MODELLING AND ANALYSIS OF ACCENT BASED RECOGNITION AND SPEAKER IDENTIFICATION SYSTEM

Kasiprasad Mannepalli[1], Panyam Narahari Sastry[2] and V. Rajesh[1]
[1]K L University, KLEF, Vijayawada, Andhra Pradesh, India
[2]CBIT, Hyderabad Telangana, India
E-Mail: mkasiprasad@gmail.com

ABSTRACT

Speech processing has emerged as one of the most important research areas of signal processing. In this research area speaker identification and accent recognition are the different types of research applications of speech processing. Nowadays the speech processing has become essential for technological developments in various aspects and this technology is also incorporated in many electronic gadgets. Accent is a kind of modulation of speech in the same language where as speaker identification is to identify the person from the known set of speakers (closed-set). The Intensity, Energy, Spectral Density, Formants and Prosodic features are some of the features that vary with a language, climate and socio economic background. It is preferred to recognize the Accent first and then to identify the speaker. Telugu as a language basically has three accents Telangana, Rayalseema and Coastal Andhra. This Work aims to differentiate the Accent of Telugu language of different speakers from Telugu speaking areas and further to identity the Speaker. The speeches of many speakers from various Telugu speaking areas are collected, which are text dependent. Then feature extraction is carried out for these speech samples. An Algorithm is proposed to recognize the Accent and also the Speaker based on the features extracted. The recognition accuracy to recognize the speaker based on accent is 78 %, whereas the best recognition accuracy for identifying the speaker is obtained as 74%.

Keywords: speaker identification, accent identification, Telugu speech.

INTRODUCTION

Spoken language is the natural way used by humans to communicate information. The speech signal conveys several types of information. From the speech production point of view, the speech signal conveys linguistic information (e.g., message and language) and speaker information (e.g., emotional, regional, and physiological characteristics). From the speech perception point of view, it also conveys information about the environment [1] in which the speech was produced and transmitted. Even though this wide range of information is encoded in a complex form that is into the speech signal, humans can easily decode most of the information. Such human ability has inspired many researchers to understand speech production and perception for developing systems that automatically extract and process the richness of information in speech and can be used for the purpose of human-machine interaction. This speech technology [2] has found wide applications such as automatic dictation, voice command control, audio archive indexing and retrieval etc. The various applications can be used to extract the information in speech of what we required. For example, the speaker information will be relevant if the goal is to recognize the accent from the words that the speaker is producing or it can be used to find out the emotion of speaker from the speech. The presence of irrelevant information like noise present in the speech may actually degrade the system accuracy.

IMPORTANCE OF SPEECH /SPEAKER RECOGNITION

Speech/Speaker recognition is a growing technology in different areas for its various applications. In order to provide a comfortable and natural form of communication which is being able to speak to your personal computer, and have it recognize and respond what you are saying is a good technological development. As it reduce the amount of typing what you have to do, leave your hands free, and allow you to move away from the screen and your personal computer can be made secure with your voice, as it can open only with your voice. Currently there is a much interest in Speech recognition, and performance is improving.

Main purpose of speech processing [3] is designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has interested engineers and scientists for centuries. Speaker recognition has a history back some four decades and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style).

METHODS OF SPEECH RECOGNITION

The earliest attempts for Speech processing were made in the 1950's, when researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, a system for isolated digit recognition for a single speaker was built [5]. In early ASR designs, a recognizer would segment the speech into basic pronunciation units, and then identify the individual corresponding to the segments. By the early 1970's, largely as a result of development in electronics and information sciences, the prospects had changed. The computer had evolved into a powerful and flexible processor. In the late of 1970's, the research tended from isolated word recognition problem to connected word recognition problem. In 1979, Sakoe and Chiba [6] offered a two-level Dynamic Programming (DP) algorithm to solve the connected word recognition problem. Rabiner and Myers in 1981 [7] proposed a template matching technique called Dynamic Time Warping (DTW) similar to DP algorithm. It was more flexible and efficient, but at the same time more complex. In 1980's ASR research took a new direction due to the introduction of statistical modeling methods; especially the Hidden Markov Model (HMM) approaches. Refinements in the theory and implementation of Markov modeling techniques have greatly improved the ASR applications [8].

Along with the development of speech recognition, speaker recognition technique had also evolved a lot in last five decades. The first attempts for automatic speaker recognition were made in the 1960s, one decade later than that for Automatic Speech Recognition. Pruzansky at Bell Labs [9] was among the first to initiate research by using filter banks and correlating to digital spectrograms for a similarity measure. Pruzansky and Mathews [10] improved upon this technique and Li *et al*. [11] further developed it by using linear discriminators. Intra-speaker variability of features, one of the most serious problems in speaker recognition, was intensively investigated by Endres and Furui [12]. For the purpose of extracting speaker features independent of the phonetic context, various parameters were extracted by averaging over long enough duration or by extracting statistical or predictive parameters. Averaged auto-correlation [13], instantaneous spectra covariance matrix [14], spectrum and fundamental frequency histograms [15], Linear Prediction Coefficients [16], and Mel Frequency Cepstral Coefficients (MFCC) [17] are some of the techniques that received serious attention from the research community.

