ACCENT DETECTION OF TELUGU SPEECH USING PROSODIC AND FORMANT FEATURES

Kasiprasad.Mannepalli[1],P.Nrahari Sastry[2],V.Rajesh[3]
[1] Research Scholar, Dept of E.C.E, K L University, Vaddeswaram, Vijayawada, Guntur (Dist), Andhra Pradesh, India.
[2]Associate Professor, Dept of E.C.E, CBIT, Hyderabad, India.
[3]Professors, Dept of E.C.E, K L University, Vaddeswaram, Vijayawada, Guntur (Dist), Andhra Pradesh, India

***ABSTRACT*—Speech automation is becoming more popular in the recent times. Speech recognition systems are increasing day by day. Earlier the speech recognition systems were developed for English language. Now these systems are being developed for many other languages. Many languages in the globe have different speaking styles or accents. The speech recognition systems may not recognize speeches with accent other than the system are trained. So it is important in the speech to text conversion systems to convert the accented speech in to text. Telugu is a language of southern part of India, has mainly three different accents namely Coastal Andhra, Rayalaseema and Telangana, in which the stress is different for the same word in these accents. In this work, text dependent speeches from Coastal Andhra, Rayalaseema, Telangana accents have been collected. Prosodic and formant features extracted from speech are used for discriminating the accents. Prosodic features are represented by durations of syllables, pitch and energy contours .These features are used to recognize the accent of the speaker using Nearest Neighborhood Classifier. The best recognition Accuracy using this model obtained 72%.**

Keywords—Speech accents, Telugu speech, Prosodic features.

I. INTRODUCTION

A language may be pronounced in different ways in different regions of that language speaking areas. The different ways of pronouncing a word in any language is known as accent. There is variation in pronouncing English language among Americans, British and Australians. Telugu is largely spoken language in southern parts of India. The grammar, stress, intonation and prosody vary from language to language. In the recent times the research on speech processing of Indian languages is increasing. Telugu language mainly has three namely Coastal Andhra (accent of Coastal Andhra region of Andhra Pradesh), Rayalaseema, Telangana. Recognition of dialects or accents of speakers prior to automatic speech recognition (ASR) helps in improving performance of the ASR systems by adapting the ASR acoustic and/or language models appropriately[1]

The dialect or Accent specific Information is present in the speech signal at different levels. At the segmental level, the accent or dialect specific information can be observed in the form of unique sequence of the shapes of the vocal tract for producing the sound units. At the supra-segmental level, the

Dialect specific knowledge is embedded in the duration patterns of the syllable sequences and the dynamics of the pitch and energy contours [2]. At the sub-segmental level, the dialect specific information may present in the shape of the glottal pulse and durations of open and close phases of Vocal folds segmental features are extracted by analyzing the speech segments of duration 20-30ms [2]. Supra-segmental features also known as prosodic features extracted from the speech segments of duration greater than 100ms. Sub-segmental features are extracted from the speech segments of duration less than 3ms [2].

Periodic information changes from language to language. These features are the rhythmic and intonation properties in speech. The examples are voice fundamental frequency (F0), F0 gradient, intensity, energy and duration etc. They are relatively simple in structures, and are believed to be effective in some speech recognition tasks [3]. There were studies on Automatic dialect or accent identification of languages of western countries and few studies are there in identification of dialects of Hindi language of India.

II. LITERATURE SURVEY

Arlo Faria in the work "Accent Classification for speech recognition" described the classification of speech from native and non-native speakers, enabling accent-dependent automatic speech recognition [4]. In addition to the acoustic signal, lexical features from transcripts of the speech data can also provide significant evidence of a speaker's accent type. Subsets of the Fisher corpus, ranging over diverse accents, were used for these experiments.

Bin MA, Donglai ZHU and Rong TONG, presented a method to extract tone relevant features based on pitch flux from continuous speech signal. The autocorrelations of two adjacent frames are calculated and the covariance between them is estimated to extract multi-dimensional pitch flux features [5].

D. Ververidis and C. Kotropoulos, in their work "A state of the art review on emotional speech databases," have used short time supra-segmental features and their statistics for analyzing the emotions [6]. Some of the prosodic features used by them include: pitch frequency *F*0, energy, formant locations and bandwidths, speaking rate and transition time, dynamics of pitch, energy and formant contours.

Fadi Biadsy, Julia Hirschberg, have examined the role of prosodic features like intonation and rhythm across four different Arabic dialects such as Gulf, Iraqi, Levantine, and Egyptian, for the purpose of automatic dialect identification [7].

Gang Liu and John L. Hansen in the work "A systematic strategy for robust automatic dialect identification" investigated a series of strategies to address the question of small and noisy dataset dialect classification task [8].

