International Journal of Soft Computing 10 (5): 287-292, 2015
ISSN: 1816-9503

Accent Detection of Telugu Speech Using Supra-Segmental Features

[1]Kasiprasad Mannepalli, [2]P. Narahari Sastry and [3]V. Rajesh
[1]Department of ECE, KL University, Vaddeswaram,
Vijayawada, Guntur (Dist.) Andhra Pradesh, India
[2]Department of ECE, CBIT, Hyderabad, India
[3]Department of ECE, KL University (Koneru Lakshmamaiah Educational Foundation),
Vaddeswaram, Vijayawada, Guntur (Dist.), India

Abstract: Speech recognition systems are used in many applications, it is crucial for the speech recognition systems to be able to deal with accented speakers. The speech recognition systems have to model the speech variances among different speakers such as dialects or accents of the spoken language, the speaker's gender to identify what was spoken by the speaker. It is more important in the speech to text conversion systems to convert the accented speech in to text. Telugu language 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 the present research, text dependent speeches from Coastal Andhra, Rayalaseema, Telangana accents have been recorded. The features like pitch, power spectral density, energy and intensity have been extracted and are used to recognize the accent using nearest neighborhood classifier. The best recognition accuracy using this model obtained as 73.3%.

Key words: Accent detection, supra-segmental features, Telugu speech, prosodic features, speakers

INTRODUCTION

The different ways of pronouncing a word in any language is known as accent. There is wide variation in pronouncing English language between Americans and British. Similarly, there are three main accents in Telugu namely Coastal Andhra (accent of Coastal Andhra Region of Andhra Pradesh), Rayalaseema, Telangana. There have been significant attempts to automatically recognize the accent of a speaker given his or her speech utterance. 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 (Liu *et al*., 2000).

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 (Rao and Koolagudi, 2011). 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 20-30 msec (Rao and Koolagudi, 2011). Supra-segmental features also known as prosodic features extracted from the speech segments of duration >100 msec. Sub-segmental features are extracted from the speech segments of duration <3 msec (Rao and Koolagudi, 2011). There were studies on automatic dialect or accent identification of languages of Western countries and few studies are there in identification of Hindi language of India. There are many languages in India where there is a much scope for identifying accents, emotion, etc. Telugu is one such language in India. Large section people from Southern part of India speaks Telugu with different regional accents.

Literature review: Faria (2005) in the research "accent classification for speech recognition" describes classification of speech from native and non-native speakers, enabling accent-dependent automatic speech recognition. In addition to the acoustic signal, lexical features from transcripts of the speech data can also

Corresponding Author: Kasiprasad Mannepalli, Department of ECE, KL University, Vaddeswaram, Vijayawada, Guntur (Dist.), Andhra Pradesh, India

provide significant evidence of a speaker's accent type. Subsets of the Fisher corpus, ranging over diverse accents were used for these experiments.

Ma *et al*. (2006) 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.

Ververidis and Kotropoulos (2006) in their research "A State of the Art Review on Emotional Speech Databases" have used short time supra-segmental features and their statistics for analyzing the emotions. Some of the prosodic features used by them include: pitch frequency F0, energy, formant locations and their bandwidths, dynamics of pitch, energy and formant contours, speaking rate and transition time.

Biadsy and Hirschberg (2009) examined the role of prosodic features (intonation and rhythm) across four Arabic dialects: Gulf, Iraqi, Levantine and Egyptian, for the purpose of automatic dialect identification.

Liu *et al*. (2010) in the research dialect identifcation: impact of differences between read versus spontaneous speech dialects of a language normally are reflected in terms of their phoneme space, word pronunciation or selection and prosodic traits. These traits are clearly visible in natural speaker to speaker spontaneous conversations.

Liu and Hansen (2011) in the research "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.

Kasiprasad *et al*. (2013) have used formant features, pitch and energy to identify the speaker in the research "Analysis and Design of Speaker Identification System using NNC" and obtained an efficiency of 78%.