As a nonparametric model, vector quantization (VQ) [18] was investigated. A set of short-time training feature vectors of a speaker can be efficiently compressed to a small set of representative points, a so-called VQ codebook. As a parametric model, HMM was investigated. An utterance was characterized as a sequence of transitions through a 5-state HMM in the acoustic feature space. Tishby [19] expanded Poritz's [20] idea by using an 8-state ergodic autoregressive HMM represented by continuous probability density functions with 2 to 8 mixture components per state, which had a higher spectral resolution than the Poritz's model. Rose *et al*. [21] proposed using a single state.

Research on increasing robustness became a central theme in the 1990s. Matsui *et al*. [22] compared the VQ-based method with the discrete/continuous ergodic HMM-based method, particularly from the viewpoint of robustness against utterance variations. They found that the continuous ergodic HMM method is far superior to the discrete ergodic HMM method and that the continuous ergodic HMM method is as robust as the VQ-based method when enough training data is available. They investigated speaker identification rates using the continuous HMM as a function of the number of states and mixtures. It was shown that speaker recognition rates were strongly correlated with the total number of mixtures, irrespective of the number of states. This means that using information about transitions between different states is ineffective for text-independent speaker recognition and, therefore, GMM achieves almost the same performance as the multiple-state ergodic HMM.

M. kasiprasad, P. Narahari Sastry, V. Rajesh [23] have achieved a recognition accuracy of 78% for speaker identification by using features like Pitch, Formants F1, F2, F3 and energy as reported in their paper.

In 2000's A family of new normalization techniques has recently been proposed, in which the values are normalized by subtracting the mean and then dividing by standard deviation, both terms having been estimated from the (pseudo) imposter score distribution. Different possibilities are available for computing the imposter score distribution: Znorm, Hnorm, Tnorm, Htnorm, Cnorm and Dnorm [24]. The text-independent speaker verification techniques associate one or several parameterization level normalizations (CMS, feature variance normalization, feature warping, etc.) with world model normalization and one or several score normalizations. High-level features such as word idiolect, pronunciation, phone usage, prosody, etc. have been successfully used in texts-independent speaker verification.

DATA ACQUISTION AND METHODOLOGY

Features used

i. Pitch

Pitch is a perceptual property that allows ordering of sounds on a frequency related scale. The variations in the fundamental frequency during the duration of utterance if followed would provide the contour which can be used as a feature for speech recognition. The speech utterance is normalized and contour is determined. The normalization of speech utterance is required because the accurate time alignment of utterances is crucial and the same speaker utterances could be interpreted as utterances from two different speakers. The contour is divided into set of segments and measured pitch values are averaged over the whole segment. The vector that contains average values of pitch of all segments is used for speaker recognition.

ii. Formants

It is defined as the spectral peaks of the sound spectrum of the voice. The frequency components of human speech formants with F1, F2, F3. The arranging of formants starting from increasing order with low frequency F1 to high frequency F3. F1 and F2 are the distinguish vowels. The two determine the quality of vowels open or close front or back. F1 is assigned higher frequency for ‘a’ and lower frequency for close vowel ‘i’ and ‘u’. F2 is assigned as higher frequency for front vowel ‘i’ and lower frequency for back vowel ‘u’.

Periodic excitation is seen in the spectrum of certain sounds especially vowels. Speech organs form certain shape to produce vowel sound and regions of resonance and anti-resonance are formed in the vocal tract, location of these resonances in the frequency spectrum depends on form and shape of vocal tract. Since speech organs is characteristic for each speaker and difference in frequencies can also found in position of their formant frequencies. As these effect the overall spectrum shape as these formant frequencies are sampled at a rate used for speaker recognition.

iii. Energy function

It is computed by splitting the speech signal into frames of N samples and computing total squared values of signal samples in each frame. For a short time interval of speech signal it is considered to be stationary and these frames are multiplied by a window.

$$E=\sum_n \{x(n) * w(n-m)\}^2 \qquad (1)$$

From the above equation (1), E describes about energy and x(n) represents the speech signal and w(n-m) represents the window. Through this equation energy of a short time speech signal is calculated.

iv. Power spectral density

The speech signal which is taken from training sample is performed mathematical auto correlation which is to provide the relation of the same sample and after performing it is to undergo Fourier transform which gives the power spectral density. This converts the time domain waveform to frequency domain to analyze the signal.

