International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Speech Recognition by Dynamic Time Warping Assisted SVM Classifier

Sanaullah Ahmad Rizvi, M Sundararajan

Abstract*: Speech recognition using sustenance vector machine assisted by Dynamic time warping (DTW) method is proposed. The input training datas are collected from 40 speakers for five unique words. Every one of the information was gathered in a profoundly acoustic and commotion confirmation condition. Mel recurrence cepstrum coefficients (MFCC's) are represented as constant property of the signal. First and second derivatives of MFCC are used for dynamic properties. Subsequent to deciding element vectors, an adjusted DTW technique is proposed for highlight coordinating. Support Vector Machine (SVM) as well as Radial basis function (RBF) are used to categorize. The model is tried for multiple speakers and a good detection rate is obtained.

Keywords: Mel frequency cepstrum coefficient, Delta feature, Delta-Delta feature, Dynamic time warping, Support vector machine.

I. INTRODUCTION

Automatic Speech Recognition (ASR) translates the speech to writings, have accomplished gigantic enthusiasm in Human-machine association [1-2]. It is currently utilized in the safety framework, medicinal services, communication, court revealing, telematics, programmed subtitling, air separating and military purposes. Numerous scientists and specialists are directing examination to increase the precision of the ASR framework by suggesting a novel system as well as calculations through various parameters. In spite of the fact that various research led for the development of ASR, hearty as well as precise ASR isn't grown conversely since the non-semantic substance of speech. People could wipe out this non-semantic substance. In any case, on behalf of a machine, it's a significant test. To confront the test, distinctive techniques are put forward and concluded that an ASR framework have the most part three stages: highlight extraction, include coordinating and information arrangement.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Sanaullah Ahmad Rizvi*, Research Scholar, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India.
M Sundararajan, Professor, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India.

Amongst them, the generally utilized algorithms are Linear prediction coefficients (LPC), Perceptualstraight forecast (PLP), Mel recurrence cepstrum coefficients (MFCC) and Relative spectral-perceptual linear prediction PLP (RASTA-PLP) on behalf of dialogue acknowledgment [3],[6-7]. LPC procedure isn't suitable for speaking to discourse since it think about flag as stationary inside a casing and consequently not investigate the confined occasions precisely [3]. Alternative element Mel recurrence cepstrum coefficient (MFCC) presented by Davis and Mermelstein in 1980, has turned into the recent method in favour of speech acknowledgment [9], [6]. MFCC is a non-parametric recurrence space method dependent on human sound-related recognition framework. It possess less calculational unpredictability and enhanced execution for speech acknowledgment framework. The execution of MFCC is relative to the signal to noise proportion. Consequently, for little noisy form, MFCC is superior to some another strategy [7]. Subsequent to extraction of features, matching of features is the important. As discourse flag is for the most part successive information, it is hard to prepare the information straightforwardly by customary neural system or previous directed learning. But the calculation, Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Recurrent Neural Network (RNN) are generally utilized for highlight coordinating reason [1-3]. DTW calculation is on behalf of the most part utilized for deciding the ideal closeness among two successions. For deciding the closeness among this two groupings, they are 'distorted' non-directly regarding time pivot [3], [10].It is a standout amongst the majority productive technique for looking at time-subordinate vector information, for example, MFCC highlight vector. SVM is predominantly an administered form. For, ASR framework non-straight characterization is essential. Regular sorts of bits employed to isolate non-straight information are polynomial portions, spiral premise parts (RBF), and direct bits. These portions examine information to pass a hyper-plane and therefore arrange the information.

The previously denoted ASR calculations are connected in English language. Utilizing HMM put together ASR with respect to whole and center TIMIT informational collections, the telephone

acknowledgment rates are 67% and 65% separately [8]. English language acknowledgment is additionally vital as complete 233 million individuals talk in English [4]. To separate English word acknowledgment, HMM oriented ASR display with MFCC is put forwarded, It increased 83.97% precision [5].

Contrasting other methods with DTW oriented ASR requires less calculational weight. In this way, SVM classifier connected to MFCC parameters are better for speech recognition calculation. This article offers an idea of SVM and DTW to recognize the speech. In addition, subsidiaries of the features of MFCC are additionally used with MFCC.