Rao and Koolagudi (2011) have explored speech features to identify Hindi dialects and emotions in their research "Identification of Hindi Dialects and Emotions Using Spectral and Prosodic Features of Speech".

Liu *et al*. (2000) in their study "Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling" proposed an accent adaptation approach using pronunciation variation modeling technology for the mandarin accent.

"Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB" was proposed by Bansal (2012). This study introduced text dependent systems that have been trained for a particular user.

Yan and Vaseghi (2002) presented a comparative study of the acoustic speech features of two major English accents: British English and American English. Detailed study of the acoustic correlates of accent using intonation pattern and pitch characteristics was performed in their research "A Comparative Analysis of UK and USA English Accents in Recognition and Synthesis".

McGilloway *et al*. (2000) in their research, "Approaching Automatic Recognition of Emotion from Voice Have Used the Peaks and Troughs in the Profile of Fundamental Frequency and Intensity, Durations of Pauses and Bursts for Identifying the Four Emotions Namely Fear, Anger, Sadness and Joy". They have reported the classification performance of 55% using discriminant analysis.

Gaikwad *et al*. (2013) in their study "Accent Recognition for Indian English using Acoustic Feature Approach" present an experimental approach of acoustic speech feature for Marathi and Arabic accents for English speaking. The detail study of acoustics correlates the accent using formant frequency, energy and pitch characteristics.

Features used

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 research, 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.

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.

Short time energy: The energy associated with speech is time varying in nature because the speech signal consist of voiced, unvoiced and silence regions. Splitting the signal into frames can be achieved 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.

Intensity: Sound intensity is the acoustic or sound power (W) per unit area. The SI-units for sound intensity are W/m^2.

MATERIALS AND METHODS

The 30 people were identified 10 from each region of Andhra Pradesh. A sentence "EVARO ANNAM THINNARU. NENU EVARINI CHUDALEDHU" (A Telugu sentence which means somebody ate food but I don't see anybody) each speaker speaks that sentence for five time and the speeches are recorded. The input speech signals that are fed to the system should be in. WAV format, since 'wavread' function in MATLAB accepts only. WAV files. The MP_3 files are converted to WAV using Media Converter Software.

Pratt 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 (Fig. 1).

Feature extraction: Figure 2 shows speech signal and its pitch contour of a speech signal form Coastal Andhra Region. The minimum, maximum, average value of pitch, standard deviation and mean absolute slope of the pitch are calculated from the plot. Figure 3 shows the intensity plot of speech signal Coastal Andhra Region. The average intensity of the speech signal is calculated from the plot. Figure 4 shows the power spectral density of the speech signal form Coastal Andhra Region. The average value of the power spectral density is taken from the plot. Figure 5 shows the energy plot of speech signal form

