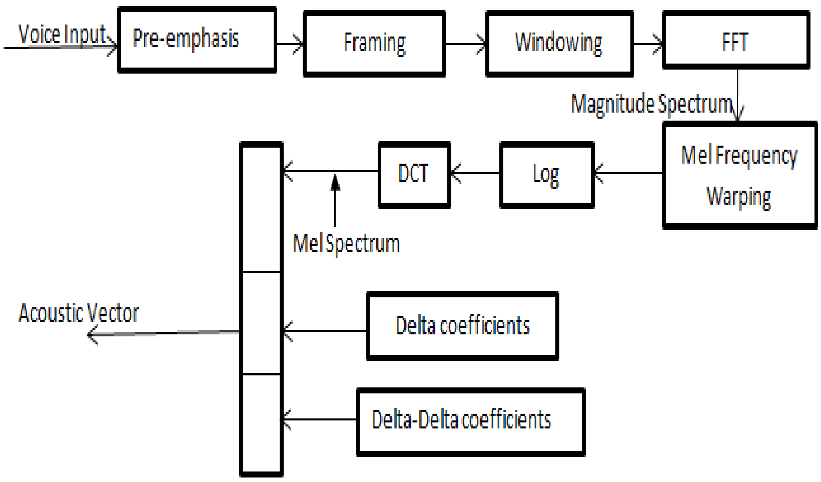


Fig.3: Model of MFCC

The combination of MFCC coefficients with their temporal derivatives enhances the ability of the system to capture the dynamic aspects of speech, making it more robust for recognizing spoken words, especially in varying acoustic conditions and speech styles. The extracted MFCC features, along with their delta and delta-delta coefficients, provide a comprehensive representation of the speech signal for subsequent pattern matching in the word recognition system. Fig.7 shows Obtained MFCC elements [11].

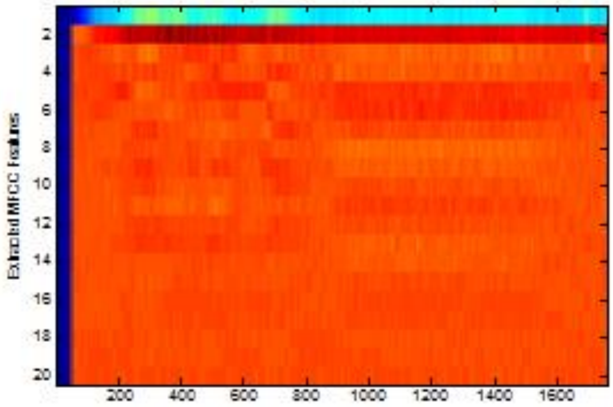


Fig.7: Obtained MFCC elements

i. Categorization:

Categorization is the final step in the word recognition process, where the extracted features (such as MFCC coefficients and their derivatives) are used for pattern matching. Various classification algorithms can be employed, and in this case, the K-Nearest Neighbor (KNN) classifier is chosen.

j. K-Nearest Neighbor (KNN) Algorithm:

KNN is user friendly and easy tool for classification and regression tasks. In the context of word recognition, the algorithm works by comparing the feature vectors of an input speech signal with those in the training dataset. The "k" in KNN represents the number of nearest neighbors that are considered during the classification process. The classification decision is based on the majority class among the k-nearest neighbors. The choice of the appropriate value for "k" can influence the performance of the KNN classifier.

k. Flow Chart of KNN Algorithm:

The KNN flow chart likely illustrates the sequential steps involved in the classification process. Fig.8: KNN Flow Chart. This may include procedures such as calculating distances between feature vectors, selecting the k-nearest neighbors, and making the final classification decision based on the majority class.The testing phase requires the use of all training data. K is a defined by users constant that specifies how many neighbors have an impact on the categorization during the process of classification.

5. Dataset

i. Dataset Development:

The Hindi dataset is created for the purpose of training and testing a speech recognition system. Hindi sentences are recorded for different speakers. The recordings are conducted in a room environment to capture realistic acoustic conditions. The speech files are stored in the .wave file format.

ii. Recording Details:

Each word is recorded as an isolated sample, and the recording time for each sample is limited to two seconds. The choice of a two-second recording time aims to avoid unnecessary silence in the recorded speech samples.

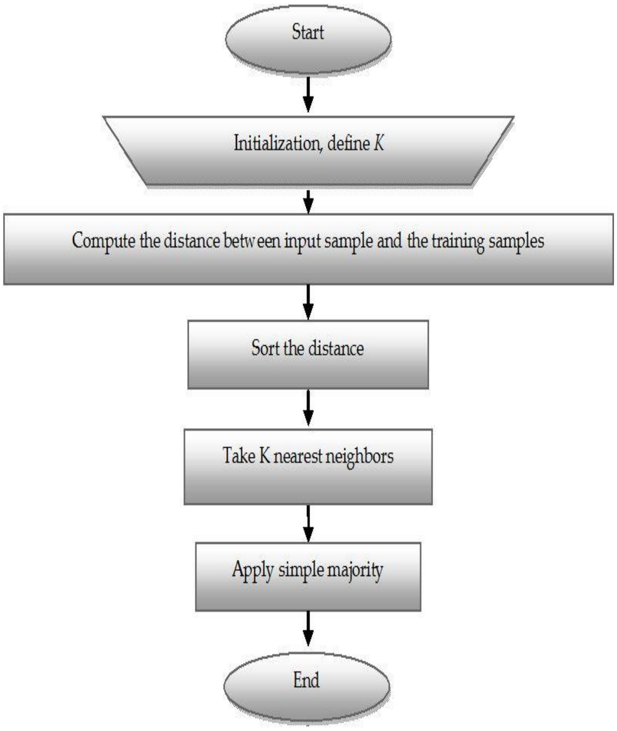


Fig.8: KNN Algorithm

This duration is deemed sufficient for capturing the pronunciation and characteristics of isolated words.

