Turkish Journal of Computer and Mathematics Education Vol.12 No. 9 (2021), 3196-3206

Research Article

Comparison Of Different Feature Extraction Techniques In Telugu Dialects Identification

S. Shivaprasad[1*], M. Sadanandam[2]

[1]Research Scholar, Department of CSE, Kakatiya University, Warangal & Assistant Professor, School of CS&AI, SR University, Hasanparthy, Warangal.

[2]Assistant Professor, Department of CSE, KU College of Engineering and Technology (Kakatiya University), Warangal.

shiva.prasad923@gmail.com[1*], sadanb4u@yahoo.co.in[2]

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

ABSRACT: The Telugu language is one of the standard and historical languages in India. Like other spoken languages, the Telugu language also contains different dialects i.e., Telangana, Costa Andhra and Rayalaseema. The research in dialects identification (DI) work is very less as compared to other speech processing applications. For the research Dialect identification (DI) in Telugu language, database is very important, but there is no standard database. To carry out the research work in DI, we created the database of Telugu Language with different dialects and we anlyzed different feature extraction techniques to implement system. In this paper, we compare the performance of different models given by applying the different feature extraction techniques such as spatial, temporal, and prosodic features in Telugu dialect identification. We have applied different classification models i.e., K-Nearest Neighbour (K-NN), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). It is observed that GMM model provides good accuracy with MFCC+ΔMFCC+ΔΔMFCC and MFCC+PITCH +LOUDNESS features compare to HMM model.

Keywords: Dialects, GMM, HMM, Telugu language, spatial features, Temporal Features, Prosodic features

1. Introduction

As per 2011 census, Telugu Language contains 82 million speakers and ranked as 4th with the height native speakers in India [1] and it is 15th in the list of languages in the Ethnologue by number of native speakers [2][3]. It is the most commonly spoken language of the Dravidian language family [4] and one of the twenty-two scheduled languages of the Republic of India. It is also the fastest-growing language in the United States, where there is a significant Telugu-speaking population [5]. Around 10,000 pre-colonial inscriptions exist in the Telugu language.

Dialects is referred as variation of a standard language that are spoken by set of people related to their geographical region. Like, other standard language, Telugu language also have the different variation in pronunciation. Mainly it contains three different dialects namely Telangana, Costa Andhra and Rayalaseema. These three dialects has different variations in pronunciations and contains different behaviour with respect to their prosodic features like rhythm, stress, intonation etc.

Telangana is a state in India situated on the high Deccan Plateau on the Indian peninsula's south-central stretch.[10] According to the 2011 census, it is the 11th largest state and the twelfth most populated state in India, with a geographical area of 112,077 km2 (43,273 sq mi) and 35,193,978 citizens [11]. The region was partitioned from Andhra Pradesh as the newly formed part of Andhra Pradesh on 2 June 2014. It currently contains 33 districts. It is primarily spoken in Warangal, Nizamabad, Khammam and Karimnagar,

Coastal Andhra is situated on the Coromandel Coast in the eastern portion of the state of Andhra Pradesh and consists of nine districts. Rayalaseema slang is spoken in four districts Anantapur, Chittoor, Kadapa, and Kurnool.

Research work in Dialect identification is very less compare to other speech processing applications including language identification even though it is equal to importance to LID. The main reason is, dialects of Telugu do not contain the standard database. Automatic speech recognition also provides less accuracy in dialect identification because there are a lot of variations in a language.

In this paper, we have designed dialect recognition system with different features and different kind of standard models like KNN, HMM and GMM. We considered Telugu Language for identify the dialects from Telugu speech utterances of human being of 3-8 sec. The performance of DI system designed from different models with variats of feature vectors.

The remaining paper is organized into four sections. Section2 provides different feature vectors used for dialects identification and Section 3 describes varies statistical models used for classification, Section 4 presents results and comparison and conclusions are drawn in
Section 5.

2. DATABASE CREATION

As there is no standard database for dialects of Telugu language, we collected different speech samples from different regions. The speakers used for dataset creation belongs to different ages from 19-50. We collected total 7h 05min speech samples from Telangana, Andhra and Rayalaseema regions. Some of speeches also collected from online and edited by using streaming audio recorder. To collect the speech, we have chosen different places like office, colleges, park and roadside. Even if small words occurred, also it will not have affected the working of the model. We used PRAAT tool to record the audio from the speakers. The Human beings who participated in the voice record, spoken their own topic with their regional dialects. After collecting the speech, we applied the preprocessing techniques to remove the noise from speech signal. After that, we stored dataset according to the dialects. The figure2 shows the basic structure to create the database from the speech samples of speakers.