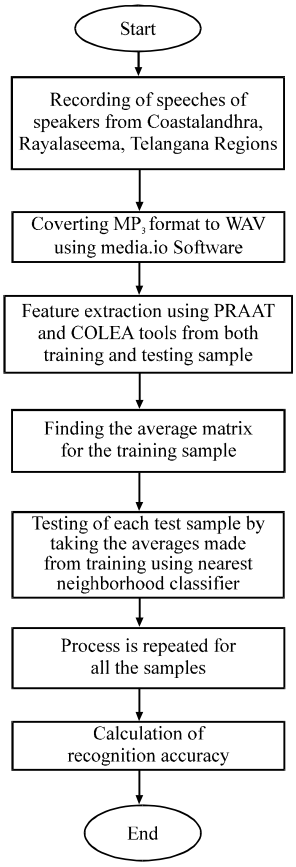

Fig. 1: Flow of the accent recognition system

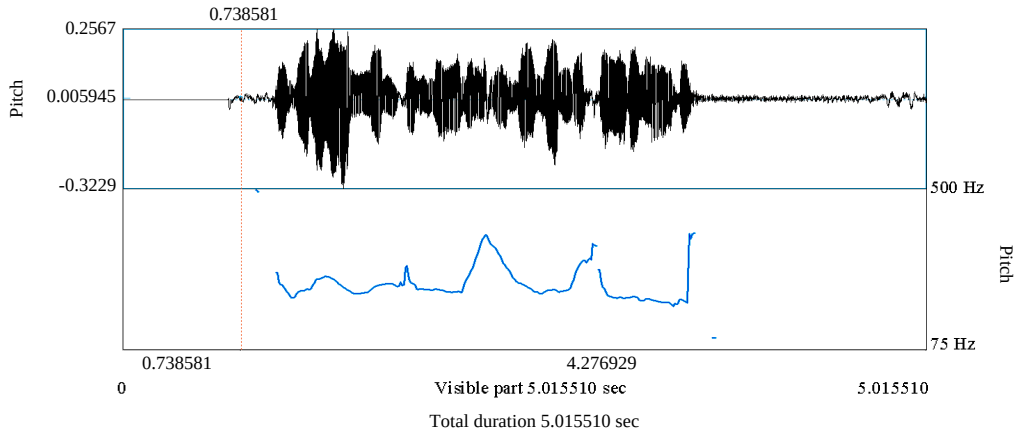

Fig. 2: Speech signal and its pitch contour of sample1 of CA training

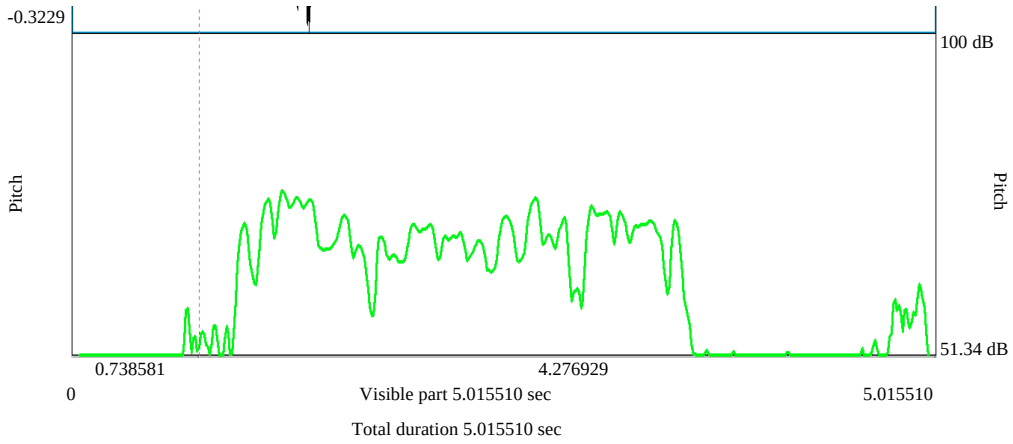

Fig. 3: Intensity of sample1 of CA training sample

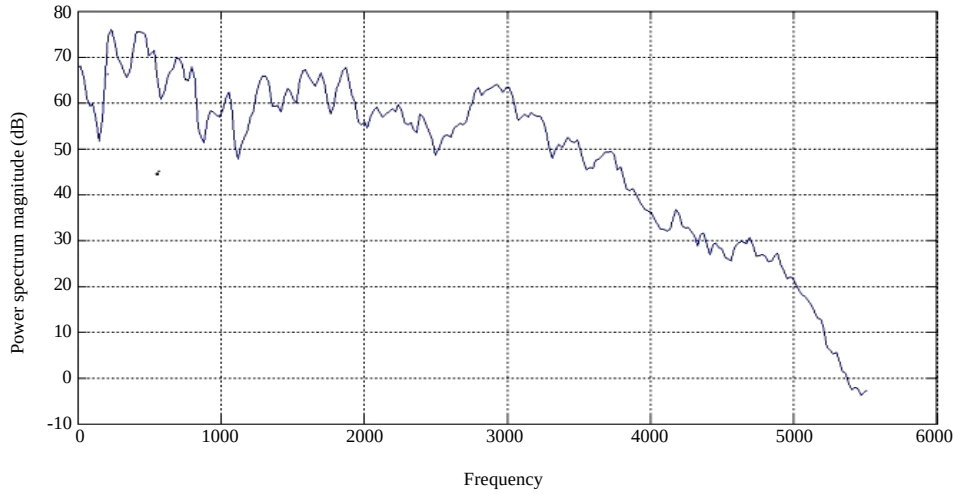

Fig. 4: Power spetral density of sample1 of CA training sample

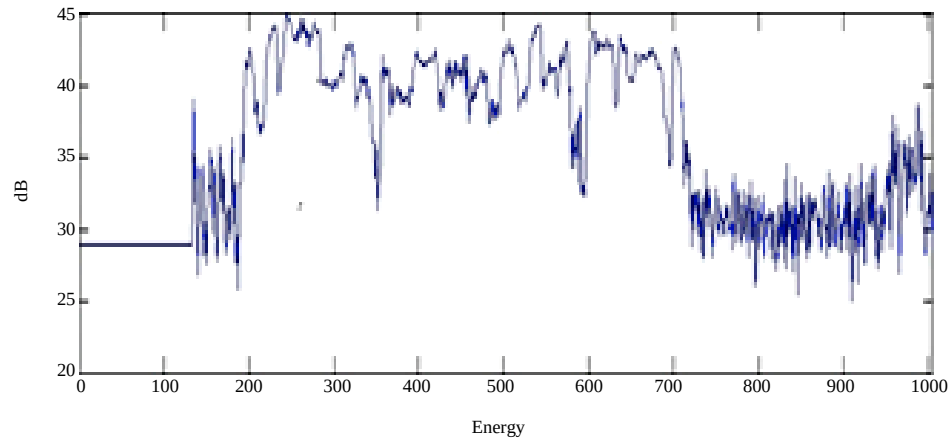

Fig. 5: Energy contour of sample1 of CA training samples

Coastal Andhra Region, the energy of the signal is calculated from the plot. The procedure is repeated for all the speech signals.

Three samples from each speaker of each region have been used for training and two samples from each speaker are used as test samples. The notations used hereafter in this study are given as:

- CA: Coastal Andhra
- RS: Rayalaseema
- TG: Telangana
- PSD: Power Spectral Density
- Std. Dev.: Standard Deviation

RESULTS AND DISCUSSION

This research is carried out in two cases. In each case, the number of parameters used is different. In case 1, minimum pitch (Hz), maximum pitch (Hz), pitch range (Hz), pitch average (Hz), Std. deviation (Hz), mean absolute slope (Hz), energy, power spectral density, intensity used. The male and female speakers are taken together for training and for testing also. The training parameters are shown in Table 1 for each region speakers.