iii. Recorded Words:

The database includes recordings of Hindi sentence, such as -

- मगर कमिनिस्टों का विश्वास ध्वंस में है चीन ने अपने हाल के प्रयोगों से ये साबित कर दिया है.
- चीन नहीं चाहता कि अमरिका उसकी गुप्त सैनिक जानकारी हासिल करे.
- अतः अगर डरना है तो अमरीका डरे.

iv. Usage for Training and Testing:

The developed database serves the dual purpose of training the speech recognition system and evaluating its performance through testing. Training involves using a subset of the database to teach the system to recognize the spoken sentence effectively. The remaining samples are reserved for testing, assessing how well the system generalizes to new, unseen data.

6. Result

This study implements Hindi sentence recognition system. Here, we matched patterns using KNN and extracted features using MFCC. Wave files are used to extract MFCC features. The KNN classifier is used to classify the features of training and testing samples. Three classes are formed because there are three sentences. The KNN classifier generates the corresponding wave files based on the classes. A graphical representation of each sentence is shown in Fig. 10. The performance of the voice recognition system is enhanced as a result of making it more reliable and effective.

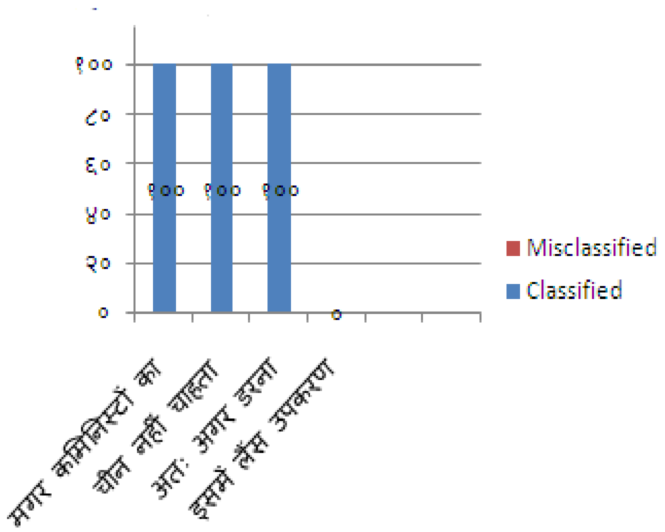


Fig.10: A visual illustration of 3 sentences

7. Conclusion

The study has been focusing on speech recognition of individual words in the Hindi language. The input of the speech was captured at 12 KHz and processed using a 25 ms Hamming window at a frame rate of 10 ms in order to acquire the feature vectors. Here, the two front-end feature extraction methods used were improved MFCC and traditional MFCC. At the back end, a KNN classifier was used. For the experiment, a set of speech data—Hindi sentences recorded by several speakers—was employed. Since Hindi cannot be searched using databases made for other languages, we assembled our own corpus of articles from popular Hindi news outlets.

References

[1] Pruthi, Tarun, Sameer Saksena, and Pradip K. Das, Swaranjali: Isolated word recognition for Hindi language using VQ and HMM." In *International conference on multimedia processing and systems (ICMPS)*, pp. 13-15. 2000.

[2] Kumar, Kuldeep, R. K. Aggarwal, and Ankita Jain, A Hindi speech recognition system for connected words using HTK." *International Journal of Computational Systems Engineering*, no. 1 (2012): 25-32.

[3] Mishra, A. N., Mahesh Chandra, Astik Biswas, and S. N. Sharan, Robust features for connected Hindi digits recognition." *International Journal of Signal Processing, Image Processing and Pattern Recognition*, no. 2 (2011): 79-90.

[4] Sinha, Shweta, Shyam S. Agrawal, and Aruna Jain, Continuous density Hidden Markov Model for context dependent Hindi speech recognition, *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1953-1958. IEEE, 2013.

[5] Aggarwal, Rajesh Kumar, and M. Dave, Using Gaussian mixtures for Hindi speech recognition system, *International Journal of Signal Processing, Image Processing and Pattern Recognition* 4, no. 4 (2011): 157-170.

[6] Saksamudre, S., and R. Deshmukh, Isolated word recognition system for Hindi Language, *International Journal of Computer Sciences and Engineering*, no. 7 (2015): 110-114.

[7] Thakur, Abhishek, and Naveen Kumar, Automatic Speech Recognition System for Hindi Utterance with Regional Indian Accents: A Review, *International Journal of Electronics & Communication Technology*, (2013).

[8] Paithane, A. N., and D. S. Bormane, Analysis of nonlinear and non-stationary signal to extract the features using Hilbert Huang transform, *IEEE International conference on computational intelligence and computing research*, pp. 1-4. IEEE, 2014.

[9] Paithane, A. N., and D. S. Bormane, Electrocardiogram signal analysis using empirical mode decomposition and Hilbert spectrum, *International Conference on Pervasive Computing (ICPC)*, pp. 1-4. IEEE, 2015.

[10] Dinde, Sneha, and A. N. Paithane, Human Emotion Recognition using Electrocardiogram Signals, *International Journal on Recent and Innovation Trends in Computing and Communication* 2, no. 2 (2004): 194-197.

[11] Wani, Priyanka, U. G. Patil, D. S. Bormane, and S. D. Shirbahadurkar, Automatic speech recognition of isolated words in Hindi language, *International Conference on Computing Communication Control and automation (ICCUBEA)*, pp. 1-6. IEEE, 2016.