IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 1, January 2016

Speech Recognition for Isolated word using Matlab

Ms. Vishakha Nandanwar[1], Ms. Darshana Chaware[2], Mr. Sushil P. Borkar[3]

Assistant Professor, Department of Electronics, Rajiv Gandhi College of Engineering & Research, Nagpur, India [1,2]

Assistant Professor, Dept. of Electronics & Telecommunication, Priyadarshini College of Engineering, Nagpur, India [3]

Abstract: This paper present the training and testing of the data collected from male and female and then recognition of isolated word by using speech processing. One set of the data recordings was used for the training runs and other a different set was used for testing. Speech recognition process in human is running from long years. In this paper the words are recognized by using Maltab software. The isolate spoken language can be study at two different levels: (1) phonetic components of spoken words, e.g., vowel and consonant sounds, (2) acoustic wave patterns. In this paper we have used the digits from one to nine and zero but any small vocabulary could be used. So this paper presents the training and testing of the different or recognition of the words spoken from male and female.

Keywords: Speech processing, Speech recognition, vocal track MATLAB.

I. INTRODUCTION

As the growth of digital signal processing and digital computers we can say that the beginnings of modern Automatic speech recognition (ASR) .The more common control systems and the popular speech to text systems can be seen today. The voice recognition plays an important role in any detection or investigation. But the recognition of the voice by computer is used in access control and security systems. Today the automatic speech recognition is coupled with text to speech processing and can be can be used for automatic spoken language translation. A large numbers of applications has been developed. The speech recognition has been studied at two different levels: (1) phonetic components of spoken words, e.g., vowel and consonant sounds, (2) acoustic wave patterns. A language can be broken down into a very small number of basic sounds, called phonemes. The acoustic waves are a type of longitudinal waves that generally propagate by means of adiabatic compression and decompression process. Longitudinal waves are waves that have the same direction of vibration as their direction of travel. The two-dimensional pattern in sound analysis is generally called spectrograms which display frequency in vertical axis and time in horizontal axis represent the signal energy. Generally, the flow of air in the vocal tract generates that consonant which generally gives the different sounds. On the other hand modifying the shape of the passages through which the sound waves, produced by the vocal chords, travel generates vowels. Isolated vocal tracks (or vocal master tracks) give us a bare truth about how good someone actually is, and whether or not it’s all just tricks and effects added in the production. It gives great pleasure to hear a singer in an isolated environment, where we can enjoy and feel their energy on a much personal level than while listening to a studio recording. A "neutral" vowel is defined as a vowel produced by a vocal tract configuration that has uniform cross-sectional area along its entire length. The difference in the sound of spoken vowels such as 'A' and 'E' are due to differences in the formant peaks caused by the difference in the shape of your mouth when you produce the sounds. The power source for consonants is airflow producing white noise, while the power for vowels is vibrations from the vocal chords.

II. SPEECH RECOGNITION

Nowadays the speech recognition is becoming more useful. Today the computer has become most talkative system means the person say anything they can repeat the same in real time. Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The speech recognition generally converts the sound waveform into words. This is the most important and relevant task in the industry. No special equipment required for the speech recognition. The speech recognition system is basically a biometric system which is used to identify the voice of particular person. The process of enabling a computer to identify and respond to the sounds produced in human speech. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Both the term voice recognition and speech recognition can be used interchangeably. Speech recognition is used to identify words in spoken language. All the voice recognition systems or the written programs for voice recognition generally make errors in the system. For example the screaming of the children, barking of the dogs, and external conversations done loudly can produce false input for the system and system will not be able to understand.

A. Component required for speech recognition

1. Speech waveform capture (analog to digital conversion)

The analog to digital signal conversion is required in speech processing. The sampling rate of 8000 generally gives a Nyquist frequency of 4000 Hertz which should be adequate for a 3000 Hz voice signal. Some systems have

DOI 10.17148/IARJSET.2016.3117

used for over sampling plus a sharp cut off filter to reduce the effect of noise.

2. Pre-emphasis filtering

This is the other important component of the speech recognition. Because the tilt of the speech has an overall spectral 5 to 12 dB per octave and pre-emphasis filter has the form of **1 – 0.99 z-1** which is normally used. By using the first order filter we come to the fact that the lower component contains more energy than the higher. If it weren't for this filter the lower component would be preferentially modelled with respect to the higher formants.

3. Feature extraction

Next component of the speech recognition is the features extraction which is used to derive the features from the Linear Predictive Coding (LPC), which is a technical attempt to derive the coefficients of a filter that would produce the utterance that is being studied. LPC is useful in speech processing because of its ability to extract and store time varying formant information.