The article is organized as follows: Suggested system details are explained in section II. Implementation assessment of the projected representationis displayed in segment III. Segment IV describes main factors of the test.

II. PROPOSED METHODOLOGY

A. Classification Methodology:

Recognition of speech words possess the following three stages: extraction of parameters, feature equivalence and finally the classification of phrase. MFCC features are extracted intended for all the edge. Dynamic time distorting is utilized for highlight coordinating of two element vector.

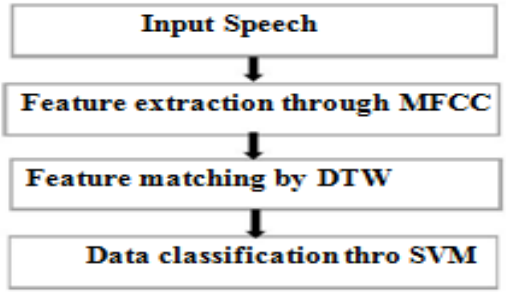

Fig.1 Block Diagram of ASR

Support vector machine is utilized to arrange the confined word. The outline to classify is appeared in Fig. 1.

B. Extraction of Features:

Any speech recognition framework, removing and choosing the best factor of signal as well as portrayal of it acts as a vital task as it altogether influences the execution of acknowledgment. MFCC highlights depend on speech recognition [3],[6]. The normal recurrence recognition capacity of a human is around the scope of 1KHz [2]. For MFCC constants assurance, the accompanying activitiesare executed on speech: Earlier stage-underscoring, separating utilizing Mel-channel store, task on sifted range, and DCT. The steps involved in MFCC is displayed in Fig. 2.

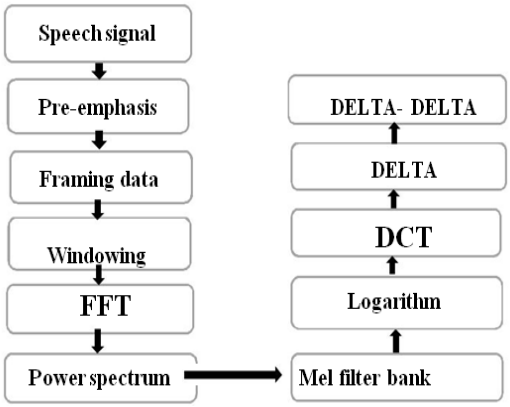

Fig.2 MFCC feature extraction

1) ***Pre-Emphasize of Signal*:** The speech signal ischaracterized by means ofa[n] and they are allowed to pass by means of the high pass filter. Considerb[n] as the high pass filter's output and when the value changes from 1.0 to 1.1, b[n] is be characterized as below:

$$b(n)=a(n)- a(-1) \quad (1)$$

By the z transform of b(n) , the transfer function is

$$H(z)=1- a \quad (2)$$

The method increases the elevated frequency parts in speech. The high pass filter produces an efficient output and is shown inFig. 4.

*2) **Framing Signal*:** Speech based signals keep on moving all the time. Hence, the entire signal is fragmented into tiny M frames, the normal rate of the frame size is 20-25ms and stride size of frame is 10ms [6]. Hence, when the first frame starts at five seconds, subsequently second frame will start in 15ms.

***3) Windowing of Frame*:** The essential to diminish the spectral leakage is windowing, because of the break of time domain. All the frames are multiplied through window to create the stability of all the points of the frame starting from the first to the final frame. It is shown in figure.3.

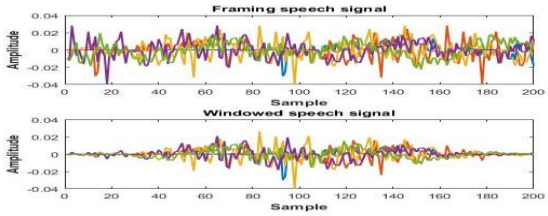

Fig.3 Windowed speech signal

***4) Power Spectrum and FFT* :** FFT is employed to change every signal form to frequency field from time field. FFT is pertained in all frames and256 FFT spectral points are taken into account to avoid severance FFT is computed .