v. Prosodic features

Prosodic features are supra segmental. They are not confined to any one segment, but occur in some higher level of an utterance. These prosodic features units are the actual phonetic “spurts” or chunks of speech. They need not correspond to grammatical units such as phrases and clauses. Prosodic units are marked by phonetic cues. Phonetic cues can include aspects of prosody such as Pitch, and Accents, all of which are cues that must be analyzed in context, or in comparison to other aspects of a sentence. Pitch, for example, can change over the course of sentence, falling intonations. Prosody helps in resolving sentence ambiguity, but when the sentence is read aloud, prosodic cues like pauses and changes in intonation will make the meaning clear. More recently, there have been renewed and more successful efforts to find ways of incorporating prosodic information into a wider variety of ASR- related tasks, such as identifying speech. Prosodic features helps in Speech processing for language modelling, acoustic modelling, speaker identification and identification of Accent and emotion of speech.

SPEECH SAMPLES

Speech Acquisition has been made from three different regions of Andhra Pradesh. The speech samples are collected from Coastal Andhra, Rayalaseema and Telangana of the three popularly known Accents in Andhra Pradesh. This work is based on text-dependent, in order to find out the sentence which is to be uttered that basically focuses on the Accent and will be familiar to the three regions, we have found out a sentence which plays a crucial role in identifying Accent and Speaker Identification. The sentence in Telugu “ఎవరోఅన్నంతిన్నారునేనుఎవరినిచూడలేదు“ (evarooannamtinnaarunenuevarinichuudaledu) is used to collectthe Speeches from these regions wiz Telangana, Coastal Andhra, Rayalaseema for 39 speakers wherein 13 speakers from each region including male and female are present. The utterance of a given sentence from each speaker is recorded for five times using HTC smart phone under noisy conditions. The speeches which are collected from these three regions are in the .amr format and these speeches are converted in to .mp3 and .wav format for convenient purpose. As Matlab can accept the speech waves in .wav format so it has been converted to the required format. Speeches are recoded from less noisy conditions for better result and conversion in to various formants is also taken in a careful way.

FEATURE EXTRACTION FROM A GIVEN SPEECH

(i) Collected the speeches in Telugu language from the various regions viz Coastal Andhra, Rayalaseema, and Telangana for 39 speakers (13 speakers from each region). Each speaker spoke the same sentence for five times. The first three times speech is used to develop the database. The other two speech signals of the same sentence are used for testing the recognition.

(ii) “COLEA” tool was used to extract the features from the speech of the above mentioned speakers. This COLEA tool successfully extracts Formants (F0, F1, F2, and F3), Energy and Power Spectral Density (PSD).

One of the speech samples is taken to describe about various features. The Figure-1 represents the Pitch of a person’s speech in a sentence as follows, where the x-axis represents the time domain and y-axis represents the frequency of the speech.

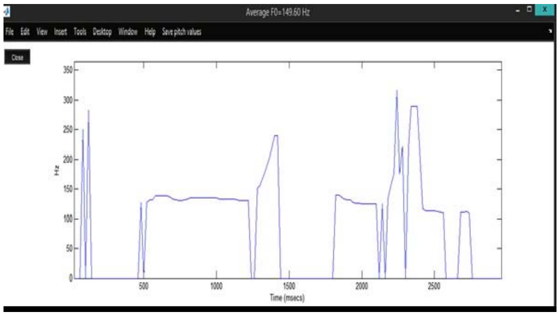

Figure-1. Pitch of a person’s speech in a sentence.

The Figure-2 shows the three formants F1, F2 and F3, where F1 will be ranging up to 900 Hz and F2 will be ranging up to 2300 Hz and F3 ranging from 2300 Hz to 3400 Hz. In this Figure-2 the x-axis represents the time domain and the y-axis represents the frequency of the speech.

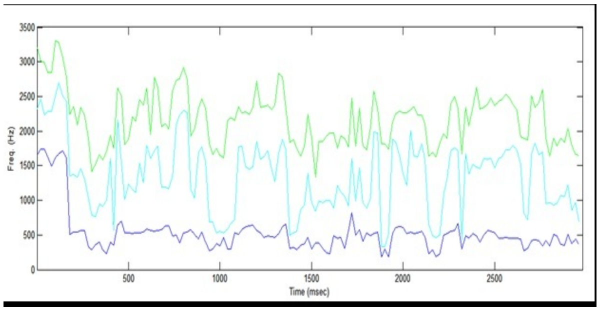

Figure-2. Formants F1, F2 and F3.

