IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AUTOMATION SYSTEM TO CLASSIFY HUMAN VOICE USING NEURAL NETWORK

Arshi Anjum[*], Zareen Khan, Sonali Bhujade, Nikhat Fatema Sheikh
[*]Electronics & Communication, Anjuman College of Engineering & Technology,Nagpur, India
Electronics & Communication, Anjuman College of Engineering & Technology,Nagpur, India
Electronics & Communication, Anjuman College of Engineering & Technology,Nagpur, India
Electronics & Communication, Anjuman College of Engineering & Technology,Nagpur, India

DOI: 10.5281/zenodo.47583

ABSTRACT

With the development of more and more identification system to identify a person, there is a need of development of system that can provide personal identification task such as gender identification automatically without any human interface. Gender identification using voice of a person is comparatively easier than from facial images. There exist several algorithms for automatic gender identification but none of them is been found to be 100% efficient. In this project, noble gender identification is been proposed. It identifies gender in near real time by utilizing all possible characteristics of human voice in terms of pitch. This feature provides input to trained Neural Network. Feed forward neural network is trained by providing database.

KEYWORDS: Neural Network ,Feature Extraction, MFCC, Speech Processing

INTRODUCTION

Speech is a natural mode of communication for people and we rely on it throughout our lives. It comes so naturally that we do not recognise how complex a phenomenon speech is. This motivates research efforts to allow speech to be use for human-computer interaction. Automatic Speech Recognition (ASR) is viewed as an integral part of future human-computer interfaces. The vocalizations can vary widely in terms roughness, nasality, pitch, volume, speed etc. Moreover, during transmission our irregular speech pattern can be further distorted by background noises. This all makes speech recognition and classification a complex process.

The aim of our project is to identify gender of a speaker based on voice of speaker using certain speech processing techniques in real time using MAT-LAB.To define basic constraints on our voice samples used in our system. The data type of voice is wave file. The voice is recorded in a quite environment. The less background noise we have, the better data input our system has. In this project, we try to select an existing word group. Every user should only speak this word group in the process of recognition. Additionally, we can select different word groups and compare their performance.

PROPOSED WORK

Initialization:

The input is a speech signal that is an analogue signal at the recording time, which varies with time. To process the signal by digital means it is necessary to sample the continuous-time signal into a discrete time discrete value signal. Even if they appear to be quite reasonable, it might be a good idea to consider pre-processing the data before initiating training. Pre-processing is a transformation, or conditioning, of data designed to make modelling easier and more robust. For example, a known nonlinearity in some given data could be removed by an appropriate transformation, producing data that conforms to a linear model that is easier to work with.
Similarly, removing detected trends and outliers in the data will improve the accuracy of the model. Therefore, before training a neural network, you should consider the possibility of transforming the data in some useful way.

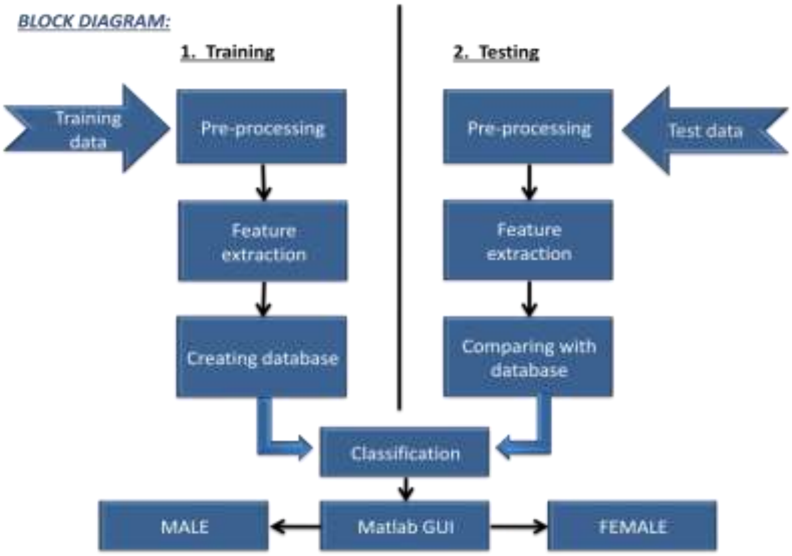

Figure 1. Block Diagram

Recording :
For recording of speech signal dsp audio recorder inbuilt in MAT-Lab is operated. For this a good quality microphone is used which have high sensitivity

Speech Processing:
In speech classification, first phase is pre-processing which deals with a speech signal, which is an analog signal at the recording time, which varies with time. . PROCESSING of speech signals, i.e. segregating the voiced region from the silence/unvoiced portion of the captured signal is usually advocated as a crucial step in the development of a reliable speech or speaker recognition system.