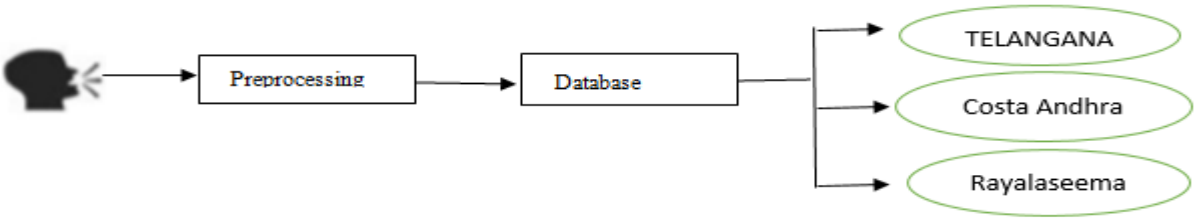

Fig.1. Block diagram of database creation

The complete dataset created for identification purpose as shown in Table.1.

Table.1 Complete dataset created.

S.NO	Dialect	Total time of speech data	Speakers for each dialect	Age of speakers	Sampling Frequency
1	Telangana	2h 35 min	22	20-55	44,100Hz
2	Coastal Andhra	2h 47 min	19	20-55	44,100Hz
3	Rayalaseema	1h 43 min	18	20-55	44,100Hz

3. Feature Extraction Techniques:

3.1 Mel frequency cepstral coefficient (MFCC)

Mel frequency cepstral coefficient is a fundamental feature extraction technique in any speech application and speech processing technique. In this, the cepstrum provides the basic information regarding the coefficients in the speech signal used to identify the human being. The cepstrum can be represented in the below equation (1)

$$C(s(t)) = F^{-1}[\log(F[s(t)])] \quad (1)$$

Where s(t) is the represent the speech signal and F is DFT applied on speech signal.

Applied F on S(t) in order to find out the spectrum of the speech signal. Logarithm is used to find out the log amplitude spectrum for which discriminate the vocal track envelope and glottal pulse. Apply the inverse Fourier transformation to find out the cepstrum of speech signal C(s(t)). The following flow chart [2] describes the steps to find out cepstrum of signal.

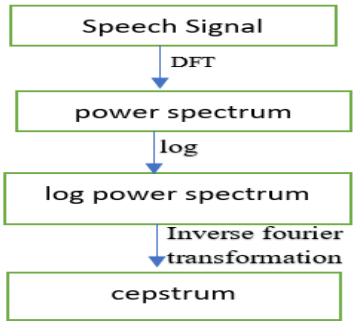

Fig.2.Basic block diagram to find out cepstrum of speech signal

Speech signal can be represented as follows

$$S(t) = E(t).H(t) \quad (2)$$

Where H(t) is vocal track frequency and E(T) is glottal pulse. To separate these two from signal apply the logarithm on both sides by applying the low pass filters.

$$\text{Log}(S(t)) = \log(E(t).H(t)) \Rightarrow \text{Log}(s(t)) = \log(E(t)) + \log(H(t)) \quad (3)$$

Then we eliminate the E(T) and considered H(t) for any application. The basic block diagram of MFCC is represented in figure [3].

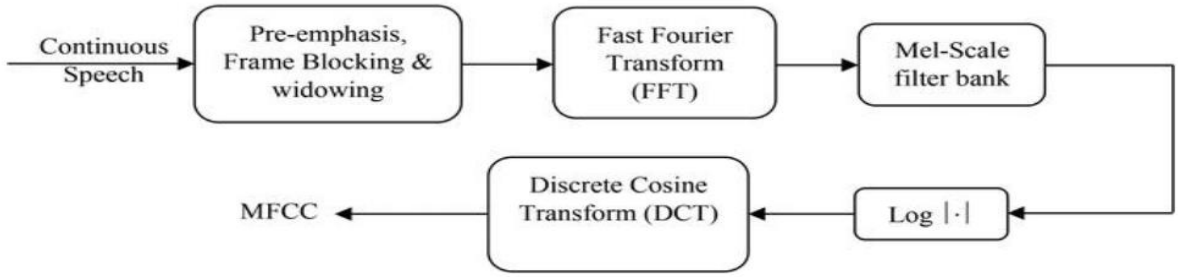

Fig.3. Block diagram of MFCC feature extraction

the signal is quasi in nature means it changing continuously in nature.In order to reduce the changes in a speech signals, it is required to split the signal into frames of 20-40 ms, where the changes are very less. In next step, Hamming windowing Technique is applied, to provide the original frequency spectrum. Then time domain signal is converted into frequency domain using discrete cosine transformation. In order to find out the power spectrum of a given speed signal, the Mel filter banks is applied. The resultant power spectrum indicates the frequencies in a frame.The first filter bank indicates how much energy is presented at 0 hertz. If the frequencies are high then the size of the filter bank also high. Once the filter bank energies are ready then find out cepstrum features of the speech signal which are very less correlated and these features are applied to any classification model. The formula used for converting frequency to Mel scale is

$$M(f)=1125 \log(1+ f/700) \text{ where f is actual frequency} \qquad (4)$$

We consider only 12 discrete cosine transformations only. Basically, there are 26 standard discrete cosine transformations but starting 12 will provide the complete information regarding the speech samples.