Kasiprasad.M, P.Narahari Sastry, V.Rajesh have used formant features, pitch and energy to identify the speaker in the work "Analysis and Design of speaker identification system using NNC" and obtained an efficiency of 78% [9].

Qin Yan, Saeed Vaseghi presented a comparative study of the acoustic speech features of two major English accents American English and British English. Detailed study of the acoustic correlates of accent using intonation patterns and pitch characteristics was performed in their work "A Comparative Analysis of UK and USA English accents in recognition And Synthesis" [10].

S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen,M. Westerdijk, and S. Stroeve in their work, "Approaching automatic recognition of emotion from voice have used the peaks and troughs in the profile of fundamental frequency and durations of pauses, intensity and bursts for identifying the four emotions namely fear, anger, sadness and joy. They have reported the classification performance of 55% using discriminant analysis [11].

Santosh Gaikwad, Bharti Gawali and K V Kale in their study "Accent Recognition for Indian English using Acoustic Feature Approach" present an experimental approach of acoustic speech feature for Marathi & Arabic accents for English speaking. The detail study of acoustics correlates the accent using formant frequency, energy and pitch characteristics [12].

III. Features used

i. Pitch:

Pitch is the fundamental frequency of the speech signal which differs person to person. Pitch can be calculated by using auto correlation or cepstrum method. In this work the pitch calculated by using correlation method. For a long sentence the value of pitch varies time to time the mean of these pitches is taken. The minimum pitch, the maximum pitch of the sentence and the range at which the pitch varies are also taken as the other features that can extracted from the pitch.

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 can be seen in the spectrum of certain sounds mainly vowels. Speech organs produce vowel sounds by forming certain shape. The regions of resonance and anti-resonance are formed in the vocal tract. Location of these resonances in the frequency spectrum depends on 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. Prosodic Features

Prosodic features are supra segmental in nature. 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 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. In the recent times, 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

iv. Power spectral Density:

It is a function of how the variation in a signal is caused by different frequency components. Thus it is strictly a variance density function that describes how the signal energy or power is distributed across frequency.

v. Short time Energy:

The energy associated with speech is time varying in nature as the speech signal consist of voiced, unvoiced and silence regions. Splitting the signal into frames is done by multiplying the signal by a suitable window $W(n)$, n=0,1,2, …,N-1, which is zero for n beyond the range (0, N-1). The energy of voiced speech is generally greater than that of unvoiced speech though there are occasions when the energy of strong fricatives is greater than that of weak vowels.

vi. Intensity:

Intensity is the Acoustic or Sound Power (W) per unit area. The SI-units for Sound Intensity are W/m^2.

IV. METHODOLOGY

A Telugu sentence "EVARO ANNAM THINNARU. NENU EVARINI CHUDALEDHU" was recorded from 10 speakers each of three regions each speaker speaks that sentence for five times.

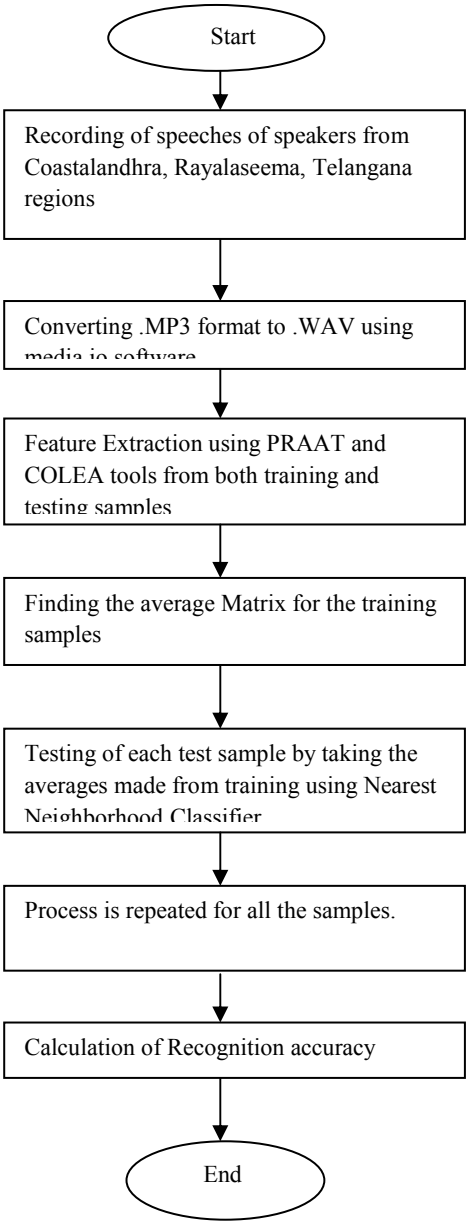

Fig 1: Flow of the Accent recognition system

Praat Tool is used to extract the Minimum Pitch, maximum Pitch, mean pitch, standard deviation, Pitch Range, mean absolute slope, Intensity in the speech. Colea tool is used to extract the Power spectral Density and the Energy .for all the recorded speeches the features have been extracted and tabulated.

