International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

SIGNAL PROCESSING AND NEURAL NETWORKS BASED SPEAKER RECOGNITION SYSTEM

Dr.T. JAYASREE[1] G.RAJARAM[2] , K.RAJKUMAR[3] , WERLAN KHARSATI[4], M.SURESH[5], L.JAISLEEN GREETEL[6]

[1]Assistant Professor, [2, 3, 4, 5, 6]Students

Govt. College of Engineering, Tirunelveli

Article Received: 11/04/2015

Article Revised on:15/04/2015

Article Accepted on:18/04/2015

JAYASREE T

ABSTRACT

Speech recognition is a technology which has close connections with computer science, signal processing, voice linguistics and intelligent systems. In real life, speaker recognition has been used very frequently. Neural network is a technology which tries to mimic human brain functions. With the development of neural network the speaker recognition has become very popular and successful. In this paper, the concept of signal pre-processing, feature extraction, neural network design and implementation, are introduced. The Mel Frequency Cepstrum Coefficients (MFCC) is the best available approximation of human ear features. Back propagation neural network is used to design the recognition system. Moreover an implementation has been made in Matlab platform. The experimental results show that the system works well and it can be improved by using more training samples.

KEYWORDS: Speech, Spectrogram, Cepstrum, Neural network, Signal

1. INTRODUCTION

As the computer technology develops, it promotes the development of society. On the other hand the development of human society entails a higher challenge to the computer development. The communications between humans and computers are wider and deeper and communication functions by using mouse, keyboard and touch screen cannot satisfy the quick, accurate and efficient interchange of information. How to send information in a more natural, more efficient and quicker way has become an urgent question. From technology research to daily life, computers are involved in every aspect of people's daily life. Computers are used to accomplish many tasks. Considering this situation, intelligent communication between computers and humans, human-computer interaction, becomes one of the most important research fields. Speech is one of the natural forms of human communication. Since childhood people can express themselves by speech, recognizing others by distinguishing their voices and under- standing others by their speech. People are very good at speaker and speech recognition.

The human brain uses neurons and synapses, modified with experience and provides a distributed form of associative memory. Motivated by this, speaker and speech recognition systems have been developed.

1. Speaker recognition is the technology of letting a machine distinguishes different speakers from each other. Depending on the different speakers different actions are implemented.

2. Speech recognition is the technology of letting a machine understand human speech and, according to the meaning of the speech, implement the intention of the human.

These technologies involve wide cross-disciplinary research; close connections to computer science, telecommunications, signal processing, voice linguistics, neural networks and intelligent systems. For the past 30 years already, speaker recognition and speech recognition have been used in industry, military, traffic, medicine and daily use, and especially in automatic control, information processing, telecommunications and electrical system. As voice control technologies, speaker recognition and speech recognition will definitely affect the technologies in automation and machine operation [1].

The time of multimedia is here, urgent requirements for the development of speech recognition will promote speech recognition technology to have a breakthrough both theoretically and in its applications. Joint speech and speaker recognition system has various applications, which variable from health care, military to daily use applications. Automation of complex operator-based tasks is one of the most popular applications, e.g., customer care, dictation, form filling applications, provisioning of new services, customer help lines, e-commerce, etc. One application which enables big advantage to people's life is voice commanding car or robot. E.G RACO technology brings Google voice commands to car which achieves the GPS tracking by voice commands. Controlling robots which can be used with different robots follow its master's commands and gives the right reaction, like voice controlled wheelchair which makes the disabled people to control the chair much easier and convenient. This can be a further implementation based on the research of this project.

2. PROPOSED METHODOLOGY

The flowchart for the proposed signal processing based speaker recognition system is shown in fig.**1**

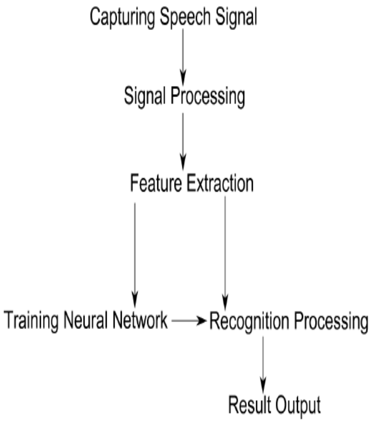


Fig 1. Flow chart of the Proposed Methodology

3. FEATURE EXTRACTION METHODS

In this section, different types of signal processing techniques for speech processing are discussed below:

3.1 SPECTOGRAM

There is a better representation domain, namely the spectrogram. This representation domain shows the change in amplitude spectra over time. It has three dimensions:

X-axis: Time (MS)

Y-axis: Frequency

Z-axis: Color intensity represents Magnitude

The complete sample is split into different time-frames (with a 50%overlap). For every time- frame, the short-term frequency spectrum is calculated. Although the spectrogram provides a good visual representation of speech it still varies significantly between samples. Samples never start at exactly the same moment, words may be pronounced faster or slower and they might have different intensities at different times [2].