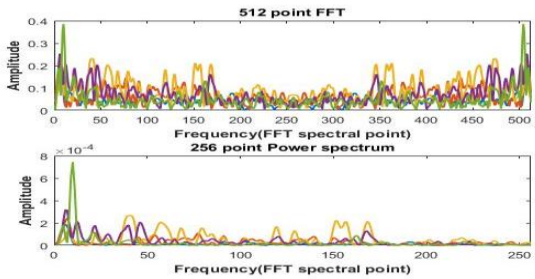

Fig.4 FFT spectral point

***5) Mel filter store*:** The Mel scale produces aassociation amongoriginal and received frequency of the communication. It has 26 feature vector and each of them has length 256. To obtainthe Mel filter store, a lower and upper frequency of 300 and 4000 Hz is selected with respect to 8 kHz sample frequency [6]. They are changed into Mel frequency. For 26 filter store, this Mel frequency is fragmented into 28 uniformly dividedvalue. The attained 26 Mel frequencies are then changed to Hertz.

After making the value of frequencies to round off, It can be find that the finishing filter store completes at bin 256, which match up to 8 kHz with a 256-point FFT size. Then a Mel filter bank is created by 26 frequency and atlast Mel filtered values are found by multiplying the power spectrum with the filter store.

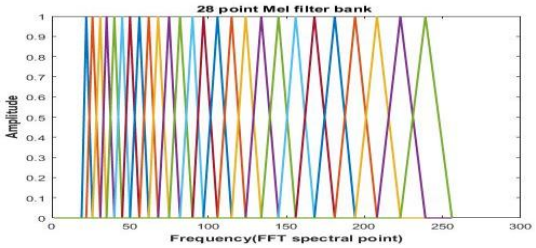

Fig.5 Mel Filter Store

6) ***Discrete cosine transform (DCT)*****:** DCT is employed to every stage of Mel separated out spectrum, T(n) which is in logarithm scale and atlast to find 26 point MFCC. An M feature DCT is described as [10].

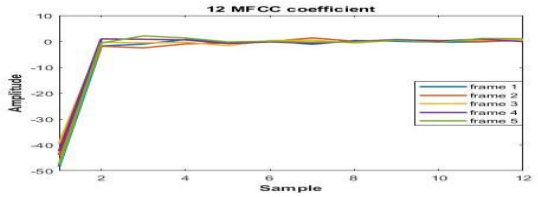

Fig.6 Mel frequency cepstrum coefficient

7) Coefficient of Delta**:** The MFCC feature vectorhas barely the influence spectrum module of every stage and these stages does not have connection among stages although speech signal has vibrant details [6].

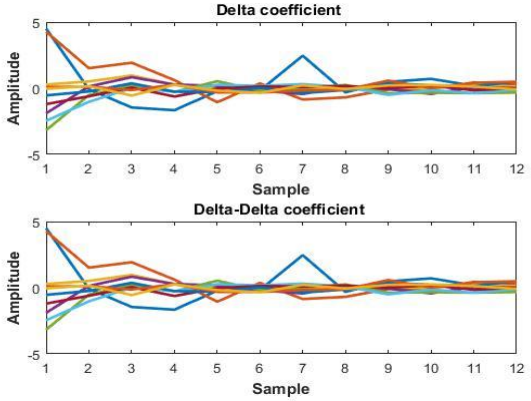

Fig.7 Delta and Delta-Delta coefficient

Each frame is associated to the previous frame . Delta coefficient produces a connection between frames by considering derived from MFCC feature vector.

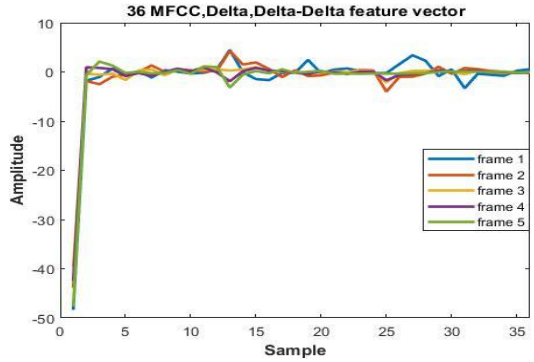

Fig.8 36 feature vectors

Delta-Delta coefficient is calculated by means of delta through the same method as followed to find delta from MFCC. Where MF is the MFCC feature vector of frame, M=2 and RC is the row count of the entire matrix. Atlast, 36 coefficients of MFCC and the feature vector are calculated for every frame.