The energy of a speech sign al is shown in 3 (a) in dB and the corresponding speech signal is shown in 3.3 (b) represented in amplitude

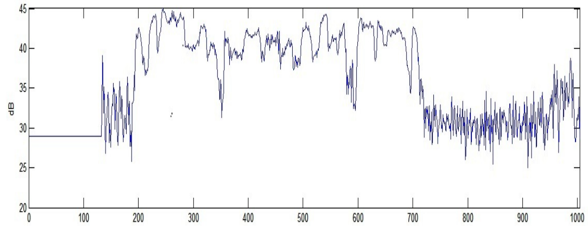

Figure-3. Energy plot of a Speech sentence.

The power spectral density is obtained by applying auto correlation to the speech signal and further processed with Fourier transform, which is shown in Figure-3. The corresponding data statistics is also shown below and the mean value of Y column is considered as a feature for each of the speech signal

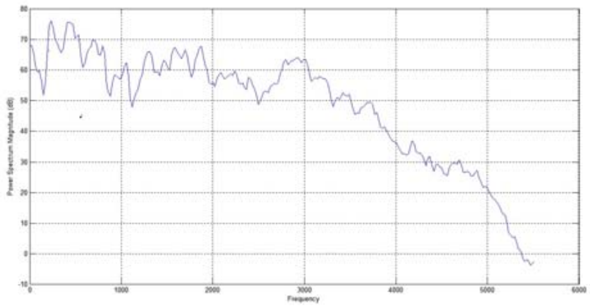

Figure-4. Power spectral density plot.

(iii) "PRAAT" tool is used to extract prosodic features from the speech samples of all the speakers.

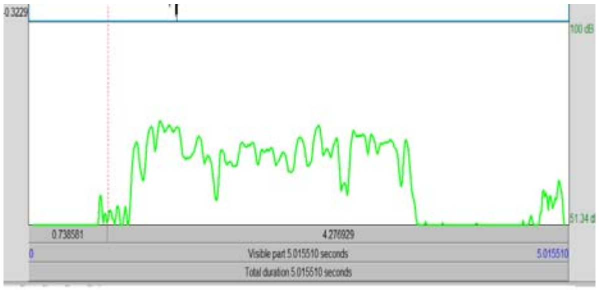
0.3229
100 dB
51.34 d
0.738581
4.276929
0
Visible part 5.015510 seconds
5.015510
Total duration 5.015510 seconds

Figure-5. Intensity of speech signal.

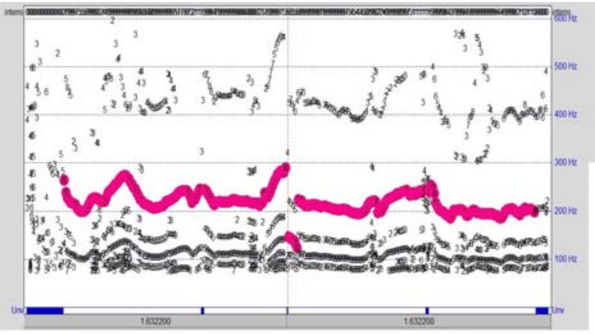

Figure-6. Prosodic features.

The Figure-6 shown above used to extract the prosodic features:

a) Digits between 0 and 9 scattered all over the drawing area. Their locations represent the pitch of which there are several for every time frame. The digits themselves represent the goodness of a candidate, multiplied by ten. For instance, if you see a "9" at the location (1.23 seconds, 189 hertz), this means that in the timeframe at 1.23 seconds, there is a pitch candidate with a value of 189 hertz, and its goodness is 0.9. The number 0.9 may be the relative height of an autocorrelation peak, a cross-correlation peak, or a spectral peak, depending on the method bywhich the Pitch object was computed.
b) A path of red disks. These disks represent the best path through the candidates, i.e. our best guess at what the pitch contour is. The path will usually have been determined by the path finder, which was called by the pitch-extraction algorithm, and you can change the path manually. The path finder takes into account the goodness of each candidate, the intensity of the sound in the frame, voiced-unvoiced transitions, and frequency jumps. It also determines whether each frame is voiced or unvoiced.

By using above Figure-7 the features Minimum, Maximum, Range, Standard deviation, and Mean absolute slope are tabulated.

ALGORITHM DEVELOPMENT

The first step in the development of this proposed method of accent based recognition system and speaker identification is to collect data from different speakers of various areas belonging to Coastal Andhra Pradesh. The various areas identified are Telangana, Coastal Andhra and Rayalaseema. For experimental evaluation thirteen speakers from each region are selected. The speeches were recorded using a high quality smart phone of HTC model. These speeches were recorded in ".amr" format, the facility available in the smart phone. The sentence selected for testing is text-dependent"ఎవరోఅన్నంతిన్నారునేనుఎవరినిచూడలేదు" (evarooannamtinnaarunenuevarinichuudaledu).