The Spectrogram of the word ‘Trojan’ is shown in Fig. 2

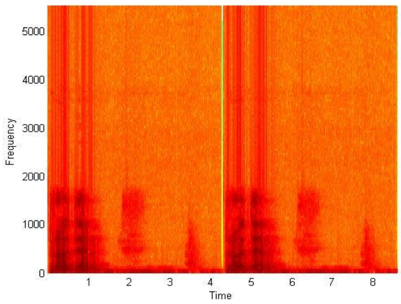


Fig 2. Spectrogram of the word ‘Trojan’

3.2 CEPSTRUM

We know that human ears, for frequencies lower than 1 kHz, hear tones with a linear scale instead of logarithmic scale for the frequencies higher that 1 kHz. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The voice signals have most of their energy in the low frequencies. It is also very natural to use a Mel-spaced filter bank showing the above characteristics. The following approximate formula is used to compute the Mel for a given frequency in Hz:

$$Mel\ (f)=2595^{10}.log(1+f/700)$$

For each tone with an actual frequency f (in Hz), a subjective pitch is measured on a scale called the 'Mel' scale. The pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mel. The Cepstrum is the forward Fourier transform of a spectrum. It is thus the spectrum of a spectrum, and has certain properties that make it useful in many types of signal analysis. One of its more powerful attributes is the fact that any periodicities, or repeated patterns, in a spectrum will be sensed as one or two specific components in the Cepstrum. If a spectrum contains several sets of sidebands or harmonic series, they can be confusing because of overlap. But in the Cepstrum, they will be separated in a way similar to the way the spectrum separates repetitive time patterns in the waveform [3].

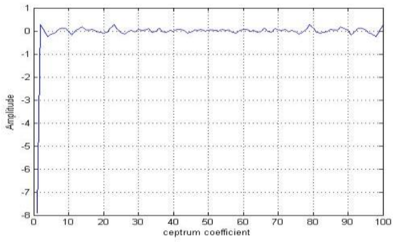


Fig 3.Cepstrum of the word 'TROJAN'

3.3 MEL FREQUENCY SCALE

It is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The name Mel comes from the word melody to indicate that the scale is based on pitch comparisons. It is a logarithmic scale similar to the way the human ear perceives sound.

As the picture shows, when the frequency is 1000 Hz, in the Mel scale it is also 1000 mel. When the frequency is below 500 Hz, the intervals are smaller compared with frequencies larger than 500 Hz. Generally speaking the Mel scale has a linear relationship with hertz when the frequency is below 1000 Hz, and a logarithmic relationship when the frequency is bigger than 1000 H

There are many formulas to convert hertz into Mel, but the most popular is:

$Mel\ (f) = 2595.log\ (1 + f/700)$

Where f is the linear frequency.

3.4 Mel Frequency Cepstral Coefficients

"In sound processing, MFC which means Mel frequency Cepstrum, is a representation of the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency". MFCC stands for Mel Frequency Cepstral Coefficients. The coefficients represent audio based on perception. They are derived from the Mel frequency Cepstrum. It is known that the human ear is more sensitive to higher frequency. The spectral information can then be converted to MFCC by passing the signals through band pass filters where higher frequencies are artificially boosted, and then doing an inverse Fast Fourier Transform (FFT) on it [4]. This results in higher frequencies being more prominent. As the Mel frequency Cepstrum can represent a listener's response system better, MFCC is always considered to be the best available approximation of human ear.

