# Isolated Telugu Speech Recognition On T-DSCC And DNN Techniques

**Archek Praveen Kumar, Neerudu Uma Maheshwari, Y.Sangeetha ,P. Jyothi**

***Abstract***— ***Communication is the major path to convey the information. Speech is the best mode for conveying the information. Human to human information can be exchanged through some particular language. But the interaction between human and machine is the major challenge which deals with ASR (Automatic speech recognition). This research recognizes speaker independent data which gives good results by using T-DSCC (Teager energy operator delta spectral cepstral coefficients) feature extraction technique and DNN (Deep Neural Networks) feature classification technique. This paper also uses CASA technique for pre-processing the speech signals. This research is done by creating the database for 10 most speak able isolated words in Telugu.***

***Keywords***— ***Isolated speech recognition, Telugu language, T-DSCC, DNN..***

## I. INTRODUCTION

In communication systems speech is the major role among human and this is efficiently used in information exchange. The further technological development is natural language of speech identification for human computer identification (HCI). Various algorithms are developed for identification of speech. Algorithm is method of translating speech signal to sequence of words which is executed by computer program [1].

The main aim of speech recognition is to establish a method for input of a speech to machine. The main application of speech recognition is identifying speech automatically.

Automatic speech identification is used to interact with human and computer. Speech is a one dimensional signal which has different properties like frequency, time, pitch, pressure and intensity. Depending on these features speech can be recognized in different ways. Speech recognition is depends on voice and speech is recognized then converted to text. Voice recognition is used in security purposes. These type of applications are depends on speaker. In conversion of voice to data it is independent on speaker. Research of speech identification has started in 1950s in different ways [2,3]. To understand the human language by computers different studies are done. From many years it has become a challenging task. The basic problem is sampling speech signal. The first system launched before 1980s which has actually speech translation.

Research on speech is to recognize the words. Words may be single or group of words. Single words are known as isolated words and group of words are known as continuous speech. Speech may be of any language [4]. In India more languages exist in this paper Telugu language is chosen for speech recognition.

There are different types for automatic speech recognition system. Those are isolated speech recognition system and continuous speech recognition system. Single Isolated word or single sound is used at a time in isolated speech. Continuous speech is natural way of speaking. It is difficult to recognize the continuous speech. In this paper we are working on isolated word speech recognition technique [5].

Speech can be recognized in different ways as pre-processing of features, extraction of features, selection of features, reduction of features, post processing of features, classification of features. It is also known as pattern recognition which contains two step procedure named as training and testing [6,7]. The constraints of speech are explained using more examples in training phase. And in testing phase, the test pattern is matched with the trained model of each data. Pre-processing deals with amplification of speech signal. After amplification extraction and classification is done by using different methods like MFCC, DNN, LPC and RASTA. After extracting selection of features is done using filtering methods like LDA and ANOVA wrapper methods and embedded methods. In post processing of features different steps like sharpening, de-noising is done. In this paper an isolated word is extracted by using two stages deep learning neural network (2-DNN) and stressed speech can be recognized by using teager energy and delta spectral cepstral coefficients (DSCC). Feature classification is done by deep neural network (DNN) and computational auditory scene analysis (CASA).

## II. BLOCK DESCRIPTION

Speech recognition of Telugu language is done by using various stages in which described by the figure 1. This is done on Telugu language. Telugu language is evolved from Dravidian language [8]. Telugu language has 60 letters where every word in Telugu language ends with a vowel.

**Manuscript published on 30 September 2019.**
***Correspondence Author(s)**

**Dr. Archek Praveen Kumar,** Professor, HOD, Department of ECE, Malla Reddy College of Engineering for Women, Hyderabad, Telangana, India.
**Neerudu Uma Maheshwari,** Associate Professor, Department of ECE, Malla Reddy College of Engineering for Women, Hyderabad, Telangana, India.
**Y.Sangeetha,** Assistant Professor, Department of ECE, Malla Reddy College of Engineering for Women, Hyderabad, Telangana, India.
**P. Jyothi,** Assistant Professor, Department of ECE, Malla Reddy College of Engineering for Women, Hyderabad, Telangana, India.