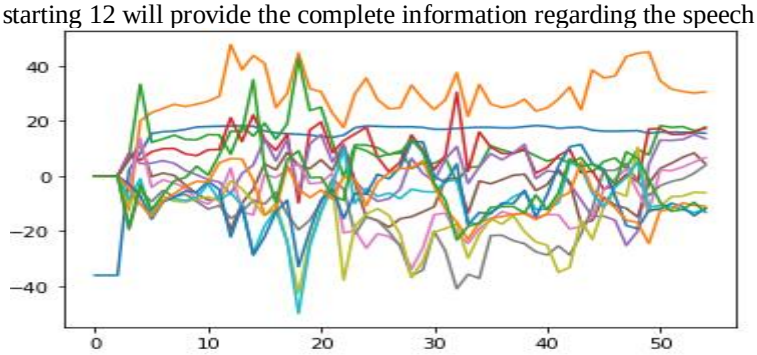

Fig.4. MFCC features of speech sample

3.2 Deltas and Delta-Deltas

Mfcc feature vectors works like human earing system represents the required frequencies information on windowing speech signal.These features are calculated from the power spectrum of speech signals. The information is also available in dynamics which is also very important i.e., the trajectory of MFCC coefficients. In order to find out dynamic information, we calculate Delta and Delta Delta coefficients. The formula for Delta coefficient is

$$d(t) = \frac{\sum_{m=1}^{M} m(P(t+m)-P(t-m)}{2\sum_{m=1}^{M} m\text{^}2} \qquad (5)$$

Where d(t), represents the Delta coefficient of t frame and it is calculated by static coefficient t+m to t-m where the value of M is 2. Delta-Delta coefficient also is calculated like this, but we will use Delta coefficients instead of static coefficients. The complete features are extracted from each frame as shown in the table [2].

Table.2. Complete MFCC Features extracted from speech signal

13	absolute	Energy (1) and MFCCs (12)
13	Delta	First order derivatives of the 13 absolute coefficients
13	Delta-Delta	Second order derivatives of the 13-absolute coefficients
39	total	MFCC feature vector

3.3 PROSODIC FEATURES

The prosodic features play a vital role to recognize the discriminate features of the standard language with respect to intensity, energy,pitch,rythem and intonation etc. The prosodic futures easily identify the differences in dialects of same language.

3.3.1 PITCH:

Frequency and pitch are very important features of any speech signal. The Pitch is the basic characteristic of any speech signal.Pitch represents that how vocal track fast vibrated the sound. If the vocal track fast vibrates the speech then the frequency is very high and the pitch also is more, if it is a low frequency then the pitch also be less. the frequency is proportional to the pitch. Frequency indicates that how a sound wave vibrates i.e., the number of wavelengths in a particular unit of time. As shown in the below figure the first diagram contains low frequency made it contains only 3 waves in the unit of time 1sec while the second diagram contains 6 waves in the unit of time 1 sec.The high and low pitch siganls as shown in below figure [5].

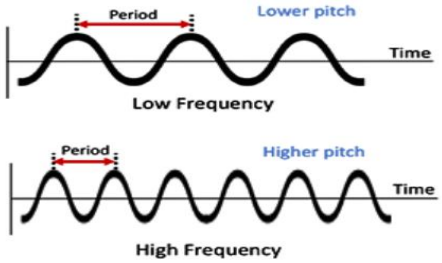

Fig.5. High and low pitch signals

From the above diagram

Pitch is propotional to Frequency and Frequency is inverse propotional to time, so that a low pitch for the audiot takes more time and a high pitch for the audio that takes less time. Pitch characters the ascent of the particular Language based on the vibration of vocal track when a human being speaks.So that, Pitch discrimantes the dialects of particular language. In this work, we extracted the pitch from three dialects of Telugu and it is observed that the pitch is very high for Telangana dialects and low for Andhra dialects.

3.3.2 LOUDNESS

Loudness is practically an ear sensation and It is a physiological process rather than a physical one. The dimension of hearing reflects loudness and the loudness of the speech signal completely depends upon the amplitude of the signal. If the amplitude of the signal is high then the speech is said to be loud. Loudness is proportional to the square of amplitude and it is measured in decibel (dB). If the speech signal is very loud as per required dB , it will provides the good information for models with clear features. The difference between loud and soft sound signals is shown in figure [6].

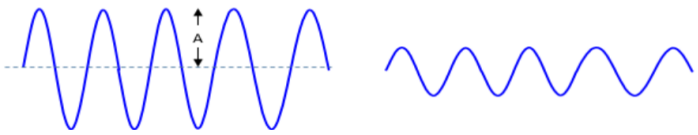

Fig.6. a. Loud sound b. Soft sound

As shown in the figure [4], the Loud sound contains more amplitude and the soft or quiet sound contains less amplitude in signal vibration.

In this, we extracted the Loudness from different speech samples and combine it with spectral features to identify the dialects of the Telugu language.