C. Matching of Features :

DTW is employed to compute the common factors of a two sets of data , which gives back the most favourable spaces among the data. Hence, DTW matches up the organized models as well as calculates the comparison among them by means of the distance among them [3]. Let, C and D are the feature vector matrix in which, the amount of rows has the 36 feature vector of every stage.

E. Data classification:

Support Vector Machine (SVM) depends on the idea of ideal isolating in hyper-plane which makes choice limit among independent modules. A hyper-plane is mostly an idea of a summed up plane. A plane with multiple measurements is for the most part viewed as hyper - plane. For non-direct arrangement, bits are utilized. Parts are capacities that can compute similitudes among perceptions. RBF is the majority well known portion decision in SVM. In this way, RBF bit is utilized in SVM classifier to make a choice limit dissecting train information.

Speech signals gained for five English words through 40 speakers are utilized in SVM to train information for illustration Decision limit and normal most extreme back is approximately 86%. From that point onward, the test information of five English words for multiple speakers are utilized to execute the assessment.

III. PERFORMANCE EVALUATION

The computation of execution assessment is imperative to check the approval of the general framework execution. The Performance for the most part relies upon the word acknowledgment time. For a solitary English word, it is tried through deciding how frequently it could effectively perceive the word on behalf of various speakers. It is determined as far as level of word acknowledgment rate which is characterized as the proportion of number of effective. Acknowledgment of word as well as the quantity of checks information of a solitary word for various speakers.

IV. CONCLUSION

Huge advancement are practiced on speech recognition meant for the segregated English words. However used for the progression isn't practically identical to specify if the words are isolated. This paper is totally centered around detached speech recognition. This article characterizes disconnected speech signal by MFCC oriented component removal with DTW and SVM classifier. We effectively increased 86.08% exactness through the suggested method. This research can be extended for ceaseless type speech recognition in future.

REFERENCES

1. RabeetFatmi ; Sherif Rashad ; Ryan Integlia, "Comparing ANN, SVM, and HMM based Machine Learning Methods for American Sign Language Recognition using Wearable Motion Sensors" , 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)
2. SalaheddineAgab ; FatmazohraChelali: "HOG and HOOF Spatio-Temporal Descriptors for Gesture Recognition", 2018 International Conference on Signal, Image, Vision and their Applications (SIVA)
3. Nidhi KalidasSawant ; SangamBorkar,"Devanagari Printed Text to Speech Conversion using OCR",2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference.
4. JayantaDey ; MdSanzid Bin Hossain ; Mohammad ArifulHaque, "An Ensemble SVM-based Approach for Voice Activity Detection" 2018 10th International Conference on Electrical and Computer Engineering (ICECE).
5. EvaggelosSpyrou ; IoannisVernikos ; RozaliaNikopoulou ; PhivosMy lonas, "A Non-Linguistic Approach for Human Emotion Recognition from Speech",2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)
6. Jeremia Jason Lasiman ; DessiPuji Lestari, "Speech Emotion Recognition for Indonesian Language Using Long Short-Term Memory",2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)
7. Karan Nathani ; AlexNoel Joseph Raj ; "Cubic SVM Classifier Based Feature Extraction and Emotion Detection from Speech Signals", 2018 International Conference on Sensor Networks and Signal Processing (SNSP).
8. HanifFakhrurroja; Riyanto; AyuPurwarianti ; ArySetijadiPrihatmanto ; CarmadiMachbub," Integration of Indonesian Speech and Hand Gesture Recognition for Controlling Humanoid Robot ",2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV).
9. ShruthiS ; Yashaswi G ; Shruti V ; Manikandan J, Design and Evaluation of a Real-Time Speech Recognition System, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
10. S. Jeyamaalmarukan,"Self-Speech Evaluation with Speech Recognition and Gesture Analysis",2018 National Information Technology Conference (NITC).