*Retrieval Number: K25440981119/19©BEIESP*
*DOI: 10.35940/ijitee.K2544.0981119*
*Journal Website: www.ijitee.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*and Sciences Publication (BEIESP)*
*© Copyright: All rights reserved.*

The slang of Telugu is quite different which accounts various considerations called as Acchulu and Allulu. All the ASR has same stages where speech recording is done and data base is created followed by feature extraction and feature classification.

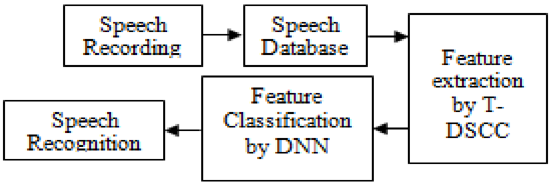

**Fig 1 : Block diagram**

Telugu language is one of the important language in south India where many speakers are spread over the world. All the words in Dravidian Telugu language ends with the vowels which is the specialty of this language. Recognition of this language is somewhat an easy task since there are limited words which are used regularly. Recognized speech is of male 250 and female 250 of age group from 6 years to 70 years. The speech signals are sampled at 16kHz frequency.

*a. Pre-Processing*

*Computational auditory scene analysis (CASA):*

Functioning of Automatic speech recognition system is reduced extremely in noisy environment to identify the speech. Different procedures are developed to recognize speech in noisy environment like distributed speech recognition and spectral subtraction [9]. Noise may be stationary and time-varying. When it is Stationary the system will be enhanced but in time varying the performance will be minimized. By using speech masking it is comfortable to isolate our speech from additive noise. The feature of human voice system can be extracted speech signal from noisy background by using computational auditory scene analysis (CASA) technique [10].

CASA technique is the best way to process the speech signal effectively in a noisy environment. Barker, Cooke and Ellis are suggested a strong method for CASA model in speech recognition.

In this method bottom up partition and assembling will be implemented by top down method. Even in typical environment it offers best functioning with this CASA techniques. But time and space is a problem in this technique. Another researcher developed an algorithm for high frequency speech.

Operation of CASA System: it is combined with speech separation from noise and speech based reconstruction. The total system is divided has different blocks to divide speech into frames, to detect the pitch and de-noising done in each frame, then speech is recognized [11].

Front end analysis is important for every speech recognition system. As the recorded speech signal is an integration of speech and noise so there are few steps which are followed in preprocessing as shown in figure 2

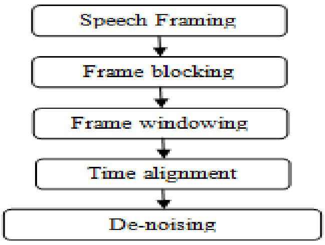

**Fig2: Pre processing flow diagram**

CASA technique initializes speech signal by auditory peripheral method followed by speech separation. Then the pitch of speech is determined and it is applied as a important factor for separating speech signal. In this technique depending upon speech feature like pitch data speech frames will be joined according to time and frequency units[12]. These streams are refined and recombined together then organized to speakers by the module.

*b. Feature extraction*

Feature extraction is one of the major parts in speech recognition. MFCC is the most important techniques used for feature extraction.

Mel Frequency Cepstral Coefficients (MFCC) features are most commonly used for speech as well as emotion recognition obtaining a good appreciation rate. Characteristically, the starting 13 cepstral constants are used because MFCC in small frequency area have a good frequency firmness, and extreme frequency constants do not get acceptable correctness. Mean Variance Normalization (MVN) is done to remove the acoustic modification from the structures additional than feeling. Delta and double-delta structures are also calculated for every frame to mend the trend report in the frame-by-frame study [13]. Thus, 39 MFCC structures are attained for every frame of every dialog sound, as explained in Fig. 3