In the next step, every voice sample is converted from ".amr format" to ".mp3 format" using media software. Further the speeches are also obtained in ".wav format" using the same media software for further processing. The ".mp3 format" is very much useful for hearing the speech sentence of any speaker, the ".wav format" is useful for further processing in the proposed algorithm using Matlab. The various features of each of the 195 speeches are extracted using "Colea" software. The selected set of features is formants F1, F2, F3, pitch, energy and power spectral density. Further, Praat tool is used to extract prosodic features to add to the above set of features selected for the proposed algorithm.

For the development of database comprising training samples, speeches of the thirteen speakers of each region i.e. Coastal Andhra, Rayalaseema and Telangana are selected and the three speech samples of each speaker are taken to train the computer system knowledge / database. The speeches of the remaining two speech samples are taken for test samples. Hence the extracted features of all the three regions and all the speeches are divided into database (training samples) and test samples.

In this proposed method, there are two objectives i.e.

a) Speaker recognition based on accent

To achieve the above said first objective the step by step algorithm is shown below in Figure-7. For the speaker recognition based on accent, the features extracted from all the speakers were arranged based on their region for the training samples and from these samples the average of each parameter of the respected region from all the speakers are averaged and procedure is repeated for the other regions and this averaged feature set is used for training the system in the proposed algorithm. However for the test samples, the extracted features are directly used or given as input in the proposed method. The database/training samples set consist of 117 speeches of three regions as described earlier. The distance between the extracted feature set of test sample and feature set of each of the database sample is calculated. Hence 78 test samples are given as input to compute distances for each test sample. Here the nearest neighbourhood (NNC) method is used to compute distances. The minimum distance is found out from the three regions. Hence it is inferred that, test sample tested is identified to be person or region corresponding to the minimum distance.

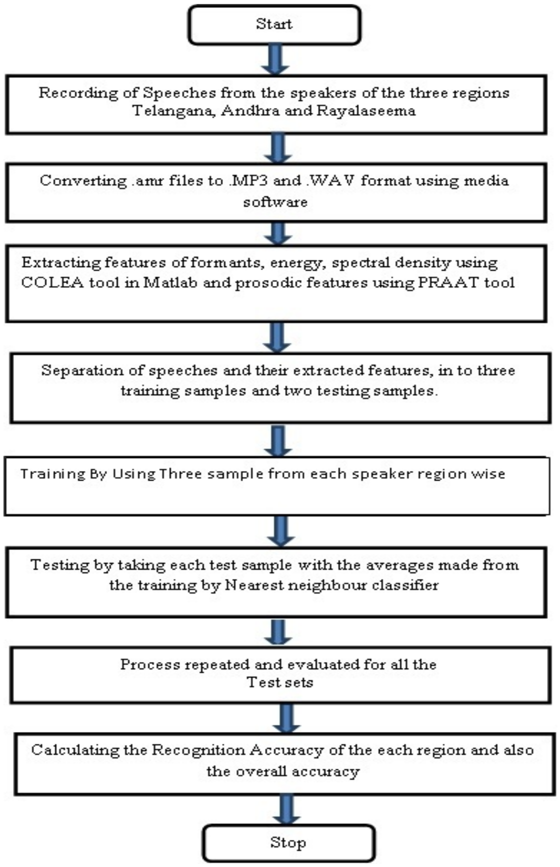
Start
Recording of Speeches from the speakers of the three regions Telangana, Andhra and Rayalaseema
Converting .amr files to .MP3 and .WAV format using media software
Extracting features of formants, energy, spectral density using COLEA tool in Matlab and prosodic features using PRAAT tool
Separation of speeches and their extracted features, in to three training samples and two testing samples.
Training By Using Three sample from each speaker region wise
Testing by taking each test sample with the averages made from the training by Nearest neighbour classifier
Process repeated and evaluated for all the Test sets
Calculating the Recognition Accuracy of the each region and also the overall accuracy
Stop

Figure-7. Step by step algorithm for accent identification.

The second objective of this proposed method is to identify the speaker from the closed dataset.

b) Speaker identification based on text

In case of speaker identification the proposed method is explained in step by step flow chart in Figure-8 as follows. The extracted features of all the speakers from the three regions are stored and from each speaker the first three samples are taken for training purpose and from the corresponding features, the mean of each feature is calculated and stored as a single feature. This procedure is repeated for all the features and for all the speakers of the three regions.