4. Classification

A library of feature vectors is provided – a "training" process usually builds it. The classifier uses the spectral feature set (feature vector) of the input (unknown) word and attempts to find the "best" match from the library of known words. Simple classifiers, like mine, use a simple "correlation" metric to decide. While advanced recognizers employ classifiers that utilize Hidden Markov Models (HMM), Artificial Neural Networks (ANN), and many others.

B .Two feature extractors were employed:

(1) Straight LPC on the digitized (and normalized) input.

(2) LPC-cepstrum, computed from the LPC coefficients.

III. BLOCK DIAGRAM

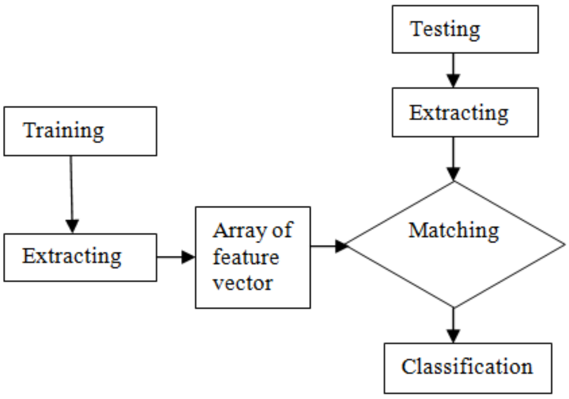


Fig.1 Block diagram function of training and testing

The above block diagram represents the functioning of testing and training. The sample of voice from the different male and female are collected. Then these samples further send for the training purpose. Next step after training is feature extraction where the features are extracted of the different samples and build the feature design of the data. And next the array of the data is designed. On the other hand at the same time the sampled data are sent to testing block for the testing purpose. Testing generally uses the routine matching to get the best match for a word to be used for testing. Again after testing the sample feature are extracted by using the extracting block. After this, the features we received from both training and testing are send for matching. This matching determine the best match along with the figure of merit After this, some threshold value is used for the figure of merits to determine up to how much information we have received. Now, the information to which test words are supplied than the program knows if it is right or wrong so the testing keeps a count of the rights, wrongs

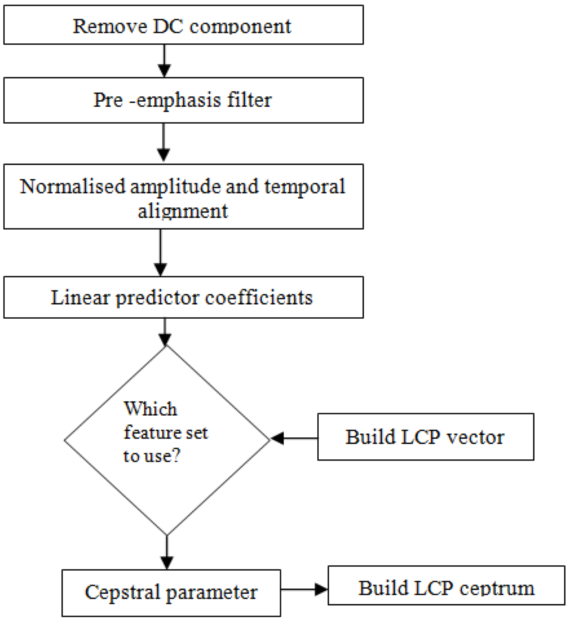

Remove DC component
Pre -emphasis filter
Normalised amplitude and temporal alignment
Linear predictor coefficients
Which feature set to use?
Build LCP vector
Cepstral parameter
Build LCP ceptrum

Fig.2 Block diagram of preprocessing and feature extracting

Firstly remove the DC component and then send it to the pre-emphasis filter and normalised the amplitude and temporal alignment and the data further send to the linear predictor to find the coefficients of the data. Next the result is check for which feature should be used and further LCP vector will be created then cepstral parameter will be found out and LCP ceptrum will be created. This data is for a combination of two speakers. Each of the sixteen runs consisted of eight trainings followed by eight testing.

IV. IMPLEMENTATION OF ALGORITHM

(1) Firstly, the correlation is measured between the features of the vocabulary words.

(2) Secondly measure the features of each word from the training sets.