The average values for each parameter were calculated from training samples of each accent and thus the average matrix was formed. Each column in the average matrix represents one of the three accents in Telugu. The average values are then compared with the test samples by finding the Euclidian distances and the least distances were obtained in other matrix. The accent corresponding to the least Euclidian distance was identified as the absolute accent. The results so formed were entered in the result sheet and thereafter the efficiency is calculated.

V. Feature Extraction

The features extracted from the COLEA and PRAAT are shown in the following figures.

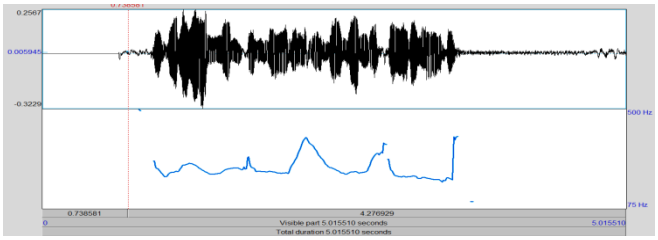

Fig 2: Speech signal and its pitch contour of sample1 of CA training

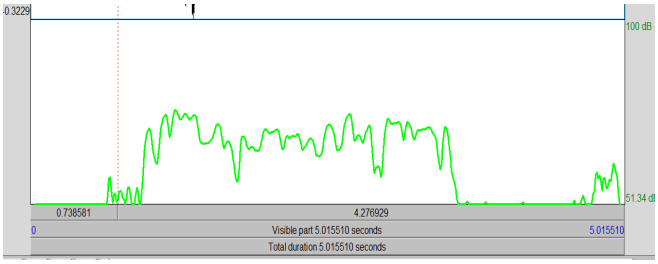

Fig 3: Intensity of sample1 of CA training sample

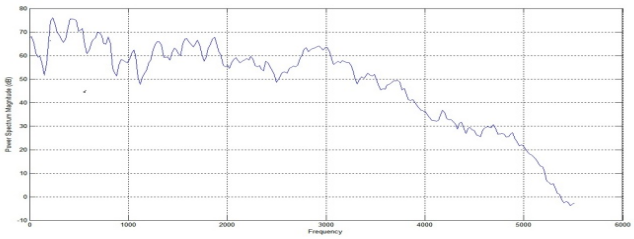

Fig 4:Power spetral Density of sample1 of CA training sample

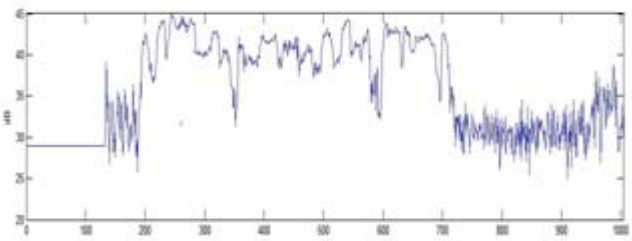

Fig5: Energy contour of sample1 of CA training samples

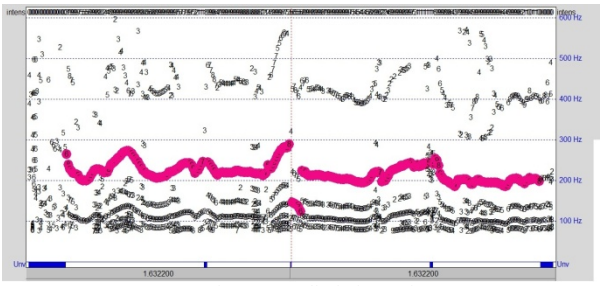

Fig 6:Prosodic information

Fig 5 shows the energy plot of speech signal form Coastal Andhra region, the energy of the signal is calculated from the plot. Fig 6 shows prosodic information. The procedure is repeated for all the speech signals. The training averages are shown in table 1.

TABLE 1.TRAINING AVERAGES OF PARAMETERS

Accent / Parameter	CA	RS	TG
Minimum Pitch (HZ)	100	115	130
Maximum Pitch(HZ)	299	309	379
Pitch Range(HZ)	199	195	249
Average Pitch (HZ)	152	155	185
Std deviation of Pitch (HZ)	31	31	39
Mean absolute slope of Pitch (HZ)	331	313	396
Energy	42	38	37
PSD	54	44	52
Intensity	79	71	73
F1(Hz)	503	441	431
F2(Hz)	1312	1235	1101
F3(Hz)	2144	2105	1908

In the Table 1 CA represents Coastal Andhra accent, RS represent Rayalaseema accent, TG represents Telangana accent PSD power spectral density. The training averages are computed by taking average of 30 samples from each region.