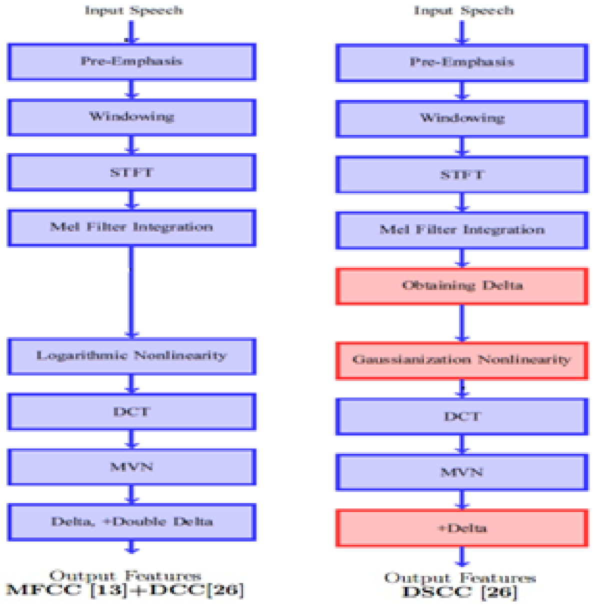

**Fig 3: MFCC and DSCC flow diagram**

In recording the excellence of the dialogue usually differs extensively. These data may have been recorded through different types of channels, such as a cell phone, and a room microphone. In addition, the data might comprise diverse stages and kinds of stabilizer sound, such as white noise, babble noise, and music. Finally, dialogues may also be recorded in different acoustic environments with different impulse responses. So, a principle acknowledgement structure would must to be hard to channel outcomes, sound, and echo Set of structures, called delta-spectral cepstral coefficients (DSCC), was necessary to improve acknowledgement correctness through accomplishing the first delta operation in the spectral area instead of the cepstral area

The DCT process wraps the 40-dimensional delta-spectral structures to a 13-dimensional vector of delta-spectral cepstral constants (DSCC) [14]. Double-delta structures are then originated from the delta-spectral structures in the cepstral area. The outlines expressive the spotless and loud dialogue indications are extra alike in the chart expressing DSCCs; this likeness referees that DSCCs should be more hard structures in sound than DCCs, possibly due to the Gaussianization nonlinearity which substitutes the logarithmic nonlinearity [15]. The mainstream of findings in the area of dialogue reaction acknowledgement have focused on the structures resulted from a linear sound creation model which adopt that air movement spreads in the spoken tract as a flat wave. Corresponding to studies by Teager , though, this supposition might not grasp since the movement is essentially distinct and attendant vortices are spread throughout the vocal tract.

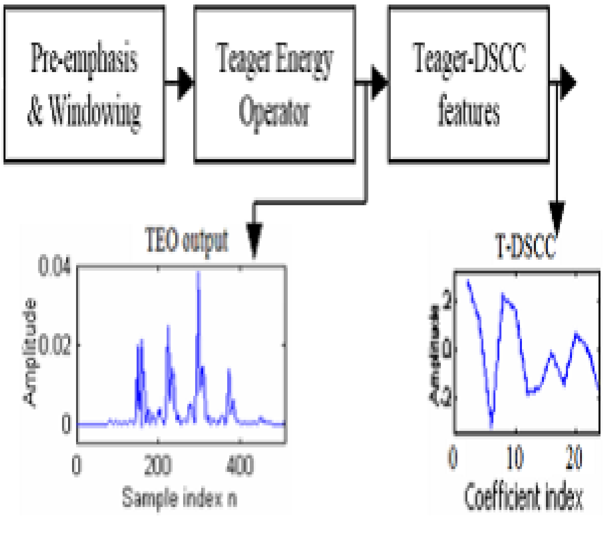

**Fig 4: Flow of T-DSCC**

Teager proposed that the real foundation of sound creation is essentially the vortex-flow connections, which are non-linear. This opinion was reinforced by the theory in fluid mechanics along with numerical replication of Navier–Stokes calculation. Hence, non-linear sound structures are essential for organization of dissimilar expressive position. In an attempt to reproduce the immediate power of nonlinear vortex-flow communications, Teager industrialized an energy operator, with the accompanying opinion that hearing is the course of noticing the energy. The simple and classy form of the operator was introduced by Kaiser.

As a formidable nonlinear operator, TEO gives a outstanding presentation in the arena of background sound clampdown and indication structure withdrawal. Hence, we suggest using TEO in the course of structure withdrawal to remove the result of sound. TEO is functional afterwards pre-processing [16]. After TEO, the typical steps of DSCC are done. We get the subsequent 26- dimensional T-DSCC structures for every frame of every sound utterance of the emotion catalogue.

### *c. Feature Classification*

Features extracted are classified for recognizing. There are many techniques in which DNN is used in this paper.