3.3.3 INTENSITY

Sound intensity is nothing but power in speech signal passes through surface in particular duration. We can also define like sound energy passes thorough the unit surface area in 1sec. Intensity is promotional to the pitch. We can measure the intensity in different S.I units.

$$\text{Intensity(I)}=\frac{Energy}{Area.time} \Rightarrow I=\frac{E}{A.T}$$

Energy is measured in Joule and area in meater^2 and time in seconds then Intensity (I)= $\frac{Joule}{m\text{^}2.sec}$

When the sound generated by source, it spreads over the area in spherically. The intensity at a particular point in the area is calculated by using the equation.

Power of signal (P)= Energy/Time and Intensity(I)=$\frac{Power}{Area}$ then I= watt/m^2

Then

I= P/4πr^2 where r is the distance between source and object in the area.

From above equation

I α 1/r^2 means intensity is high at the source point and less when the distance is increased.

In this, we extracted the intensity levels of speech signal to find out the dialects because three different regions have the different intensity levels for same word. The same word is spoken the different in different regions of Telugu. To discriminate clearly, we extracted the intensity from speech signals of different regions.

3.3.4 ENERGY

Energy of speech signal is produced by vibrating the vocal folds and when these vibrations reached to our ear drum, and then we can hear the sound. If the energy is in between 16Hz to 20kHz then human can hear the sound.

If the sound energy is more than 20kHz we called the sound is ultrasound which humans cannot hear. The Energy of speech signal is proportional to pitch and also depending upon the amplitude and frequency of signal.
Energy(E) is proportional to amplitude^2.frequency^2
From above equation the energy of speech signal is high for high frequency signals. If signal carries more energy then, it contains the more information. The energy of sound if combination of potential and kinetic energies.

$$W = W_{\text{potential}} + W_{\text{kinetic}} = \int_V \frac{p^2}{2\rho_0 c^2}\,\mathrm{d}V + \int_V \frac{\rho v^2}{2}\,\mathrm{d}V,$$

From the above equation
V is the volume
P is the sound pressure
c is speed of sound
ρ_0 is the density without sound
ρ density of the medium.
In this paper, we extracted the energy of speech signals belongs to different regions and apply to the models. It is observed that energy is high for Telangana and low for Andhra dialects. Rayalaseema contains the medium energy. This energy plays vital role when the one word is spoken same in three different regions.

3.3.5 FORMANTS

In the field of both speech production and perception, the formants or vocal-tract resonances play a dominant role. Formants originate in the vocal tract as it represents vocal track frequency. Based on the opening size and shape, the air within the vocal tract vibrates at varying pitches. Formants values will be changed with the size and shape of vocal tract. The formats of the speech signal as shown below figure [7].

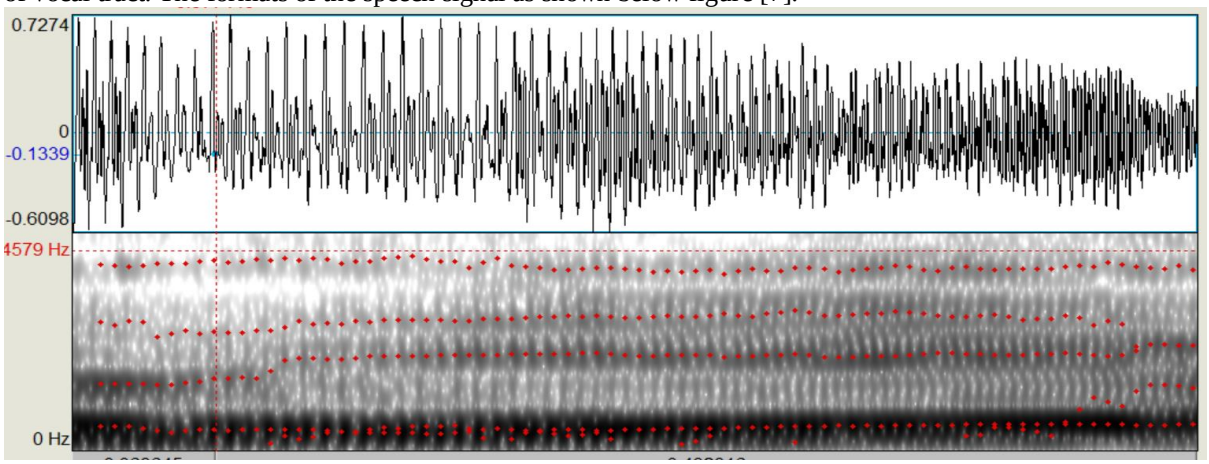

Fig.7. Formants representation of Speech signal

In the above figure [7], the red lines are formants of the speech signal. There are many number of formats but only first formants are providing the exact information of the speech. Basically we considered 4 formats represents by f0,f1,f2 and f3.f0 is also called as fundamental frequency. In the formats there are straight and dot lines are presented. The dot lines are not clearly represents the information.

