International Journal of Current Trends in Engineering & Research (IJCTER)
e-ISSN 2455–1392 Volume 2 Issue 2, February 2016 pp. 151 - 156
Scientific Journal Impact Factor : 3.468
http://www.ijcter.com

Isolated Word Recognition

Julna Nazer[1], Sajeer K(Asst Prof)[2]

[1,2]Department of Computer Science and Engineering
MES College of Engineering, Kuttipuram

Abstract—Isolated word recognition is a technical process that allows translation of spoken words into a readable text format. The objective of this work is to explore how the neural networks are used to recognize individual words in a speech. These general techniques can be developed further extended to other applications. Back-propagation neural network algorithm make use of input training samples and their respective desired output values to learn and recognize specific patterns, by modifying the activation values of its nodes and weights of the links connecting its nodes. Such a trained network is later used for feature recognition in automatic speech recognition systems. System is trained with same words with different dialects in order to identify the words that are spoken with different accents.

Index Terms—word recognizer, dialects, accent, neural network architecture, back propagation algorithm

I. INTRODUCTION

An isolated word recognition system identifies each and every single word, when uttered by humans. It is a computer driven transcription of spoken words into a readable text formant [1].It is the process of converting an acoustic signal, captured by microphone or telephone, into set of words. This transformation or translation of speech to text must be done in real time with high accuracy. One main thing that should be noted is that this type of translation should be independent of vocabulary size, external as well as internal noises and speaker characteristics. Major drawback of these kinds of systems is they're not able to understand the utterances with of same dialect pronounced differently. Sometimes the same word can be uttered differently with various situations or circumstances. So that the pitch, frequency, amplitude of voice signals varies accordingly. These changes in voice signals are very difficult to understand by a system, and speech to text conversion is quite difficult in this case. Mainly there are two types of systems: Speaker dependent and Speaker independent systems. From the application point of view, the use of word recognizer is nothing but the fact that talking is faster than typing. Automatic speech recognition consists of mainly following steps [2]:

1. Pre-processing
2. Feature extraction
3. Decoding
4. Post-processing

There are four main approaches in word recognition: The acoustic-phonetic approach, the pattern recognition approach, the artificial intelligence approach and neural network approach. Hidden Markov Model (HMM) and Gaussian Mixture models are also adopted in word recognition.

Word recognition is mainly a pattern recognition problem. Word recognition involves extracting features from the input wave signal and classifying them to classes using pattern matching model. Performance of these systems is measured on the basis of recognition accuracy, complexity and robustness. The deviation of operating conditions from those assumed during training phase may result in degradation of performance.

- **TYPES OF SPEECH UTTERED**

Separation of speech recognition system in different classes can be made based on what type of utterance they have ability to recognize.

A. Isolated speech

Isolated word recognizer usually set necessary condition that each utterance having little or no noise on both sides of sample window. It requires single utterance at a time. Often, these types of speech have "Listen/Not-Listen states", where they require the speaker to have pause between utterances. Isolated word might be better name for this type.

B. Connected word

Connected word require minimum pause between utterances to make speech flow smoothly. They are almost similar to isolated words.

A. Continuous speech

Continuous speech is basically computer's dictation. It is normal human speech, without silent pauses between words. This kind of speech makes machine understanding much more difficult.

B. Spontaneous speech

Spontaneous speech can be thought of as speech that is natural sounding and no tried out before. An ASR system with spontaneous speech ability should be able to handle a diversity of natural speech features such as words being run at the same time.

Basically speech recognition is the ability of machine or program to identify words and phrase from spoken language and convert them in to machine readable format. It is also known as Automatic Speech Recognition or computer speech recognition and speech to text conversion. This is another area of word recognition. Speech recognition is more complicated than isolated word recognition. With the rapid evolution of computer hardware and software, speech recognition technology is moderately key technology in computer information processing technology. It is widely used in voiced-activated telephone exchange, medical services, banking services, industrial control every side of society and people's lives.

II. PRESENT WORK

In speaker dependent mode, any word that is uttered must be identified correctly. Words are trained with different pronunciations so that for a system it is easy to understand the word that is uttered.

III. IMPLEMENTATION DETAILS

The present system is implemented in MATLAB R2014A. The basic unit for sentence parsing and understanding is word. In order to identify the sentences, the words must be identified properly. The following dataset which consists of 5 words that are can be uttered 6 times individually to create 6 different dialects of each word. So there will be a total of 30 dialects. These dialects or words will be identified when uttered with various accents. The table below shows the sample words with their pronunciations.