(3) Then measure Euclidean distance between the averages of the words and the features of the test word in the training sets

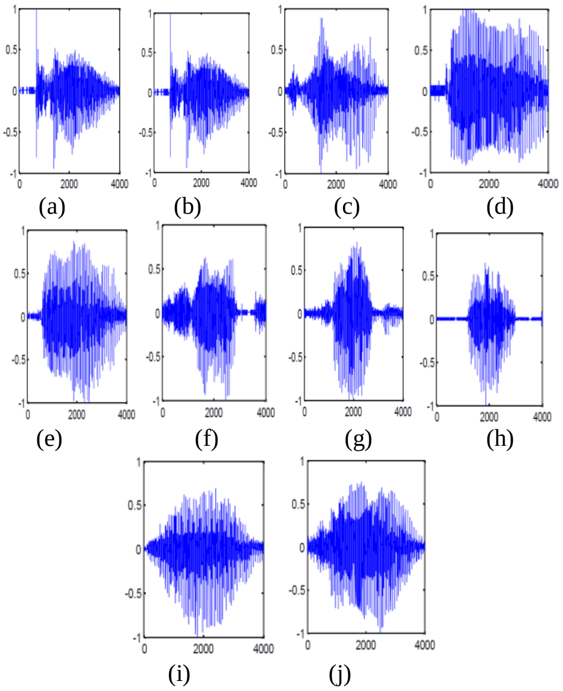
(4) Then take nonparametric k-nearest neighbour approach (a natural extension of the single nearest or closest).

(5) The Mahalanobis distance between the features of the test word and the words in the training sets.

(6) Then the correlation method will work averages for each vocabulary word and for the training feature vectors.

DOI 10.17148/IARJSET.2016.3117

V. EXPERIMENTAL RESULT



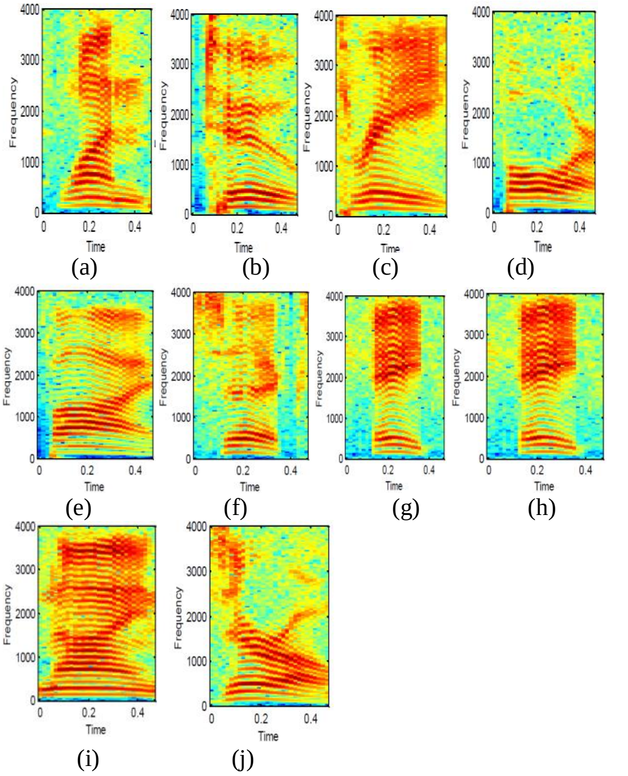
(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig.3 The above fig. a, b, c, d, e, f, g, h, i, j represent the sample data of one, two, three, four, five, six, seven, eight, nine, zero respectively.

The above fig. is sample plots taken from the numbers training data. The figure a, b, c, d, e, f, g, h, i, j represents the sample data of one, two, three, four, five, six, seven, eight, nine, zero respectively. The plots are time vs. signal amplitude for the ten vocabulary words from set: Male Training.



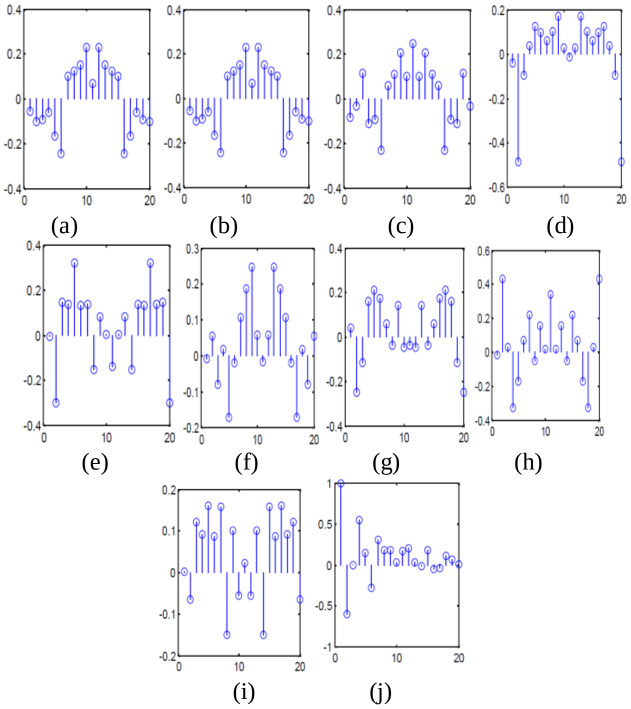
(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig.4 The above fig. a, b, c, d, e, f, g, h, i, j represent the spectrogram of one, two, three, four, five, six, seven, eight, nine, zero respectively.