3.4 Zero Crossing Rate (ZCR):

ZCR indicates the number of times the amplitude of the speech signal passes through the timeline axes or horizontal axes. It is mainly used in the separation of voice and unvoiced part from the speech signal. If the sign wave moving from positive to negative then the value of ZCR is 1 otherwise value 0.ZCR value is very high for the unvoiced part and low for the voiced part of the speech signal. The number of zero-crossing points is calculated for a speech signal is calculate the ZCR rate and divided by the number of frames as shown in the below equation (6).

$$Z(i)= \frac{1}{2Wl}\sum_{m=1}^{Wl}|sgn[si(m)] - sgn[si(m-1)])| \qquad (6)$$

where sgn () is the sign function and the value of the function can be defined by like

$$\text{sgn[si(m)]} = \begin{cases} 1, & si(m) \geq 0, \\ -1, & si(m) < 0. \end{cases}$$

If the amplitude of the signal moving from positive to negative then the value is 1 otherwise the value is -1. W_L is representing the number of frames and Si is the original speech signal. The following figure [8] shows the original signal and corresponding ZCR.

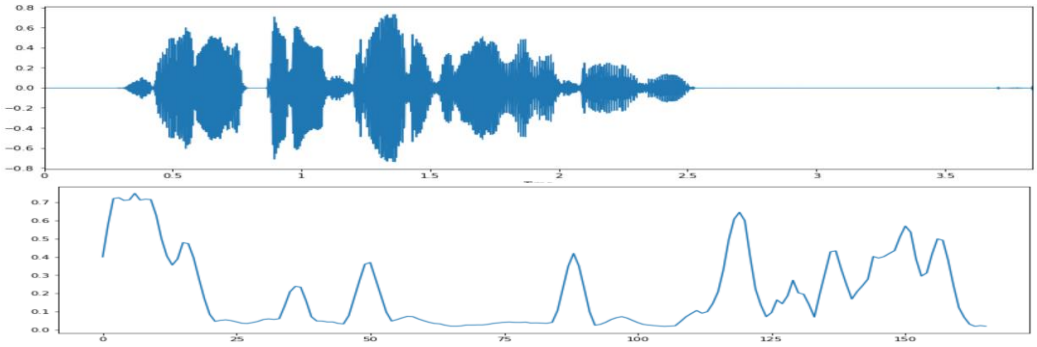

Fig.8.(a) Original signal (b). ZCR values

4. The Hidden Markov model (HMM)

It is a probabilistic model that depends upon the Markov chain property. The markov property indicates that the future sequence or state completely depends upon the current state, not on previous states. This property also used to assign the values to random variables and states. It is a popular statistical model in speech processing and recognition applications. The Hidden Markov model consists of four elements namely states, observable sequence, transition and initial probabilities.

$S=s_1, s_2 \ldots s_N$ Number of hidden states

$O=o_1, o_2 \ldots o_T$, Number of observable sequence.

A= a11,a12,a13..a1n…ann, is transition probability matrix where aij represents the probability moving from state i to state j. The all the moves corresponding to a_{ij} are equal to 1.

B=b1, b2…bn is the emission probability.

Π= π1 π2 π3… πn. represents the initial probabilities. Π1 represents the probability to start initially from satae1. If π2=0 is indicating that state2 is not initial state. In our experiments, we have three hidden states i.e., Telangana, Costa Andhra and Rayalaseema. And we are used spectral and prosodic features like MFCC, PITCH and Loudness as observable states. So, there is 3 hidden states and 3 observables state, so $T= 3^3=27$ possible observations are occurred. To, find out the state of the event, calculate maximum likelihood. This will calculate by using the joint probability as shown below

$$P(O,Q) = P(O|Q) \times P(Q) = \prod_{i=1}^{T} P(o_i|q_i) \times \prod_{i=1}^{T} P(q_i|q_{i-1}) \quad (8)$$

In the above equation can be explained like

P (MFCC, PITCH, LOUDNESS, TS, CA, RS) = P(MFCC/TS) *P(PITCH/CA) *P(LOUDNESS/RS) *λ

λ = P(TS)*P(CA/TS) *P(RS/CA) (9)

Where P(S) will not depend on any previous state we considered it as initial probability and remaining part of λ, indicates the transition probabilities and the first part is emission probabilities of states.

4.2 GAUSSIAN MIXTURE MODEL

GMM is a probabilistic model and it works based on the mean as well as the variance of the data. The Gaussian mixture models contain some random distribution one for each cluster. It produces the output like, all the data points belong to some distribution and together follow the normal distribution. For example, if we have three Gaussian distributions G1, G2, G3 then these contain means (μ1, μ2, μ3) and variance (σ1, σ2, σ3) respectively and for the single-dimensional data points the probability density function given in eq. (10).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (10)$$

If more number of dimensional data, we consider covariance of the data matrix. It is used to find the shape of the cluster also. For N-dimensional data points, PDF is defined in eq. (11).

$$f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \quad (11)$$

In GMM we used Expectation-Maximization (E-M) algorithm to assign the data point to a particular cluster. It is used to estimate the maximum likelihood of model parameters. The E-M method contains two steps. E-step is used to estimate the missing value by using available data and M-step is used, after finding the E-step, then modify the model parameters like mean and the number of values in each distribution and covariance. This procedure applied in an iterative manner.