WORD	PRONUNCIATION
left	/'left/
right	/'raite/
forward	/'fowed/
backward	/backwed/
stop	/'stop/

Table 1: A sample data set of 5 words

A. Dataset Design

Using good quality recording equipment, the speech signals are recorded in a low noise environment. The signals are sampled at 8 kHz frequency. When the input data is surrounded by silence, reasonable results can be achieved in isolated word recognition. The speech signal also contains data that is unnecessary like noise and non speech, which need to be removed before feature extraction. The resulting speech signals will be passed through an endpoint detector to determine the beginning and the ending of a speech data. Here the words in the dataset are left, right, forward, backward, stop. These words can be uttered 6 times with different pronunciations.

B. Feature Extraction

The feature extractor (FE) block is designed in speech recognition in order to reduce the complexity of the problem before the next stage starts to work with data. Its aim is to use a priori knowledge to transform an input in the signal space to an output in a feature space to achieve some desired criteria. Here 13 features are extracted using LPC coefficients. The main steps in feature extraction are depicted in following figure.

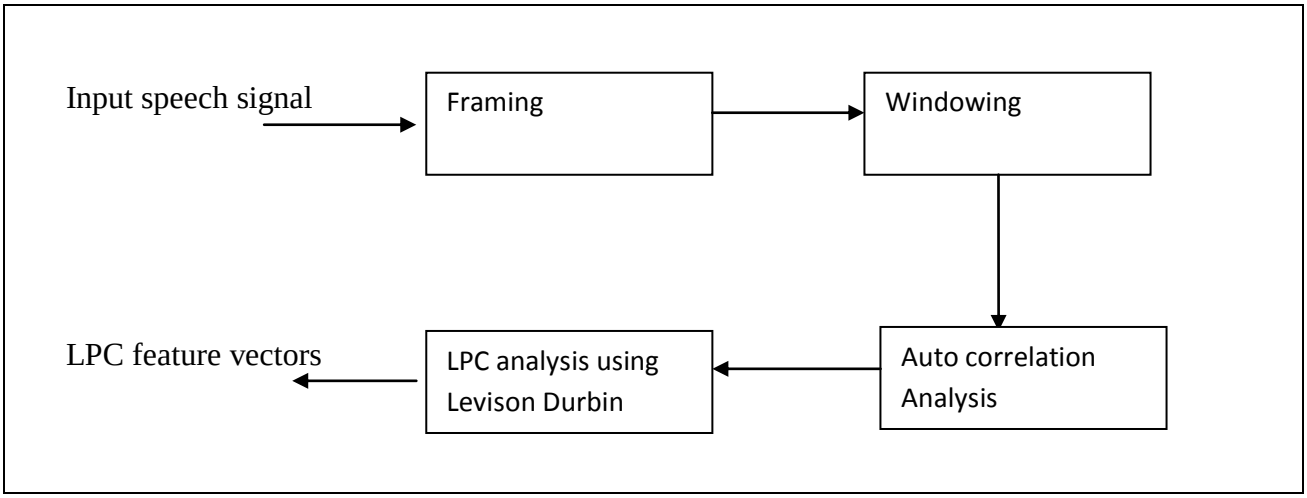

Figure: 1 Steps in feature extraction

a. Framing:

The speech signal is dynamic or time-variant in the nature. According to Rabiner (1993), the speech signal is assumed to be stationary when it is examined over a short period of time [3]. In order to analyze the speech signal, it has to be blocked into frames of N samples, with adjacent frames being separated by M samples. If $M \leq N$, then LPC spectral estimates from frame to frame will be quite smooth. On the other hand if $M>N$ there will be no overlap between adjacent frames

b. Windowing:

Each frame is windowed in order to minimize the signal discontinuities or the signal is tapered to zero at the starting and ending of each frame. A window size of 0.005 is chosen.

c. Autocorrelation Analysis:

Auto correlation analysis can be used to find fundamental frequency or a pitch of the signal. It can also be used for finding repeating patterns in a signal or identifying the missing fundamental frequency. The technique relies on finding the co-relation between the signal and a delayed version of itself.

d. LPC Analysis:

Linear predictive coefficient (LPC) is a tool used in audio signal processing and also for speech processing which shows the spectral envelope of a digital speech signal in a compressed form, with the information of a linear predictive mode [4]. The speech signal is analyzed by LPC by estimating the formants, their effects are removed and intensity and frequency of the remaining buzz are estimated. The process of removing the formants is known as inverse altering. After the subtraction of the altered modeled signal, the remaining signal is called the residue. This residue signal, the formants and the number which represents the intensity and frequency of the buzz are stored or transmitted elsewhere. The speech signal is synthesized by LPC by reversing the process using the buzz parameters and the residue in order to produce a source signal.

The function [a, e] =lpc(x, N), A = LPC(X, N) finds the coefficients, A= [1 A (2) ... A (N+1)], of an Nth order forward linear predictor.

Xp (n) = -A (2)*X (n-1) – A (3)*X (n-2) - ... – A (N+1)*X (n-N)

Such that the sum of the squares of the errors err (n) = X (n) – Xp (n) is minimized.