ALGORITHM DEVELOPMENT FOR SPEAKER IDENTIFICATION

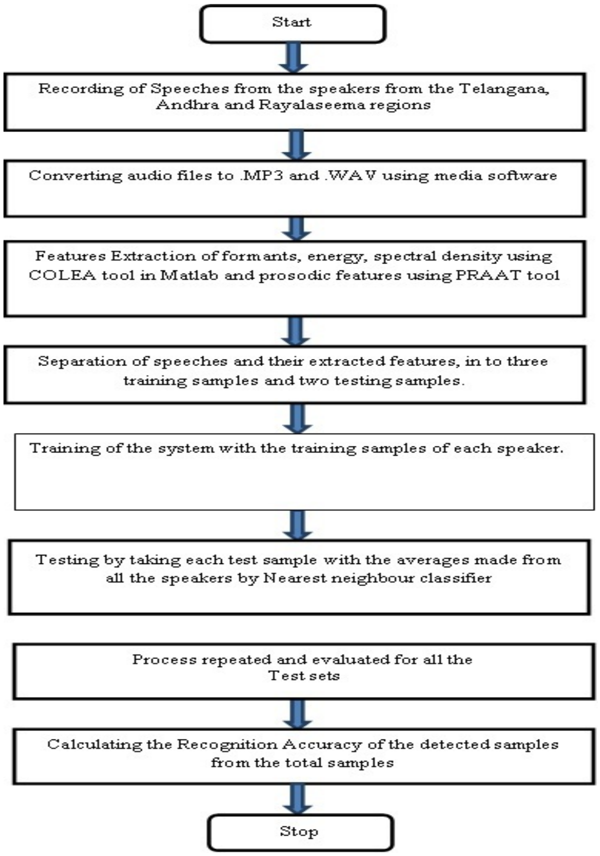
Start
Recording of Speeches from the speakers from the Telangana, Andhra and Rayalaseema regions
Converting audio files to .MP3 and .WAV using media software
Features Extraction of formants, energy, spectral density using COLEA tool in Matlab and prosodic features using PRAAT tool
Separation of speeches and their extracted features, in to three training samples and two testing samples.
Training of the system with the training samples of each speaker.
Testing by taking each test sample with the averages made from all the speakers by Nearest neighbour classifier
Process repeated and evaluated for all the Test sets
Calculating the Recognition Accuracy of the detected samples from the total samples
Stop

Figure-8. Step by step algorithm for speaker identification.

So for each speaker in training the averaged sample value of corresponding features are trained to computer knowledge. So from each region thirteen speakers are taken and from each speaker three samples so the total 195 sample values are in training set used to train computer system / knowledge and the remaining two samples of the extracted features are taken as test samples and the same procedure is repeated in taking the average as explained above and the total 78 test samples are given as input. Now the total thirty nine speakers are given as an input in the proposed algorithm. From each test sample the Euclidean distance is calculated for all the training samples and from it the minimum distance is calculated and gives the speaker identification. This procedure is repeated for remaining test samples and the overall recognition accuracy is calculated.

Mathematical model for the proposed algorithm

For finding the feature vector of a speech, all the features extracted using colea and praat should be added. The feature set of speech of a speaker is

$$F_i = \begin{pmatrix} v_1 \\ v_2 \\ . \\ . \\ v_n \end{pmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.2)$$

Where n is number of features and i is ranging from $1 < i < n$

In the above equation (2), F_i represents a set of feature vectors for any given speech of a speaker.

Similarly the feature sets of all the training / testing samples are found.

$$T_m=[F_1, F_2, F_3 \ldots\ldots\ldots\ldots\ldots\ldots F_n] \qquad (3)$$

Equation (3) is a matrix of speakers.

In the next step, we find the Euclidean distance between test speaker column vector and each of the column vector of Training / database speaker. Euclidean distance can be represented mathematically between 2 vectors equation (4) and (5) as follows:

$$A_i= [A_1, A_2, \ldots\ldots\ldots\ldots\ldots\ldots A_n] \qquad (4)$$
$$B_i= [B_1, B_2, \ldots\ldots\ldots\ldots\ldots\ldots B_n] \qquad (5)$$

Where n = number of speakers
The Euclidean distance equation is given below

$$D =\surd\ (A_1-B_1)^2+ (A_2-B_2)^2+ (A_n-B_n)^2 \qquad (6)$$

From the Equation (6) describes the calculation of Euclidean distance for vectors described by equations (4) and (5). Hence Euclidean distance between test speaker vector and each of the speaker from the data base are computed and put in a row matrix as,

$$D_j= [D_1, D_2, D_3 \ldots\ldots\ldots\ldots D_n] \qquad (7)$$

In the next step we find the minimum value of D in Equation (7) and the corresponding speaker is said to be recognized as the given test speaker. Further we display the test speaker on the screen.

RESULTS AND DISCUSSIONS

Results

In the Table-1 it is very clear twenty test samples out of twenty six test samples were successfully recognized by the algorithm. Therefore the percentage recognition accuracy for the speech from the Coastal Andhra region is (20/26) 77%. Similarly for the Rayalaseema region it is twenty four test samples out of twenty six test samples were successfully recognized by the algorithm. Therefore the percentage recognition accuracy for the speech from the Rayalaseema region is (24/26) 92% and for Telangana region seventeen test samples were successfully recognized out of twenty six test samples. Therefore the percentage recognition accuracy for the speech from Telangana region is (17/26) 65 %.