A DNN is an ANN with several layers between the input and output layers. The DNN finds the correct mathematical influence to turn the input into the output, whether it be a direct link or a non- direct link [17]. The network moves through the layers analyzing the chance of each output. Each mathematical influence as such is measured a layer, and composite DNN have many layers, hence the name "deep" networks.

DNNs can typical composite a non- direct link. DNN constructions create compositional models where the thing is expressed as a coated alignment of primitives [18]. The extra layers enable alignment of structures from lower layers, potentially exhibiting composite data with fewer units than an equally execution surface network. Deep constructions include many variants of a few basic approaches as shown in figure 5

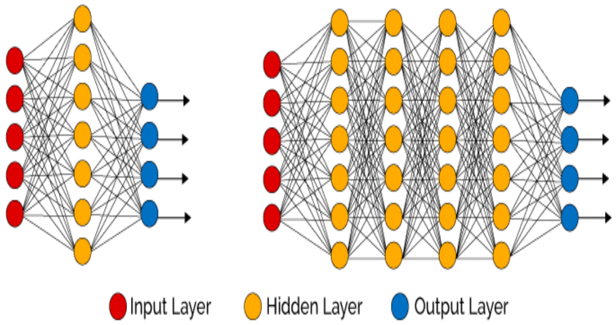

**Fig 5 : Deep learning Neural Network**

Each construction has found success in specific domains. It is not always possible to compare the functioning of various constructions, unless they have been evaluated on the same data sets.

## III. PROPOSED ALGORITHM

An algorithm is proposed according to all the techniques proposed by the authors where a flow of process is shown in figure 6. This process is almost used my many researchers in recognition of the data but only the techniques varies.

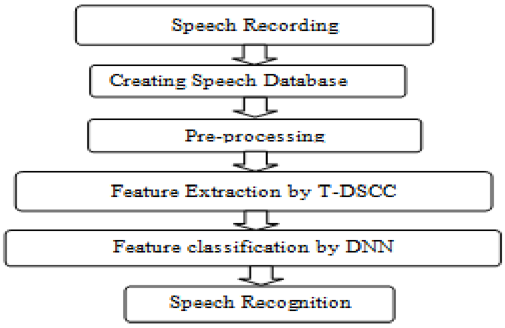

Fig 6 :Proposed algorithm

## IV. RESULTS AND DISCUSSION

The results are obtained with 97.32% by considering various techniques at different stages.

Pre-processing- CASA

Feature extraction- TDSCC

Feature Classification- DNN

Considering all these techniques provides best results. Ten most speak able words in daily life are considered and database is created. Total 50000 speeches are taken. These speeches are recorded by the combination of male and female speakers of different age.

Table II: Features extracted

| Features by TLPC | No Features |
|---|---|
| TLPC coefficients | 12 |
| Energy | 1 |
| Mean | 2 |
| Entropy of energy | 1 |
| Spectral centroid | 1 |
| Spectral spread | 12 |
| Spectral entropy | 1 |
| Spectral flux | 1 |
| Spectral roll off | 3 |
| Spectral density | 4 |
| Fundamental frequency | 12 |
| ZCR | 1 |
| Peak amplitude | 1 |
| Standard deviation | 1 |
| Total | 52 |

The figure 7 and figure 8 shows the recording of the isolated word “ATTHA”and its recognition accuracy

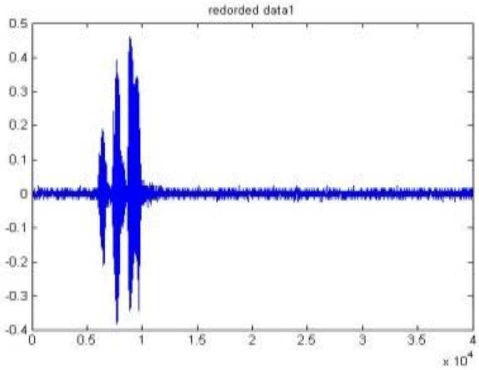

Fig 7: Recorded Isolated word “ATTHA”