The whole MFCC processing procedure can be divided into three main steps:

First the Fast Fourier Transform is used to convert speech signal from time domain to frequency domain. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windowing. These filters follow the Mel scale. In this thesis 24 filters are used. So the outputs after filtering are:

$$Signal\ k,\ k = 1, 2, 3.....24$$

Finally, take the logs of the outputs, and take the discrete cosine transform (DCT) of the list of Mel log powers. The MFCCs are the amplitudes of the resulting spectrum. We use log because our ears work in decibels. Discrete cosine transform (DCT) will be applied to each Mel Spectrum to convert the values back to real values in the time domain. We take the DCT because it is good for compressing information. In Voice box there is a function called Melcepst that can calculate the Mel Cepstrum. Actually it can also do the framing and windowing, **so the previously mentioned enframe function** (described under pre-processing) is not even needed. Where 11025 is the sampling frequency, 'M' means Hamming window and 12 is the number of Cepstral coefficients, 24 the number of filter banks, 256 the frame length and 256-64=192 is the frame overlap. All these numbers have been discussed earlier. The result MFCCs is an N * 12 matrix of 12 MFCCs per each N frame. The number N will naturally vary for each voice sample. The MFCCs are our features that will be fed to the neural network. The NN cannot have a variable number N of inputs, so we need to do something about that.

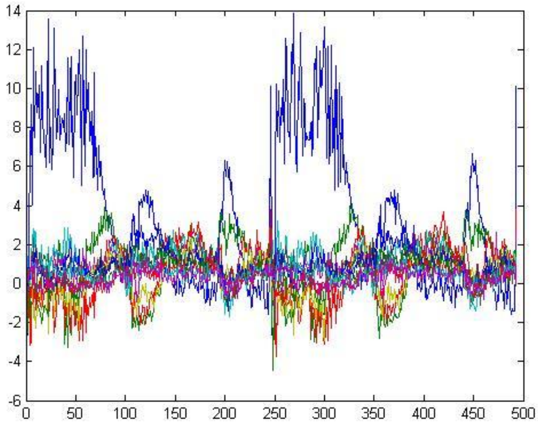


Fig 4. MFCC of the word 'TROJAN'

4. ARTIFICIAL NEURAL NETWORK FOR SPEAKER RECOGNITION

In this network, a feed forward network is used. An output from MFCC is taken as 12. Only one hidden layer is used which is 15. For speaker recognition, the output layer represents the speakers. Fig 5 shows the neural network layers.

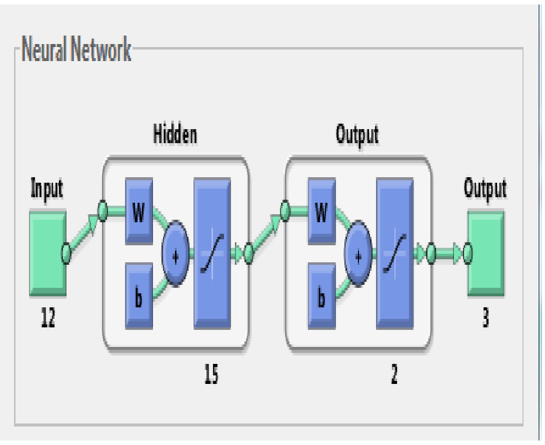


Fig 5. Artificial Neural Network layers

4.2 Network performance

The performance results of ANN are shown in fig. 6. It consists of three curves for training, validation and testing phases [5].

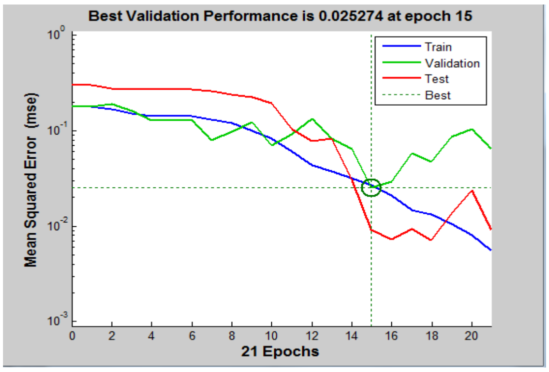
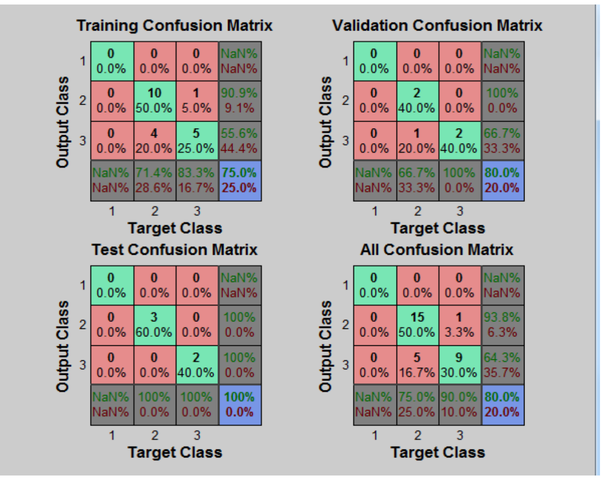