4.3 K-Nearest Neighbour(K-NN)

The KNN algorithm is supervised learning used for the classification as well as Regression algorithms. This algorithm also called as lazy learner algorithm because it will not learn from the training data set. The K-NN algorithm in the training phase stores the data given as input and in the testing phase, when the speech sample is given as input then it will classify speech samples based on the similarity. The K-NN algorithm is a non-parametric algorithm because it will not do any assumptions about the parameters based on underlying data. In order to find out the similarity between speech samples, we used Euclidian distance. Classification of a new data points by KNN model as shown below figure 9.

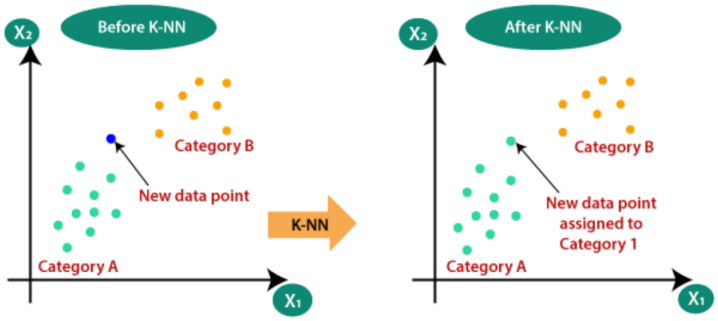

Fig.9. K-NN model

In this, we used the K-NN model to classify the dialects of the Telugu Language. For this purpose, extracted different prosodic features like Pitch/Loudness/Intensity/Energy/Formants from speech samples are given as input. The value of the K is 1 in this identification dialect also called a 1-NN model. It is a very simple machine learning algorithm mostly used in almost all speech-processing applications. The accuracies produced by the 1-NN model as shown in table[4].

4.4. Methodology

We proposed a methodology, which consists of two phases namely Training phase and testing phase. The speech samples, which are collected and pre-processed for database creation, are used for both Training and Testing separately. The block diagram of proposed method for dialect identification as shown in figure [10].