Table-1. Confusion matrix.

Accent recognition	Coastal Andhra	Rayalaseema	Telangana
Coastal Andhra	20	4	2
Rayalaseema	1	24	1
Telangana	5	4	17

The presentation of confusion matrix is represented in the form of bar graph in Figure-9 as follows, where the x-axis represents the region and the y axis represents the recognition accuracy.

Recognition accuracy

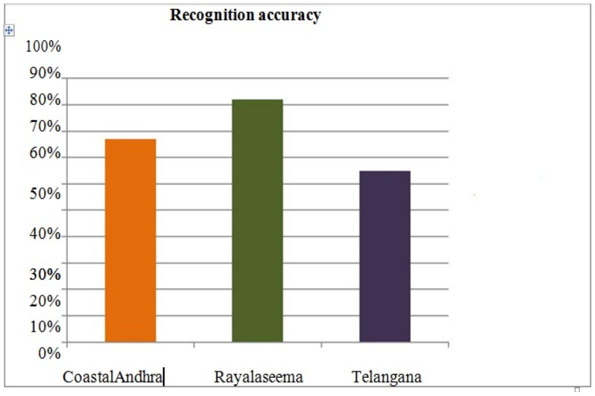

Figure-9. Accent recognition

Comparison with proposed methods

a) Speaker identification

Nitisha and Ashu Bansal [25] worked on Speaker Recognition for Hindi words. They have developed text dependent systems that have been trained for users. They have used MFCC technique for feature extraction and Vector Quantization model for feature vectors modelling. The speech signals were entered into the system using microphone, sound card etc. These were then stored as analog signals (in the form of. wav files) which were further readable by various electronic systems. The analog signals were then digitized to the digital signals. They have used Hindi words for training the system ek (one), do (two), teen (three), char (four)… and the system accuracy was found to be 72 %. [25]

In the proposed method, a text dependent Speaker Identification was developed using Telugu sentence. The sentence was recorded using the mobile phone, HTC. The technique used was the Nearest Neighbourhood Classifier (NNC) for feature matching and extraction. The Euclidian Distance between the training set and the testing sets for each feature was found out. The speaker corresponding to the least Euclidian distance was identified as the absolute speaker. The system has found to possess an accuracy of 74 %.

Table-2. Comparison with published method.

Description	Published method	Proposed method
Speech language	Hindi words	Telugu sentence
Speech recorded	Ek, Do, Teen, Char etc	“ఎవరో అన్నం తిన్నారు నేను ఎవరిని చూడలేదు“ (evaroo annam tinnaaru nenu evarini chuudaledu)
Input system	Microphone	Mobile Phone
Type of identification	Text dependent	Text Dependent
Features selected	Mel Frequency Cepstrum Coefficient (MFCC)	Formants, Energy and Prosodic features.
Classifier	Vector Quantization (VQ)	Nearest Neighbourhood Classifier (NNC)
Speech signal type	First analog and then digitized to digital signals.	Digitized signal
Percentage accuracy of the speakers	72 %	74 %

b) Accent recognition

In the published method k sreenivasa rao and shashidhar [26] used to detect the Accent of the Hindi language from the five prominent regions of India. They are Chattisgharhi, Bengali, Marathi, General, and Telugu. The work focussed on text-dependent where they the continuous speech of the user. From each region they recorded five male and five female summing to a total of 50 users and the time used to record is ranging from five to ten minutes. The features they selected are MFCC (Mel Frequency Spectral Coefficients), Prosodic features and Energy contours and used to detect by using AANN (Auto Associative Neural Network) and SVM (Support Vector Machine) and found out recognition accuracy 81% for the above user's voice collected.

In the proposed method we used to detect the Accent of the Telugu language from the three prominent regions of Andhra Pradesh. The work focused on text-dependent where the sentence is selected from the nativity of Telugu Language. From each region we recorded ten male and three female summing to a total of 39 speakers and the time used to record is for five seconds. The features we selected are Formants, Prosodic features, Energy and Intensity. With these features we used to detect by using NNC (Nearest Neighbourhood Classifier) and found out recognition accuracy 78 % for the above users voice collected.

