International Journal of Engineering & Technology, 7 (2.7) (2018) 594-596

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Text independent emotion recognition for Telugu speech by using prosodic features

Kasiprasad. Mannepalli[1*],Suman.Maloji[1] , Panyam. Narahari Sastry[2], Swetha.Danthala[1], Durgaprasad. Mannepalli[3]

[1]K. L. university, Guntur (Dist), India
[2] CBIT, Hyderabad, India
[3]Kakinada Institute of Engineering & Technology, Kakinada, India
**Corresponding author E-mail: mkasiprasad@gmail.com*

Abstract

The human speech delivers different types of information about the speaker and speech. From the speech production side, the speech signal carries linguistic information such as the meaningful message and the language and emotional, geographical and the speaker's physiological characteristics of the speaker information are conveyed. This paper focuses on automatically identifying the emotion of a speaker given a sample of speech. the speech signals considered in this work are collected from Telugu speakers. The features like pitch, pitch related prosody, energy and formants. The overall recognition accuracy obtained is 72% in this work.
Keyword: emotion recognition, Telugu speech emotion, Speech processing

Keywords: *Emotion Recognition, Telugu Speech Emotion, Prosodic Features*

1. Introduction

Spoken language is the most common mode of communicating messages for the humans. The speech signal conveys linguistic information like message along with spoken language and speaker emotion, region and physiology of the human speech production system. Emotion, on the flip side is an individual mental state. There are various emotion which effect the flow of the speech [1]. Emotion recognition is crucial in making the speech recognition efficient

A fundamental challenge for current research in the area of speech science and technology is understanding and modelling individual variations in spoken language [2]. Individuals have their own style of speaking, depending on many factors like emotion, dialect and accent of the speech. Due to prosody of the language, speaker identification also becomes challenging task [3].

This paper focuses on automatically identifying the emotion of a speaker given a sample of speech and demonstrates how such a technology can be employed to improve the Recognition system.

The main aim of this work is to build a speech emotion recognition system that identifies emotions of the speaker based on his speech patterns within the closed-set data and with the extracted features the goal of speech recognition system is to Development of Database and Testing samples consisting of different emotions.

1) To develop an algorithm to identify speech emotion.
2) Propose a classification methodology for emotion speech recognition.

Many researchers in the area of speech processing are working in the different applications such as speaker identification, emotion recognition, speech coding etc. the researchers are working with different features such as MFCC[4-7], source features[8-12] with different classification techniques[14-17].

2. Features

Pitch:

The speech signal has to be windowed and then the auto correlation function has to be defined as short time auto correlation function as shown in the equation 1 below:

$$R_{xx}(m) = \lim_{N->\infty} \frac{1}{N} \sum_{n=-N}^{N'-1} [x(n+p)w(n)][x(n+m+p)w(n+m)] \quad (1)$$

Where *w(n)* is the suitable window generally hamming window, where *N* is the length of the speech signal, that is being analyzed and *N'* is the number of samples used for computation of auto correlation.

The pitch of the speech signal is found out by using the Auto Correlation Function (ACF) periodicity.

Energy Function

The energy of a speech signal is computed by splitting the speech signal into frames by using windowing and then computing total squared values of signal samples in each frame. The frames are multiplied by a window as shown in the equation 2

$$E = \Sigma_n \{x(n) * w(n-m)\}^2 \quad (2)$$

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Formants:

Linear Prediction Coding (LPC) method is used in estimation of formant estimation in light of the fact that the determination can be set by windowing the speech signal and obtaining the LPC coefficients
The mathematical expression for LPC can be shown in the z-domain with the equation 3.

$$H(Z) = \frac{1}{1 - \sum_{K=1}^{P} a_k Z^{-k}} \quad (3)$$

The poles of the vocal tract can be given by the roots of linear predication coding and the formants are associated with respective vocal tract poles.