VI. Results and Discussion

The testing is done by taking 20 samples from each region as test samples. Out 20 speech signals of Coastal Andhra region 14 are recognized as Coastal Andhra accent, 3 are recognized as Rayalaseema, 3 are recognized as Telangana accent. Out 20 speech signals of Rayalaseema region 2 is recognized as Coastal Andhra accent, 16 are recognized as Rayalaseema,4 are recognized as Telangana accent. Out 20 speech signals of Telangana region 3 are recognized as Coastal Andhra accent, 3 are recognized as Rayalaseema, 15 are recognized as Telangana accent. The overall efficiency is 72% and the same is shown in Table 2.

TABLE 2: RESULT(IN PERCENTAGE RECOGNITION ACCURACY)

	CA	RS	TG	
CA	75	15	15	
RS	10	70	20	
TG	15	15	70	
	Overall Efficiency in %			72

In the published method k .sreenivasa rao and shashidhar used to detect the Accent of the Hindi language from the five prominent regions of India. They are Chhattisgarhi, Bengali, Marathi, General, and Telugu. 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).

TABLE 3: COMPARISON WITH PUBLISHED METHOD

Description	**Published Method**	**Proposed Method**
Speech Language	Hindi words	Telugu sentence
Speech recorded	Spontaneous responses of a question	"ఎవరో అన్నంతిన్నారు నేను ఎవరిని చూడలేదు" (evaroo annam tinnaaru nenu evarini chuudaledu)
Type of Identification	Question dependent	Text Dependent
Features Selected	Mel Frequency Cepstrum Coefficient(MFCC)	Formants, Energy and Prosodic features
Classifier	AANN, SVM	Nearest Neighbourhood Classifer(NNC)

Percentage Accuracy	Spectral&AANN-70%. Spectral,Prosodic with AANN-80%.	72 %

VII.Conclusions:

1. In this work speeches from Coastal Andhra, Rayalaseema, Telangana regions have been collected. A database with these speeches has been prepared.
2. Prosodic features and formant features have been extracted for both training and testing samples.
3. The average has been found for each parameter from each region for the training samples
4. The classification is done by NNC and the Overall Efficiency obtained is 72%

.

REFERENCES

[1] Mingkuan Liu, Bo Xu, Taiyi Hunng, Yonggang Deng, and Chengrong Li , 2000. "Mandarin accent adaptation based on context independent/Context-dependent pronunciation modeling". In Proceedings of the Acoustics, Speech, and Signal Processing, ICASSP '00, pages: II1025-II1028, Washington, DC, USA. IEEE Computer Society.

[2] 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 Volume 9 - Number 4 .ISSN: 1690-4524.

[3] Nilu Singh, R.A Khan, Raj Shree "MFCC and Prosodic Feature Extraction Techniques:A Comparative Study"

International Journal of Computer Applications (0975 – 8887) Volume 54– No.1, September 2012

[4] Arlo Faria, 2005. "Accent Classification for Speech Recognition" In proceeding of: Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13.

[4] Bin MA, Donglai ZHU and Rong Tong, 2006. "Chinese Dialect Identification Using Tone Features based on pitch flux" ICASSP, pages I1029 –I1032.

[6] D. Ververidis and C. Kotropoulos, 2006. "A state of the art review on emotional speech databases," in Eleventh Australasian International Conference on Speech Science and Technology, Auckland, New Zealand.

[7] Fadi Biadsy, Julia Hirschberg, 2009. "Using Prosody and Phonotactics in Arabic Dialect Identification" interspeech09.

[8] Gang Liu and John L. Hansen, 2011. A systematic strategy for robust automatic dialect identification. In EUSIPCO, pages 2138-2141

[9] Kasiprasad.M, P.Narahari Sastry, V.Rajesh, 2013. "Analysis and Design of speaker identification system using NNC" ICACM-2013, Elsevier Digital Edition. Pages: 381-387.

[10] Qin Yan, Saeed Vaseghi, 2000. "A Comparative Analysis of UK and US English accents in recognition And Synthesis" 0-7803-7402-9/02©2002 IEEE

[11] S.McGilloway, R.Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, 2000. "Approaching automatic recognition of emotion from voice: A rough benchmark," (Belfast), 2000.

[12] Santosh Gaikwad, Bharti Gawali and K V Kale, 2013. Article: Accent Recognition for Indian English using Acoustic Feature Approach. International Journal of Computer Applications 63(7):25-32, February . Published by Foundation of Computer Science, New York, USA.