OVERALL CONCLUSIONS AND FUTURE SCOPE

CONCLUSIONS

(i) A database / knowledge sets were generated for 39 speakers with 13 speakers from each Telugu speaking region i.e. Coastal Andhra, Rayalaseema and Telangana successfully.
(ii) Various features like Pitch, formants, energy and power spectral density for all the speeches were successfully extracted.
(iii) Prosodic features like minimum, maximum, range mean absolute slope and standard deviation for all the speeches were also extracted.
(iv) The Accent based recognition accuracy for identifying the region (Costal Andhra, Rayalaseema, and Telangana) is found 78 % with the proposed algorithm.
(v) The recognition accuracy obtained is 74% for speaker identification with the proposed algorithm using Nearest Neighbourhood Classifier.

4.2 FUTURE SCOPE

a) Detecting more features from the speech signal can help in improving the accuracy for Accent recognition and speaker Identification.
b) The recognition accuracy can be improved by increasing the number of training samples
c) This work focussed on Accent Recognition and Speaker Identification by using the Text-dependent data. This can be extended to Text-independent speech in order make the system more robust.

REFERENCES

[1] Homayoon Beigi. 2011. Fundamentals of Speaker Recognition. Springer, New York. ISBN:978-0-387-77591-3.

[2] Lawerence and R. Rabiner. Applications of speech recognition in the area of Telecommunications.

[3] D.A. Reynolds, L.P. Heck. 2000. Automatic Speaker Recognition. AAAS 2000 Meeting, Humans, Computers and Speech Symposium. p. 19.

[4] Nengheng. 2005. Biometric Identification and classification based on acoustic waves. Science of Forensic phonetics, springer, Heidelberg.

[5] K. H. Davis, *et al*. 1952. Automatic recognition of spoken digits. J.A.S.A. 24(6): 637-642.

[6] Sakoe and Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE, Trans. Acoustics, speech and signal Proc Vol. ASSP-26.

[7] C.S. Myers, L.R. Rabiner. On the use of dynamic time warping for word spotting and connected word recognition. Bell system technical journal.

[8] An HMM based approach for off-line Unstrained Handwritten Word modeling and recognition. IEEE on Pattern analysis and machine intelligence.

[9] S. Pruzansky. Pattern-matching procedure for automatic talker recognition. J.A.S.A. p. 35.

[10] Pruzansky. S, Mathews. M.V. Talker recognition procedure based on analysis of variance. J.A.S.A. 36: 2041-2047.

[11] Li. *et al*. Improvement in filter banks with correlating digital spectrograms. Vol. 44 E.C.T.I

[12] Enders and Furui. Intra speaker variability in Speaker recognition. ECTI transactions on computer and information technology.

[13] P. D. Bricker, *et al*. 1971. Statistical techniques for talker identification, B.S.T.J., 50, pp. 1427-1454.

[14] K. P. Li and G. W. Hughes. 1974. Talker differences as they appear in correlation matrices of continuous speech spectra, J.A.S.A. 55, pp. 833-837.

[15] B. Beek, *et al*. 1977. An assessment of the technology of automatic speech recognition for military applications. IEEE Trans. Acoustics Speech and Signal Processing, ASSP-25, pp. 310-322.

[16] M. R. Sambur. 1972. Speaker recognition and verification using linear prediction analysis. Ph. D. Dissert, M.I.T.

[17] P. Mermelstein and S. Davis. 1980. Comparison of parametric representation for mono syllabic word recognition in continuously spoken sentences. In IEEE Transactions on Acoustic Speech and Signal Processing. 28(4): 357-366.

[18] F. K. Soong, *et al*. 1987. A vector quantization approach to speaker recognition. AT and T Technical Journal. 66: 14-26.

[19] N. Tishby. 1991. On the application of mixture AR Hidden Markov Models to text independent speaker recognition. IEEE Trans. Acoustic Speech and Signal Processing, ASSP. 30(3): 563-570.

[20] A. B. Poritz. 1982. Linear Predictive Hidden Markov Models and the speech signal. Proc. ICASSP. 2: 1291-1294.

[21] R. Rose and R. A. Reynolds. 1990. Text independent speaker identification using automatic acoustic segmentation. Proc. ICASSP. pp. 293-296.

[22] T. Matsui and S. Furui. 1992. Comparison of text independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. Proc. ICSLP. pp. II-157-160.

[23] M. Kasiprasad, P. Narahari Sastry, V. Rajesh. 2013. Analysis and design of Speaker Identification System using NNC. Icacm, Elsevier digital edition.

[24] Kazuhiro Nakadai, Ken-ichiidai. 2001. Real time multiple speakers tracking by multi-modal integration for mobile robots. Euro speech.

[25] Nitisha and Ashu Bansal. Speaker Recognition Using MFCC Front End Analysis and VQ Modelling Technique for Hindi Words using MATLAB. Hindu College of Engineering, Haryana, India.

[26] K. Sreenivasa Rao and Shashidhar. G Koolagudi. 2011. Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech. Systems, Cybernetics and Informatics. 9(4). ISSN: 1690-4524.