Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.296 – 301

RESEARCH ARTICLE

VOICE RECOGNITION

Himani Chauhan[1], Asst. Prof. Sumitra Samal[2], Ankita Ghoshal[3]

[1]Computer Science and Engineering (CSVTU), India

[2]Computer Science and Engineering (CSVTU), India

[3]Computer Science and Engineering (CSVTU), India

[1] himani.chouhan@ssipmt.com; [2] s.samal@ssipmt.com; [3] ankita.ghoshal@ssipmt.com

Abstract— Speech is the main communication between human beings. Since the time of the invention of the computer people have been trying to let the computer understand natural speech. Speech Recognition is a technology which has close connection with computer science, signal processing and intelligent systems.

Keywords— Feature extraction, Feature matching, Mel Frequency Cepstral Coefficient (MFCC)

I. INTRODUCTION

Speaker recognition is one of the method to identify the person by features of the voice[1].The task of the speaker identification is to determine the identity of a speaker by machine. To recognize the voice, the voices must be familiar both in the case of human beings as well as machines. The second component of the speaker identification is testing, the task of comparing an unidentified utterance to the training data and making the identification. The speaker of the test utterance is referred to as the target speaker[2]. It uses to provide any authentication to any system on the basis of acoustic features of voice instead of images. The behavioral aspects of human voice is used of identification by converting a spoken phrase from analog to digital format, and extracting unique vocal characteristics such as pitch, frequency, tone and cadence to establish a speaker model or voice sample[3].

A. Signal Capturing And Pre-Processing

Signal capturing is the first step that needs to be taken in voice recognition. After this a Pre-processing of signals will handle the signal processing before taking features from the signals. Normally the pre-processing of the speech signal contains pre-emphasis, framing and windowing. Normalization is also done at some point.

1. Capturing The Signal

In the processing part the first step is to capture the signal we need. The signal from a human is an analog signal. When storing the speech to a computer, the analog signal needs to be

digitized. So the first component of speech processing is speech signal measurement. When people talk to a microphone, the analog signal pressurizes the air, then the analog electric signals goes to the microphone. Analog speed digitization has two steps: sampling and quantification.

I. Sampling

The signals we got is analog signals. To process these signals in computers, we need to convert the signals to digital form which consists of digits 1 or 0. While an analog signal is continuous in both time and amplitude, a digital signal is discrete in both domains. As for a human being, the adult male will have a fundamental frequency from 85 to 180 Hz, and the frequency of a typical adult female varies from 165 to 255Hz. and children and babies have even higher fundamental frequencies. The Shannon-Nyquist sampling theorem states that: If the signal is band-limited, and the sampling frequency is higher than twice the signal bandwidth, the original continuous signal can be completely reconstructed from the samples. With lower sampling frequencies, aliasing will occur, which produces distortion from which the original signal cannot be recovered. Human speech is generally below 5 kHz, so a sampling rate of over 10 kHz is required.

II. Quantification

Quantification means that the signal that is discrete in time domain but continuous in amplitude is turned into signal that is discrete in amplitude. When doing this the amplitude values are cut into limited extent. According to the sampling precision, the samples that are in one extent will be given the same amplitude value. The dynamic range of the voice is limited by quantification. Its unit is bit.

2. Pre-Processing

I. Silence Removal

The speech signals usually contain many areas of silence or noise. The silence signal is useless for recognition, because it contains no information. And if we keep the silence signal it will make the processing signal larger and take more time and space when getting information or features from the signal. So only the signal part that contains the actual speech segments is useful for recognition.

II. *Normalization*

Normalization is a method for adjusting the volume of audio files to a standard level, as different recording levels can cause the volume to vary greatly from word to word. The recording sound samples with different volumes and possibly some DC offset, should naturally not influence the detection system. A simple way of avoiding this is to normalize the signal in some way, e.g. scaling and offsetting the signal so that it fals between levels -1 and 1. So a normalization is needed and can be applied before any other processing. The features that we will extract later on, e.g. the Mel-Frequency transform, depend on the power of the signal. This implies that speaking loudly will be seen differently than quietly. By normalizing the recording signal, this effect can be reduced [14].