This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals moreover, extraction of the voiced part of the speech signal by marking and/or removing the silence and unvoiced region leads to substantial reduction in computational complexity at later stages

Filter Used:
We here in our project we are using Gaussian filter for filtering our speech signals. Using Gaussian filter the noise is suppressed and the noise is smoothed out.

Gaussian filtering is been extensively studied in image processing and computer vision.

MFCC (Mel Frequency Cepstral Coefficient):
The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature type for automatic speech recognition (ASR).

For speech signal based gender identification, the most commonly used features are pitch period and Mel-Frequency Cepstral Coefficients (MFCC). The main intuition for using the pitch period comes from the fact that the average fundamental frequency (reciprocal of pitch period) for men is typically in the range of 100-146 Hz, whereas for women it is 188-221 Hz. However, there are several challenges while using pitch period as the feature for gender identification.

First, a good estimate of pitch period is obtained only from the voiced portions of a clean non-noisy signal. Second, an overlap of the pitch values between male and female voice naturally exists, thus making it a non-trivial problem to solve.

MFCC extracts the spectral components of the signal at 10ms rate by fast Fourier transform and carries out the further filtering based on the perceptually motivated Mel scale. This decide the gender of the speaker by evaluating the distance of MFCC feature vectors and reported identification accuracy of about 98%.

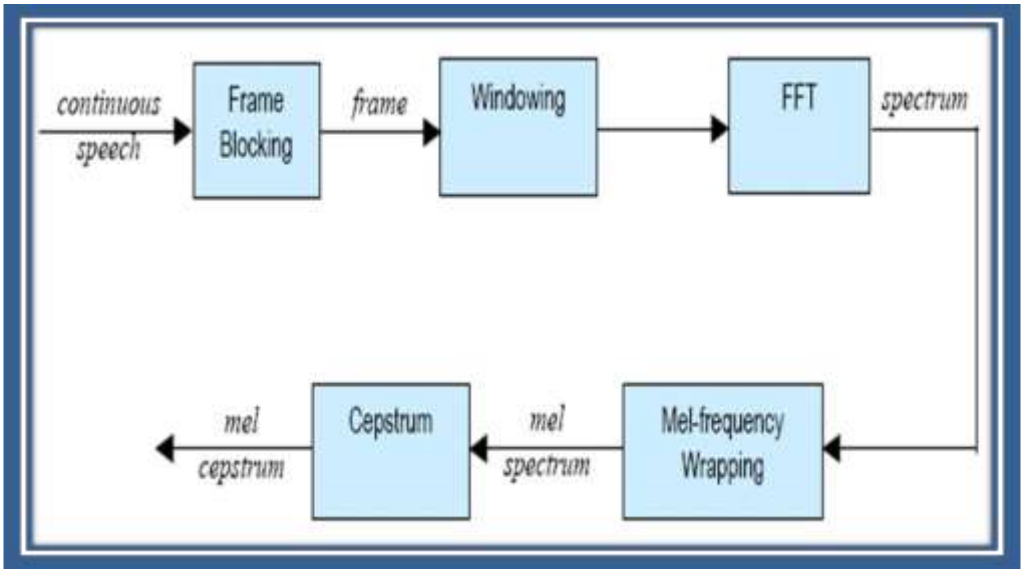

Figure 2.Block Diagram of MFCC

Noise Sensitivity

MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise. Some researchers propose modifications to the basic MFCC algorithm to improve robustness, such as by raising the log-Mel-amplitudes to a suitable power before taking DCT, which reduces the influence of low-energy components

Neural Network

An artificial Neural Network is an information processing system that is inspired by the way biological nervous system such as Brain, process information. A Neural Network is composed of large number of highly interconnected processing elements called neurons/nodes working together in order to solve specific problems. Neural Network learns by examples. Such networks are configured for a specific application such as pattern recognition or data classification through a learning process. Each connection between neurons is re presented by weight. Learning process involves adjustment of these weights based on error functions.

Neural Network possess the characteristic of deriving meaning from complicated or imprecise data and can be used to extract pattern and perform classification task that are too complex to be noticed by either humans or other computer techniques. Neural Network has an ability to learn how to do tasks based on the data given for training or initial experiences.