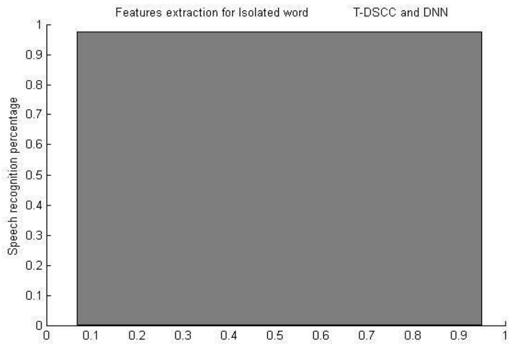

Fig 8 : Recognition accuracy

## V. CONCLUSION

The paper deals with various procedures starting from recording, creating the data base, pre-processing, feature extraction and feature classification. Selecting the suitable technique for Telugu Isolated words (10 words) generated 97.32% of accuracy. These results can be compared by the referred papers which are published already and best results are obtained. All the isolated words are recorded in wave format and later processed. This finally concludes the automatic speech recognition for Telugu language isolated words.

## REFERENCES

1. P. Ramesh babu–Digital Signal Processing; Fourth edition; SciTech Publications, 2003.
2. A. K. Yadav, R. Roy, R. Kumar, C. S. Kumar and A. P. Kumar, "Algorithm for de-noising of color images based on median filter," *2015 Third International Conference on Image Information Processing (ICIIP)*, Waknaghat, 2015, pp. 428-432.
3. A. P. Kumar, N. Kumar, C. S. Kumar, A. K. Yadav and A. Sharma, "Speech recognition using arithmetic coding and MFCC for Telugu language," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 265-268.
4. S. Sunny, D. Peter, K. Jacob, “A comparative study of wavelet based feature extraction techniques in recognizing isolated spoken words” *International Journal of signal processing systems*, Vol. 7, No. 10, 2013
5. S. Sunny, D. Peter, K. Jacob, “ Combined feature extraction techniques and Naïve Bayes classifier for speech” *Computer science and information technology, Academy & industry research collabration centre*, Vol. 3, 2013
6. A. P. Kumar, N. Kumar, C. S. Kumar, A. K. Yadav, “Speech compression by adap tive Huffman coding using Vitter algorithm”, *International Journal of Innovative Sciences*, vol. 2, No. 5, May2015.
7. S. Sunny, D. Peter, K. Jacob, “Performance of different classifiers in speech recognition” *International Journal of Research and Engineering Technology*, Vol. 2, No. 4, 2013
8. M. Rajesh, A. Hema, V. R. Rao, “ Continouse speech recognition using joint features derived from modified group delay function and MFCC”, *semantic scholor journal*, 2012
9. A. P. Kumar, R. Roy, S. Rawat, A. Sharma, “Telugu speech feature extraction by MODGDF and MFCC using Naïve Bayes classifier”, *International Journal of Control Theory and Applications*, vol. 9, No. 21, Dec 2016.
10. R. Thangarajan, A.M. Natarajan, M. Selvam, “Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language”, *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, vol. 4, No. 3, March 2008.

11. R.L.K.Venkateswarlu, R. Ravi Teja, R. Vasantha Kumari "Developing efficient speech recognition system for telugu letter recognition" *IEEE transactions on communications*, vol 29, pp. 34-39, 2013
*12.* Y. Gotoh, M. Hochberg, "Incremental map estimation of HMM for efficient training and improved performance", *In Proceeding of IEEE International conference on Achoustic Speech and Signal processing*.
13. N. Kalyani, K.V.N Sunitha, "Syllable analysis to build a dictation system in Telugu language", *International journal of computer science and information technology*, Vol. 6, No. 3, 2009
14. C. P. Dalmiya, V. S. Dharun, K. P. Rajesh, "An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications*" IEEE Conference on Information and Communication Technologies*, 2013.
15. S. Singh, E. G. Rajan, "Vector quantization approach for speaker recognition using MFCC and inverted MFCC", *International Journal of Computer Applications*, Vol. 17, No. 1, March 2011.
16. M. Hossan, S. Memon and M. Gregory, "A novel approach for MFCC feature extraction", *International Conference on Signal Processing and Communication Systems*, pp. 1-5, 2010
17. K. Hornik, M. Stinchcombe and H. White, "Multilayer feedforward networks are universal approximators" *Neural Networks* 1989 vol. 2, pp. 359-366.
18. George E. Dahl, Dong Yu, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 20, No. 1, Jan 2012