III. Pre-emphasis

Usually speech signal is pre-emphasized before any further processing. By looking at the spectrum of voiced segments we can see that the energy in the voice samples distributes more in the lower frequencies than in the higher frequencies. Thus in order to boost the amount of energy in the high-frequencies, the goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Speech that comes from the mouth will have a decay of 6dB per octave, a pre-emphasis filter is used to eliminate the -6db per octave decay [15] of the spectral energy.

IV. Windowing

An audio signal is an unstable signal, meaning that the statistical properties across time are not constant. However in a very short period of time the properties can be regarded as constant. Then a short piece of signal is cut out of the whole speech signal. This is done by multiplying the speech samples with a widowing function to cut out a short segment of the speech signal. The time for which the signal is considered for processing is called a window, and the data acquired in a window is called a frame. Features are extracted once every M ms, which is called frame rate, while the window duration is N ms. Typically N is bigger than M. Thus two consecutive frames have overlapping area. The overlapping segments are used for speech analysis. The choosing of frame length and frame shift are very important, as it can have different effect on eliminating noise as well. Normally the frame length is 256, when using a sample rate of 11025 Hz and the frame shift is 128. But according to the previous research [17], ¾ frame overlap can get the best simulation result.

3. Feature Extraction

In second step the obtained voice sample is used to frame in MFCC implementation. The pre emphasized voice signal is framed in order to get the stationary part of speech. The speech signal is divided into frames of 30~20 ms with optional overlap of 1/3~1/2 of frame size. Framed speech signal is then multiplied with the hamming window in order to remove the discontinuities in the signal. Hamming window returns (8) the symmetric points of integral values framed signal into the column vector w.

$w(n,\alpha) = (1-\alpha) - \alpha\cos(2\alpha n/N-1))0\leq n\leq N-1$ (8)

where α shows different curves of hamming window. Its value usually as 0~ 0.5 and window length is L=N. It is obtained by multiplying each frame to the hamming window.

After windowing FFT is applied to convert (9) the signal into frequency domain from time domain and also used to obtain the magnitude frequency response of each frame. In doing so it is assumed that the signal in frame in periodic and continuous when wrapping around. In the opposite case of this, there are some discontinuities at the frames start and end points that causes detrimental effects in frequency response. This can be overcome by multiplying each frame by the hamming window that will help to remove discontinuities at the start and end points of frames.

Y=fft (b) (9)

where b is the windowed form of signal.

In Mel filter or triangular band pass filter the magnitude of frequency response is multiplied with the 40 number of triangular band pass filters in order to obtain the log energy of triangular band pass filter on Mel scale. These filters are equally spaced on the Mel scale (10) and use to calculate the linear frequency.

$$m = 2595 \log_{10} (1+f/700) \quad (10)$$

The frequency response on Mel scale is reflecting the similar effect of human subjective auditory perception. Triangular band pass filter is used to flatten the magnitude spectrum and to reduce the size of the features occupied. Frequency wrapping is applied here to keep the useful informational part of the Mel.
At the end by applying DCT cepstral features of voice signal are obtained (11). It is used to convert the log Mel scale cepstrum into time domain from frequency domain.
(11) where N is the length of the computed Mel frequencies. The series starts from n and k=1, Because MATLAB vectors starts from 1 instead of 0. The result is known as MFCC. These are the 40 acoustic features of human voice that are used to recognize the person depending upon the filter to be applied.