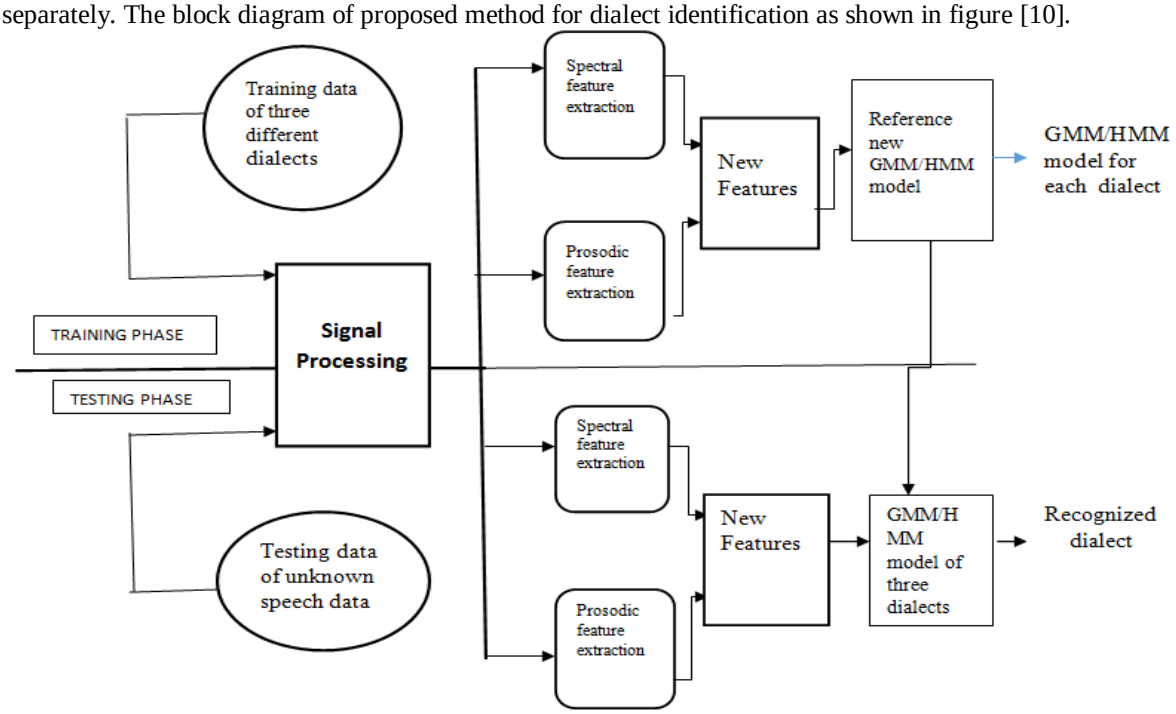

Fig.10. Proposed method for dialect identification

In the above diagram, to identify the dialects of the Telugu language, the proposed methodology has two phases training phase and the testing phase. In the training phase, our created data set (70% of data) is supplied as input. Then extract the different features from speech samples like MFCC/ZCR/Prosodic features. In order to increase the accuracy and also clearly identify the words which are spoken similar way in different dialects, we combined the Spectral features with prosodic features like MFCC+Pitch/MFCC+ZCR/MFCC+Pitch+Loudness to form a new feature vector. These feature vectors are given input to the different classification methods like GMM AND HMM model. There are three different models each one for dialect (i.e. GMM for Telangana, Costa Andhra, and Rayalaseema like HMM also) is generated as output from the training phase.

In the testing phase of the proposed system, the same set of features MFCC/ZCR/Prosodic features are extracted from speech samples whose class label is unknown. These feature vectors are supplied to build models in the training phase as input. The models calculate the likelihood score from each feature vector with respective the different dialects. Denoted by β1, β2, β3 where β1 is score generated by Telangana GMM model, β2 is score generated by Costa Andhra GMM model and β3 is likelihood score generated by Rayalaseema GMM model. Out of these 3 values, which value is higher, then the given speech sample belongs to that particular dialect.

5. Results

In this, we analyze and compare the performance of different models with different feature vectors are analysed and results are depicted in following tables.

Table.3.Accuracies produced by GMM and HMM model with original features

Feature extraction techniques	Model	Accuracy	Best Model
MFCC	HMM	72.2	HMM
	GMM	70.2	
MFCC+ΔMFCC	HMM	73.833	GMM
	GMM	81.6	
MFCC+ΔMFCC+ΔΔMFCC	HMM	82.6	BOTH
	GMM	82.6	
MFCC+PITCH	HMM	84.05	BOTH
	GMM	84.05	
MFCC+PITCH +LOUDNESS	HMM	86.95	GMM
	GMM	88.4	
ZCR	HMM	73.33	HMM
	GMM	63.3	
MFCC+ΔMFCC+ΔΔMFCC+ZCR	HMM	81.66	GMM
	GMM	83.3	

From the above table we can observe that in our work we had deployed various feature extraction techniques in order to retrieve the essential features. Those features are provided as input to both of our models namely HMM, GMM. Based upon the accuracies achieved, we can identify the best model between them for the selected feature extraction technique. At first, we used MFCC (13 features) as feature extraction technique then we observed HMM as the best model as it gave 72.2% accuracy whereas GMM gave 70.2% accuracy. Secondly, we calculated ΔMFCC features from MFCC and concatenate with MFCC i.e., MFCC+ΔMFCC (13MFCC+13 Delta MFCC). Total 26-feature vector is applied to models. We found out GMM as the best model as it gave 81.6% accuracy while HMM gave 73.833% accuracy. Then we made use of MFCC+ΔMFCC+ΔΔMFCC (13MFCC+13 Delta MFCC+13 Delta-Delta MFCC) and MFCC+PITCH for feature extraction purposes where we noted both HMM and GMM exhibited equal performance with 82.6% and 84.05% respectively. Later we chose MFCC+PITCH +LOUDNESS for feature extraction then GMM outperformed with 88.4% where as HMM gave 86.95% accuracy. ZCR when selected for feature extraction HMM was resulted as the best model with 73.33% and GMM recorded 63.3% accuracy. GMM was resulted as the best model as it gave an accuracy of 83.3% where as HMM gave 81.66% accuracy when MFCC+ΔMFCC+ΔΔMFCC+ZCR was selected as feature extraction technique. From the above all the features, GMM provides the good accuracy with 88.4% for MFCC+PITCH +LOUDNESS.

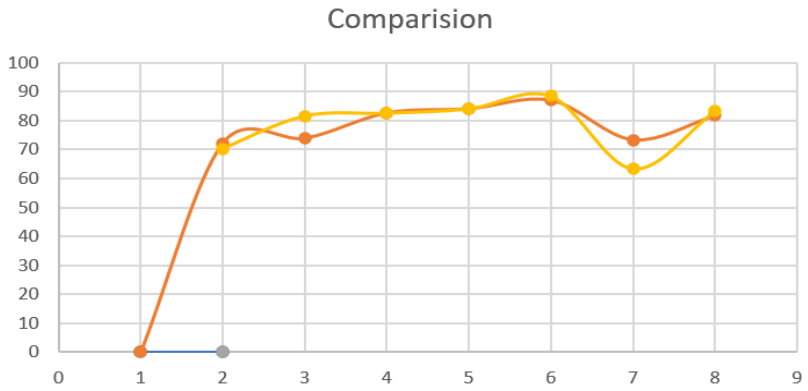

Fig.11. Comparison of Models

The above figure [8] shows the differences between accuracies given by models. The brown colour indicates the HMM model accuracies and yellow colour indicates the GMM model accuracies. By observing the graph, GMM model almost produced slightly more accuracy compare to HMM model, except the ZCR feature extraction technique.