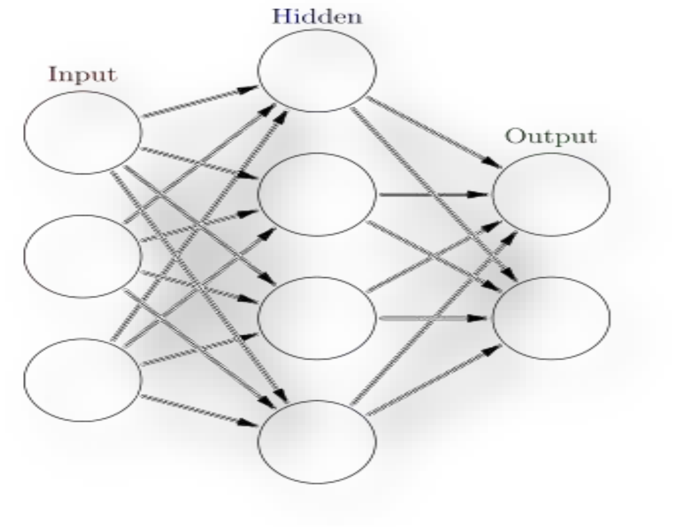

Figure 3.Structure of Neural Network

TRAINING PROCESS

Any network must be trained in order to perform a particular task. In training process, training data set is presented to the network and network's weights are updated in order to minimize errors in the output of the network. The entire set of training examples must be shown to the network many times in order to achieve satisfactory result. Back Propagation Neural Network uses back propagation algorithm for training the network. The principal advantage of back propagation is simplicity and reasonable speed. Back propagation algorithm used can be divided into following three steps.

1.Selection and Preparation of Training Data:

A good set of examples are necessary for any neural network to infer the characteristic of the input data. The best training procedure is to compile a wide range of examples (for more complex problems, more examples are required) which exhibit all the different characteristics you are interested in. It is important to select examples that do not have major dominant feature which are of no interest. It is good idea to add some kind of noise or other randomness (such as scaling factor) in some of the examples from training set. This helps to account for noise and natural variability in real data and tends to produce a more reliable network. Training set used in classification contains 60 voices out of which 30 voices are of male and 30 voices are of female. All the voices are taken in identical conditions with slight variation in distance between microphone and person.

2. Modification of Connection Weights:-

The training data set consists of input signals assigned with corresponding target (desired output). The network training is an iterative process. In each iteration weights, coefficients of nodes are modified using new data from training data set. Modification is calculated using back propagation algorithm described below.

Step 1:- Choose an input data from training set and compute output of each node in hidden and output layer using activation function.

Step 2:- Output signal of the network is compared with the desired output value (the target), which is found in training data set. The difference is calculated using mean square error and called error signal of output layer node.

Step 3:- It is impossible to calculate error signals for nodes in hidden layer directly as desired output values for these nodes are not known as in case of nodes in output layer. To calculate error signals to such nodes, idea is to propagate error signals in output layer nodes back to nodes of hidden layer that provide input to that particular node in output layer.

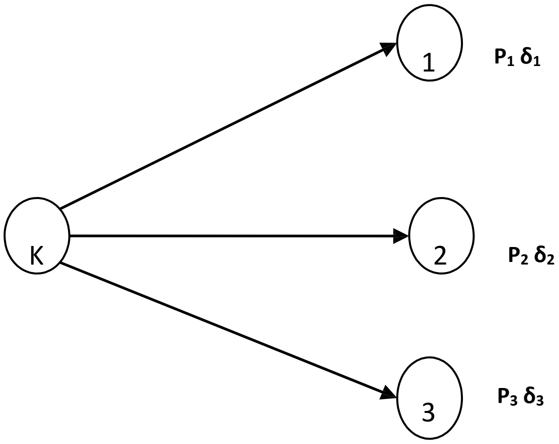

Figure 4. Calculation of error signal for nodes in hidden layer

Repetitions:
Once the above procedure is completed for all examples in, training, set, same procedure must be repeated many times until the Mean Square Error(MSE) drops below a special value. When this happens, the network is performing satisfactorily and the training session is completed.

RESULTS AND CONCLUSION

The Feed Forward Neural Network has been used. This Neural Network is the Back Propagation Neural Network with sigmoid function as neural activation function. This neural network is trained to produce output 0 for female and 1 for male.

Out of 20 samples 17 are correctly detected ,thus the efficiency of the system comes out to be 85%

REFERENCES

1. P. J. Antsaklis, K. M. Passino, and S. J. Wang, "Towards intelligent autonomous control systems: Architecture and fundamental issues," J. Intelligent Robotic Syst. vol. 1, pp. 315-342, 1989; also, a shorter version appeared in Proc. Amer. Contr. Con5 (Atlanta, GA), June 15- 17, 1988, pp. 602-607.
2. 0. Mayr, The Origins of Feedback Control. Cambridge, MA: M.I.T. Press, 1970.
3. F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, ‘‘An emotionaware voice portal,’’ in Proc. ESSP, 2005, pp. 123–131. [2] J. Luo, Affective Computing and Intelligent Interaction, vol. 137. New York, NY, USA: Springer-Verlag, 2012
4. O. Pierre-Yves, ‘‘The production and recognition of emotions in speech: Features and algorithms,’’ Int. J. Human-Comput. Stud., vol. 59, no. 1, pp. 157–183, 2003
5. D. Gerhard, ‘‘Pitch extraction and fundamental frequency: History and current techniques,’’ Dept. Comput. Sci. Univ. Regina, Regina, SK, Canada, Tech. Rep., 2003.