3 Methodology

Eight people were identified. The work is text-independent speaker identification and the sentences given for recording the speech of the speakers were "*నీళ్ళు తీసుక రా*" (neellu thisukura), "*ఇదిగో ఇటు వచ్చి కూర్చో*" (edhigo itu vachi kurcho), "*ఉచిత సలహాలు ఇవ్వద్దుర*" (uchitha salahalu ivvaddhura). Twenty five speech samples were recorded from each speaker for each of the three sentences using the phone in five different emotions (happy, angry, sad, bore and neutral). These files which were initially in .mp3 format were converted into .wav format using *media.io*. The .wav files were given as inputs to MATLAB to obtain the time domain plots.
The nine features were extracted from each sample. The data sheet for 600 samples was prepared for each of the three sentences for each emotion. The average estimates for every emotion was figured and in this manner the average matrix was developed. Every column speaks to one emotion. These averages are then compared and the test input samples by finding the Euclidian distances and the minimum distances were gotten in other matrix. The emotion relating to the minimum Euclidian separation was recognized as the outright emotion.

4 Results and Discussions

In this work three sentences from Telugu language were selected for developing the training and testing. These sentences are given in the previous section. The sentences are chosen in such a way that the emotional difference can be identified when these speeches were recorded in different emotions.

Further Five (5) emotions (angry, bore, happy, normal and sad) were selected to analyze the recognition accuracy for eight different speakers. Every speaker spoke 5 times for each emotion. Therefore the total number of speech samples for Eight (8) people with five emotions and five times of three sentences 8X5X5X3 equals 600 samples (speakers-8, emotions-5, iterations-5 and speaches-3).

For every speaker, given an emotion, five speeches were recorded for every sentence.

The features extracted from the Telugu emotion speech samples are:

1) Pitch and its related prosody
2) energy
3) Formants.

For every emotion the average value was found from these five samples from eight speakers of that respective emotion. This value is used for the training phase in the proposed work and testing is carried out by taking any of the combinations of two test samples.

From the table 1 it can be obtained that twenty-five test samples out of forty-eight samples tested were successfully recognized by this proposed algorithm for angry emotion. Therefore, the percentage recognition accuracy for the anger emotion of speech is (25/48) 52.08 %.

On the contrary, for the Bore emotion it is thirty-five test samples out of forty-eight samples tested, were correctly recognized by the algorithm. Hence, the percentage of bore emotion recognition accuracy is (42/48) 87.50 %. For Happy emotion thirty-seven tested samples were correctly recognized out of forty-eight samples given. Therefore, the percentage of happy is (37/48) 77.08 %. For Normal emotion thirty-six test samples were successfully recognized out of forty-eight test samples. Therefore, the percentage recognition accuracy is (36/48) 75.00 % and for sad emotion thirty-five samples were correctly recognized out of forty-eight samples tested which gave a recognition accuracy of (35/48) 72.91 %.

Table 1 Efficiency table for individual emotion

Emotions	No. of Samples	Correctly Identified Samples	Percentage of accuracy
Angry	48	25	52.08
Bore	48	42	87.50
Happy	48	37	77.08
Normal	48	36	75.00
Sad	48	35	72.91

5. Comparison with existing method

This work is compared with the work done by Nitisha and Ashu Bansal [13]. They have created content ward frameworks that have been prepared for a specific emotion. All recognition systems contain three important modules: Data acquisition, extraction of features and development of recognition algorithm. The comparison is shown in Table 2.

Table 2: Comparison of the existing and the proposed systems

Description	Published Method	Proposed Method
Language	Hindi	Telugu sentences
Speech sentences recorded	Ek, do, teen, char etc. (The numbers one, two three, four uttered in Hindi)	"*నీళ్ళు తీసుక రా* " (neelluthisukura) , "*ఇదిగో ఇటు వచ్చి కూర్చో* " (edhigoituvachikurcho) , "*ఉచిత సలహాలు ఇవ్వద్దుర* " (uchithasalahaluivvaddhura)
Speech Recorded system	Recording using Microphone	Mobile Phone
Type of Identification	Text Dependent emotion	Text Independent emotion
Features Used	Mel Frequency Cepstrum Coefficient	Pitch, Intensity and Three Formants.
Classification model	Vector Quantization	Nearest Neighborhood Classifier

In the proposed method, a text independent Speech emotion recognition was developed using **Telugu sentences**. The sentences were recorded using the mobile phone of SAMSUNG make. The classification used was the Nearest Neighborhood Classifier (NNC) for feature matching. The emotion of the trained samples with which the test samples has least difference was identified as the correct emotion. The system has found to possess accuracy as high as **72.91%**.