Fig 6. Performance results of ANN the word 'TROJAN'

4.3 Confusion matrix

It is a table that consists of numbers of rows and columns which reports the percentage of correct and incorrect classification. From the figure given below, it is shown that the green colour gives the percentage of correct classification and the black pink colour shows the percentage of incorrect classification. The confusion matrix results of ANN based speaker recognition system is shown in fig. 7.



Training Confusion Matrix

Output Class	1	2	3	
1	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	10 50.0%	1 5.0%	90.9% 9.1%
3	0 0.0%	4 20.0%	5 25.0%	55.6% 44.4%
	NaN% NaN%	71.4% 28.6%	83.3% 16.7%	75.0% 25.0%
	Target Class			

Validation Confusion Matrix

Output Class	1	2	3	
1	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	2 40.0%	0 0.0%	100% 0.0%
3	0 0.0%	1 20.0%	2 40.0%	66.7% 33.3%
	NaN% NaN%	66.7% 33.3%	100% 0.0%	80.0% 20.0%
	Target Class			

Test Confusion Matrix

Output Class	1	2	3	
1	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	3 60.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	2 40.0%	100% 0.0%
	NaN% NaN%	100% 0.0%	100% 0.0%	100% 0.0%
	Target Class			

All Confusion Matrix

Output Class	1	2	3	
1	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	15 50.0%	1 3.3%	93.8% 6.3%
3	0 0.0%	5 16.7%	9 30.0%	64.3% 35.7%
	NaN% NaN%	75.0% 25.0%	90.0% 10.0%	80.0% 20.0%
	Target Class			

Fig 7. Confusion matrix for ANN based speaker recognition (word 'TROJAN')

5. CONCLUSION

Speech recognition using signal processing and ANN are a hotspot of international academic circles. This paper is showing that neural networks can be very powerful speech signal classifiers. The pre-processing quality is giving the biggest impact on the neural network performance. The Mel Frequency Cepstrum Coefficients are a very reliable tool for the pre-processing stage with good results. The multilayer feedforward Neural Network with backpropagation algorithm is achieving satisfying results when mel frequency cepstrum coefficients are used.

6. REFERENCES

[1]. Han Yi, Wang Guo yin, and Yang Yong, " Speech emotion recognition based on mfcc ", .Journal of Chongqing University of Posts and Telecommunications, 2014.

[2]. Scott Chin, Kelvin Lau, and Lindsey Leu, "A speaker verification system. Report", Department of Electrical and Computer Engineering ELEC499a, 2012.

[3]. B. Plannerer, "An introduction to speech recognition Tutorial", University of Munich,2014.

[4]. Wen Lin, "Based on retro fitted mfcc speech recognition system research and design". master thesis, 2013.

[5]. Amit-Degada, " Digital coding of analog signal",lecture notes, Sardar Vallabhbhai NationalInstitute of Technology.

[6]. WangWei-Zhen, "Research of speech recognition based on neural network". master thesis, 2014

Jayasree T completed her BE degree from Barathidasan University, Trichy in 1997 and ME from Govt. College of Technology, Coimbatore in 1999. She completed her Ph.d degree from Anna University Chennai in 2011 in the area of signal processing. She has more than 15 years of teaching experience. Presently, she is working as an Assistant professor in the Department of ECE in Govt. College of Engineering, Tirunelveli. She has published many papers in National and International Journals and conferences. Her area of interest is signal processing applications in power quality, speech processing, Bio-medical signal processing.

Rajaram is doing Final year BE in the Department of ECE in Govt. College of Engineering, Tirunelveli.

Rajkumar is doing Final year BE in the Department of ECE in Govt. College of Engineering, Tirunelveli.

Werlan Kharsati is doing Final year BE in the Department of ECE in Govt. College of Engineering, Tirunelveli

Suresh is doing Final year BE in the Department of ECE in Govt. College of Engineering, Tirunelveli.

Jaisleen Greetel is doing Final year BE in the Department of ECE in Govt. College of Engineering, Tirunelveli.