The testing is done by taking 20 samples from each region as test samples. In case 1 out 20 speech signals of Coastal Andhra Region 11 are recognized as Coastal Andhra accent, 3 are recognized as Rayalaseema, 6 are recognized as Telangana accent. Out 20 speech signals of Rayalaseema Region 1 is recognized as Coastal Andhra accent, 14 are recognized as Rayalaseema, 5 are recognized as Telangana accent. Out 20 speech signals of Telangana Region 4 are recognized as Coastal Andhra accent, 4 are recognized as Rayalaseema, 12 are recognized as Telangana accent. The overall efficiency is 62% and the same is shown in Table 2.

In case 2 minimum pitch, mean pitch, SD of pitch, power spectral density, intensity and energy have been used and the training averages have been taken for male and female speakers separately. Table 3 shows the averages of the parameters used in case 2 for training. In Table 3, M indicates male and F indicates female.

The testing is done by taking 20 samples from each region as test samples. In case 1, the pitch is used as threshold for differentiating male speaker and female speaker and if it is male speaker the test sample features compared with male speaker averages of training samples and vice-versa.

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 3 is recognized as Coastal Andhra accent, 15 are recognized as Rayalaseema, 2 are recognized as Telangana accent. Out 20 speech signals of Telangana Region 3 are recognized as Coastal Andhra accent, 2 are recognized as Rayalaseema, 15 are recognized as Telangana accent. The overall efficiency is 73.3% and the same is shown in Table 4.