6. Conclusions

Telugu emotion speech database is prepared from 8 speakers with emotional speeches of anger, neutral, happy, sad, bore successfully. The average value of each feature of all the samples have been given to the system for training and testing is done by trying various combinations of two samples at a time and also for all the collected speech samples at a time. In the case of two test samples which got higher efficiency, the recognition accuracy obtained for "Bore" is 87.50% which is the highest among all the other emotions tested. Expected results are in line with the natural phenomenon and the recognition accuracy of emotion "angry" is 52.08%, since the speaker unwillingly spoke the sentence in line with the natural phenomenon. The overall efficiency of emotion recognition system by using the above features and nearest neighborhood classifier (NNC) is found to be as high as72.91 %.

7. Future Scope

1) The emotion recognition system performance may be increased by using MFCC features.

2) This system performance can be increased by using Deep Neural networks for classification.

3) The number of speech samples can be increased to get better results.

4) The data acquisition system can be developed to automatically record the speeches of any Indian language in particular Telugu.

References:

[1] Jia Rong, Gang Li, Yi-Ping Phoebe Chen "Acoustic feature selection for automatic emotion recognition from speech", Elsevier, Information Processing and Management volume 45 (2009).

[2] Hagai Aronowitz and David Burshtein, "Efficient Speaker Recognition Using Approximated Cross Entropy (ACE)" IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 7, September 2007, pp. 2033-2043.

[3] Yuan-Fu Liao and Yau-Tarng Juang "Latent Prosody Analysis for Robust Speaker Identification", IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 6, August 2007, pp. 1870-1883

[4] Ning Wang et al, "Robust Speaker Recognition Using De-noised Vocal Source and Vocal Tract Features", IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, No. 1, January 2011, pp. 196-205.

[5] Nobutoshi Hanai and Richard M. Stern, "Robust speech recognition in the automobile" Carnegie Mellon University, Pittsburgh.

[6] Seiichi Nakagawa et al, "Speaker Identification and Verification by Combining MFCC and Phase Information", IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 4, May 2012, pp. 1085-1095.

[7] Marco Grimaldi and Fred Cummins "Speaker Identification Using Instantaneous Frequencies", IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 6, August 2008, pp. 1097-1111.

[8] Ji Ming et al, "Robust Speaker Recognition in Noisy Conditions", IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 5, July 2007, pp. 1711-1723.

[9] Jamel Price and Ali Eydgahi, University of Maryland Eastern Shore, "Design of Matlab®-Based Automatic Speaker Recognition Systems".

[10] Karthikeyan Umapathy et al "Au

[11] dio Signal Feature Extraction and Classification Using Local Discriminant Bases" IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 4, May 2001, pp. 1236-1246.

[12] Khalid Saeed and Mohammad KheirNammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image", IEEE Transactions on Industrial Electronics Vol. 54, No.2, April 2007, pp. 887-897.

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

[14] Kishore, P.V.V., Kishore, S.R.C. And Prasad, M.V.D., 2013. Conglomeration Of Hand Shapes And Texture Information For Recognizing Gestures Of Indian Sign Language Using Feed Forward Neural Networks. International Journal Of Engineering And Technology, 5(5), Pp. 3742-3756.

[15] Ramkiran, D.S., Madhav, B.T.P., Prasanth, A.M., Harsha, N.S., Vardhan, V., Avinash, K., Chaitanya, M.N. And Nagasai, U.S., 2015. Novel Compact Asymmetrical Fractal Aperture Notch Band Antenna. Leonardo Electronic Journal Of Practices And Technologies, 14(27), Pp. 1-12.

[16] Karthik, G.V.S., Fathima, S.Y., Rahman, M.Z.U., Ahamed, S.R. And Lay-Ekuakille, A., 2013. Efficient Signal Conditioning Techniques For Brain Activity In Remote Health Monitoring Network. Ieee Sensors Journal, 13(9), Pp. 3273-3283.

[17] Kishore, P.V.V., Prasad, M.V.D., Prasad, C.R. And Rahul, R., 2015. 4-Camera Model For Sign Language Recognition Using Elliptical Fourier Descriptors And Ann, International Conference On Signal Processing And Communication Engineering Systems - Proceedings Of Spaces 2015, In Association With Ieee 2015, Pp. 34-38.