4. MFCC

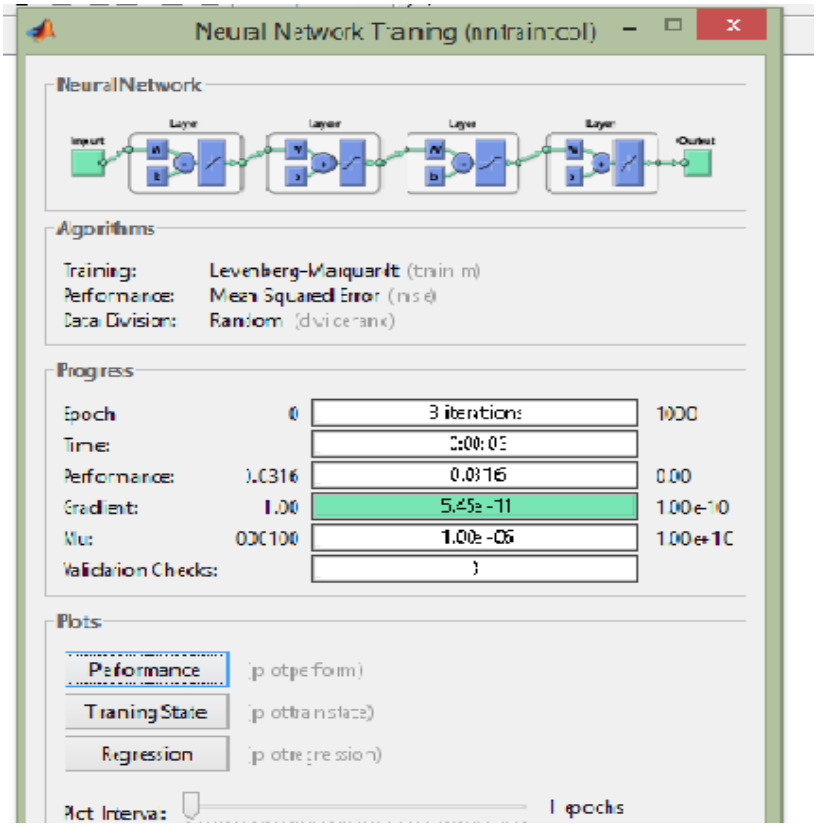
MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency cepstral these coefficient are based on the linear cosine transform of the log power spectrum on the non linear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$M = 2595 \log_{10} (1+f/700) \quad (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applied the Mel filter or triangular band pass filter to the transformed signal. Finally transformation to the discrete form by applying DCT presents the Mel cepstral coefficients as acoustic features of human voice.

5. Artificial Neural Network

Neural networks have many similarities with Markov models. Both are statistical models which are represented as graphs. Where Markov models use probabilities for state transitions, neural networks use connection strengths and functions.



In this paper speech features will be sequentially presented at neural network inputs and will be classified at the output of the networks. This process is visualized in fig. classification process in the NN.

In this paper we have used back propagation neural network[4] for the recognition system. It has been successfully applied to many pattern classification problems including speaker recognition [6].Back propagation is the generalization of the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions [7]. Properly trained back propagation networks tend to give reasonable answers when presented with inputs that they have never seen. Typically, a new input leads to an output similar to the correct output for input vectors used in training that are similar to the new input being presented [8]. The generalization property makes it possible to train a network on a representative set of input/target pairs and get good results without training the network on all possible input/output pairs [9].

B. CONCLUSION

It is concluded that the proposed research uses the technique of MFCC to extract unique and reliable human voice feature pitch in the form of Mel frequency and trained recognized using Neural Network. Research the theory of speech recognition, including its history, developing trend, signal processing, feature extraction and compression.

REFERENCES

[1] B.S. Atul, "Automatic Recognition of Speaker from there voices", proceedings of the IEEE vol. 64, PP. 460-475, 1976.

[2]G.R. Doddington, "Speaker Recognition-Identifying people by their voices , "proceedings of the IEEE, vol.73, no. 11, PP 1651-1664,1985.

[3]Ms. Arundhati S. Mehendale and Mrs. M.R. Dixit " Speaker Identification" Signals and Image Processing: An International Journal (SIPIG), vol. 2, no. 2,June 2011.

[4] Jamel Price, Sophomore Student, Dr. Ali Eydgahi, " Design of an atomatic speech recognition system using MATLAB" Cheapeaker Information Based Aerona consortium August 2005.

[5] Ehab F. , M. F. Badran , Hany Selim " Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes " Electrical Engineering Department , Assiut University.

[6] Zahorian , S.A.[1999], "Reusable binary paired partition neural network for text independent speaker identification, Proc. ICASSP-99, PP-II : 849,852.

[7] R.R Lippmann , " Review of neural network for speech recognition" neural computation , vol. 1, no. 1 , PP 1-38, 1989.

[8] Chougule , S. and Rege, P., " Language independent speaker identification " IEEE Explore PP 364-368, May 2006.

[9] A comparision of different spectral analysis models speech recognition using neural network.