Table.4. Accuracies given by KNN model with different prosodic features.

Model	Feature extraction Method	Accuracy
K-NN	Pitch	75
	Intensity	60
	Energy	61.6
	Formant(f1)	65
	Formant(f2)	46.6
	Formant(f3)	48.5
	Formant(f4)	56.6
	Formant(avg.)	43.3
	Pitch+intensity	76.6
	Pitch+Intensity+f1	71.66
	Pitch+Intensity+Energy	76.6

In above table gives the different accuracies produced by the KNN model with different prosodic features like pitch, Intensity, etc. to identify the dialects of the Telugu language. The K-NN model, where K represents the number of data points to consider and it is an integer number. In our research, we consider K value is the 1 called 1-NN model. The prosodic features provide the important cues regarding vocal tract, pressure, shape. These features are very well used in speech processing applications. When we applied Pitch features to the KNN model, it produces 75% accuracy. Through the intensity and energy features, the KNN model produces 60% and 61.6% accuracy. We extracted the different formats from speech applied to the model. It is observed that formats provide less accuracy compared to the remaining features. It produces 65%, 46.6%, 48.5% and 56.6% respectively. Some words in different regions pronounced in the same way. In order to clear discriminate those words, we combine the prosodic features called hybrid features. In this, we combine the Pitch +Intensity Pitch+Intensity+Energy and Pitch+Intensity+f1.Pitch+Intensity+Energy and Pitch+Intensity both features produce the highest and same accuracy. It produces 76.6% and Pitch+Intensity+f1 produces 1.66. The below figure [12] shows the accuracies produced by the KNN model.

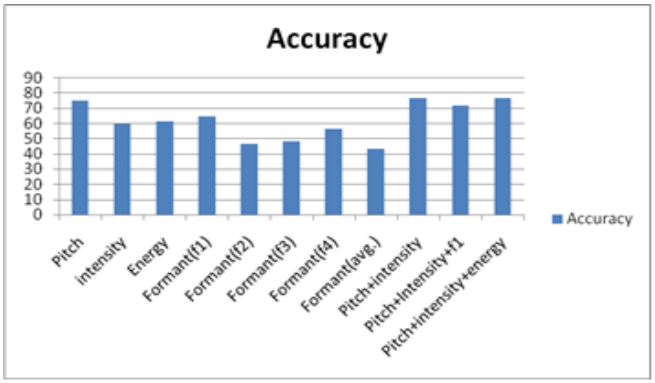

Fig.12. Accuracies by different prosodic features

6. Conclusion

In this, we are comparing the performance of different models by applying the different feature extraction techniques to find the dialects of the Telugu Language. For this purpose, we created a standard database. The features are used here are Spectral features like MFCC and its types, Prosodic features like Pitch, loudness, Intensity, etc., and temporal features like ZCR. The models used in this is the HMM and GMM and K-NN Models. From overall observation, the GMM model slightly produces good results compared to the HMM model. In the future, we can also apply the deep learning models to improve the accuracy.

REFERENCES

1. "Abstract of speakers' strength of languages and mother tongues – 2000". Census of India, Archived from the original on 29 October 2013.
2. "Telugu gets classical status". The Times of India. 1 October 2008. Archived from the Original on 4th November, 2008.

3. S.Shivaprsad M.Sadanandam "Identification of regional dialects of Telugu language using text independent speech processing models" International Journal of Speech Technology Vol23 issue1 2020
4. Mehrabani, Mahnoosh, and John H. L. Hansen. "Automatic analysis of dialect/ language sets", International Journal of Speech Technology, 2015
5.Nagaratna B. Chittaragi, Ambareesh Prakash & Shashidhar G. Koolagudi Dialect Identification Using Spectral and Prosodic Features on Single and Ensemble Classifiers , 2018.
6. "Infographic: A World of Languages" Retrieved 2 June 2018.
7. Imene Guellil, Faical Azouaou. "Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect", IEEE Intl Conference on Computational Science and Engineering. DCABES 2016.
8.V V Sreeraj, Rajeev Rajan. "Automatic dialect recognition using feature fusion", International Conference on Trends in Electronics and Informatics (ICEI), 2017.
9. Hoang Trang, Tran Hoang Loc, Huynh Bui Hoang Nam "Proposed combination of PCA and MFCC feature extraction in speech recognition system", International Conference on Advanced Technologies for Communications (ATC 2014), Hanoi, pp. 697-702, 2014.doi: 10.1109/ATC.2014.7043477.
10.Oh-Wook Kwon, Kwokleung Chan, Te-Won Lee. "Speech feature analysis using variational Bayesian PCA", IEEE Signal Processing Letters, vol. 10, no. 5, pp. 137-140, May 2003, doi:10.1109/LSP.2003.810017.
11. L. Gang, H. John and L. Hansen, "A Systematic Strategy for Robust Automatic Dialect Identification", *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 2138-2141, 2011.