The following are sample plots are spectrogram of data. The figure a, b, c, d, e, f, g, h, i, j represents the spectrogram data of one, two, three, four, five, six, seven, eight, nine, zero respectively. The plots are time vs. signal frequency for the ten vocabulary words.



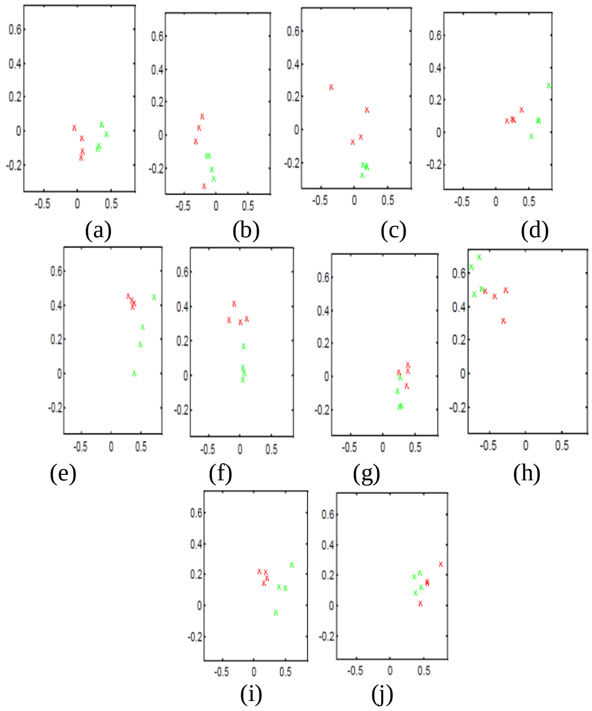
(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig.5 The above fig. a, b, c, d, e, f, g, h, i, j represent the LPC ceptrum plot of one, two, three, four, five, six, seven, eight, nine, zero respectively.

The above fig is LPC ceptrum plot for the data. The figure a, b, c, d, e, f, g, h, i, j represents the sample data of one, two, three, four, five, six, seven, eight, nine, zero respectively.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j)

Fig.6 The above fig. a, b, c, d, e, f, g, h, i, j represent the two dimensional plot of one, two, three, four, five, six, seven, eight, nine, zero respectively.

The above fig is a two-dimensional representation of the fourth vs. the seventh LPC feature (labeled 0-9) spoken four times by the same male speaker. Each of the four runs is printed in its own color.

VI. CONCLUSION

In this paper we have successfully used the numbers from one to nine and a zero. And we have successfully also used the vowel a, i, o, u. With this vowel sounds we have successfully added these familiar words like red, green, blue. This data is used for a combination of both the two speakers' one for the male and other for the female. Each of the runs consisted of eight trainings followed by eight testing. Each tape contains one utterance of each of the ten vocabulary words. We have successfully individually recorded and tapes were used for eight training and eight testing for each of two vocabularies. So in this paper we have successfully tested the words recorded from the bole male and female. Further this can be implemented for the long words. It can be used for the recognition of the speech that will be helpful in the investigation purpose.

REFERENCES

[1] K.H.Davis, R.Biddulph, and S.Balashek, 1952, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am.,24(6):637.642.

[2] B.Lowrre, 1990, The HARPY speech understanding system ,Trends in Speech Recognition, W.Lea,Ed., Speech Science Pub., pp.576-586.

[3] Jean Francois, Jan.1997, Automatic Word Recognition Based on Second Order Hidden Markov Models , IEEE Transactions on Audio, Speech and Language processing Vol.5,No.1.

[4] S. Furui, 2004, Speech-to-text and speech-to-speech summarization of spontaneous speech, IEEE Trans. Speech & Audio Processing, 12, 4, pp. 401-408.

[5] J.W.Forgie and C.D.Forgie, 1959, Results obtained from a vowel recognition computer program ,J.Acoust.Soc.Am., 31(11),pp.1480-1489.

[6] K.Nagata, Y.Kato, and S.Chiba, 1963, Spoken Digit Recognizer for Japanese Language , NEC Res.Develop., No.6.

[7] L.R.Rabiner, February 1989, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition , Proc.IEEE,77(2):257-286.

[8] K.P.Li and G.W.Hughes, 1974, Talker differences as they appear in correlation matrices of continuous speechspectra , J.Acoust.Soc.Am. , 55,pp.833-837.

[9] Giuseppe Riccardi, July 2005, Active Learning: Theory and Applications to Automatic Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4.

[10] Schomaker, L. and Segers, E. A method for the determination of features used in human reading od cursive handwriting. IWFHR, Korea, 1998, 157--168.

DOI 10.17148/IARJSET.2016.3117