Table 1: Training averages of parameters used in case 1

	Accent		
Parameters	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

Table 2: Result in case 1 (all the values are in %)

Accents	CA	RS	TG	Overall efficiency (%)
CA	55	15	30	62
RS	5	70	25	-
TG	20	20	60	-

Table 3: Training averages of parameters used in case 2

	CA		RS		TG	
Parameters	M	F	M	F	M	F
Min. pitch (Hz)	93	168	107	182	103	169
Avg. pitch (Hz)	141	250	146	232	142	248
Std.dev. of pitch (Hz)	29	54	29	53	38	41
Energy	42	36	38	38	36	38
PSD	54	49	44	46	51	52
Intensity	81	67	71	67	73	73

Table 4: Result case 2 (all the values are in %)

Accents	CA	RS	TG	Overall efficiency (%)
CA	70	15	15	73.3
RS	15	75	10	-
TG	15	10	75	-

CONCLUSION

- In this research speeches from Coastal Andhra, Rayalaseema, Telangana regions have been collected
- 9 features have been extracted for both training and testing samples
- All the nine features are used in case 1 and average has been found for each parameter from each region for the training samples
- The overall efficiency in case 1 is 62%
- The 6 features are used in case 2 and training and testing sets are used separately for male and female
- The overall efficiency in case 2 is 73.3%

REFERENCES

Bansal, A., 2012. Speaker recognition using MFCC front end analysis and VQ modeling technique for Hindi words using MATLAB. Int. J. Comput. Applic., 45: 48-52.

Biadsy, F. and J.B. Hirschberg, 2009. Using prosody and phonotactics in Arabic dialect identification. http://academiccommons.columbia.edu/catalog/ac:159968.

Faria, A., 2005. Accent classification for speech recognition. Proceedings of the 2nd International Workshop on Machine Learning for Multimodal Interaction, July 11-13, 2005, Edinburgh, UK., pp: 285-293.

Gaikwad, S., B. Gawali and K.V. Kale, 2013. Accent recognition for Indian English using acoustic feature approach. Int. J. Comput. Applic., 63: 25-32.

Kasiprasad, M., P.N. Sastry and V. Rajesh, 2013. Analysis and design of speaker identification system using NNC. Proceedings of the 2nd International Conference on Applied and Computational Mathematics, May 14-16, 2013, Athens, Greece, pp: 381-387.

Liu, G. and J.H. Hansen, 2011. A systematic strategy for robust automatic dialect identification. Proceedings of the 19th European Signal Processing Conference, August 29-September 2, 2011, Barcelona, Spain, pp: 2138-2141.

Liu, G., Y. Lei and J.H. Hansen, 2010. Dialect identification: Impact of differences between read versus spontaneous speech. Proceedings of the 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark, pp: 2003-2006.

Liu, M., B. Xu, T. Hunng, Y. Deng and C. Li, 2000. Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, June 5-9, 2000, Istanbul, Turkey, pp: 1025-1028.

Ma, B., D. Zhu and R. Tong, 2006. Chinese dialect identification using tone features based on pitch flux. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, May 14-19, 2006, Toulouse, pp: 1.

McGilloway, S., R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve, 2000. Approaching automatic recognition of emotion from voice: A rough benchmark. Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion, September 5-7, 2000, Newcastle, UK., pp: 207-212.

Rao, K.S. and S.G. Koolagudi, 2011. Identification of Hindi dialects and emotions using spectral and prosodic features of speech. Int. J. Syst. Cybernetics Inform., 9: 24-33.

Ververidis, D. and C. Kotropoulos, 2006. A state of the art review on emotional speech databases. Proceedings of the 11th Australasian International Conference on Speech Science and Technology, December 6-8, 2006, Auckland, New Zealand.

Yan, Q. and S. Vaseghi, 2002. A comparative analysis of UK and US English accents in recognition and synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, May 13-17, 2002, Orlando, FL., USA., pp: I-4103-1-416.