X can be a vector or a matrix. If X is a matrix containing a separate signal in each column, LPC returns a model estimate for each column in the rows of A. N specifies the order of the polynomial A (z) which must be a positive integer. N must be less or equal to the length of X. If X is a matrix, N must be less or equal to the length of each column of X. [A, E] = LPC (X, N) returns the variance (power) of the prediction error. LPC uses the Levinson-Durbin recursion to solve the normal equations that arise from the least-squares formulation. This computation of the linear prediction coefficients is often referred to as the autocorrelation method.

1. Feature Extraction using MFCC

The Mel frequency Cepstral coefficient (MFCC) is the most widely used features in speech recognition today. The Mel scale was developed by Stevens and Volkmann [1940] as a result of a study of the human auditory perception. It was used by Mermelstein and Davis [1980], to extract features from the speech signal for improved recognition accuracy. MFCC'S are one of the more popular parameterization methods used by researchers in the speech technology. It is capable of capturing the phonetically important characteristics of speech. The coefficients are largely independent, allowing probability densities to be modeled with the diagonal co-variances matrices. Mel scaling has been shown to offer better discrimination between phones, which is an obvious help in recognition. It has good discriminating properties. MFCC'S are based on the known variation of the human ears critical band widths with frequency to capture the phonetically important characteristics of speech, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used. This is expressed in the Mel-frequency scale. Mel –frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz.

IV. NEURAL NETWORK ARCHITECTURE

Neural network architecture is used to implement the system. A typical neural network consist set of input layers, hidden layers and output layers. Input node receives the input and forward to hidden layer. All the computations are taking place in hidden layer. After computing, results are forwarded to output layer. Artificial neural networks have a labeled directed graph structure where nodes perform some computations [5]. They consist of a set of nodes and a set of connections connecting pair of nodes. Each connection carries a signal from one node to another. Label represents the connection strength or weight indicating the extent to which signal is applied or diminished by a connection. Different choices for the weights result in different functions being evaluated by the network. Weights of the network are initially random and a learning algorithm is used to obtain the values of the weights to achieve the desired task.

A. Back propagation algorithm

Each input unit receives an input signal and broadcasts this signal to the each of the hidden units. Each hidden unit then computes its activation and sends its signal to each output unit. Each output unit computes its activation to form the response of the net for the given input pattern. During training, each output unit compares its computed activation with its target value to determine the associated error for that pattern with that unit. Based on this error, the factor δk (k = 1… m) is computed. δk is used to distribute the error at output unit back to all units in the previous layer. It is also used (later) to update the weights between the output and the hidden layer. In a similar manner, the factor δk (j= 1... p) is computed for each hidden unit. It is not necessary to propagate the error back to the input layer, but δk is used to update the weights between the hidden layer and the input layer.

B. Architecture of back propagation networks.

Back propagation is the abbreviation of backward propagation of errors. The back propagation algorithm assumes feed-forward neural network architecture. In this architecture nodes are partitioned into layers numbered 0 to 12(i.e. 12+1 lpc parameters are chosen). Here the layer number indicates the distance of a node from the input nodes. The input layer numbered as layer 0 is the lowermost layer, and the output layer numbered as layer 12 is the topmost layer. Nodes in the hidden layers neither directly receive inputs from nor send outputs to the external environment. Number of hidden layers is 6 because the dataset of words are trained 6 times individually and number of output nodes are 5. (i.e. 5 words are in dataset). C. Training the network Spoken words were recorded as six samples per word. Thus, total 30 different recordings were recorded. Then calculate LPC coefficients for all the input wave files. Here supervised learning method is used to create target vectors i.e. desired output vectors for inputs. Thus, there are 30 target vectors. The network is trained with back-propagation phase until the termination criteria is met.

V. RESULT

30 samples are used for training the network. The weights are modeled by the training phase. All the words in the data set are recognized in a low noise environment. Words in the data set are identified with various pronunciations or with different accents. The recognition rate of training phase reaches greater than 80 percent.

VI. CONCLUSION

This system shows that neural networks can act as very powerful speech signal classifiers. Back-propagation training for a multi-layer feed-forward network results in neural net- work architecture whose weights are modeled in such a way that it acts as an independent word speech recognizer. The

flow of error takes place in backwards direction, modifying the weights in such a way that the back propagation network classifiers the input samples with a reasonable accuracy. Here uses 3000 epoch of 30 samples for training the back-propagation.

REFERENCES

[1] Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno,"Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models" IEEE Transactions on Audio, speech and language processing , VOL. 23, NO. 2, FEBRUARY 2015.

[2] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, Automatic estimation of dialect mixing ratio for dialect speech recognition, IEEE in Proc. Interspeech 13, 2013.

[3] N. Hirayama, S. Mori, and H. G. Okuno, Statistical method of building dialect language models for ASR systems," IEEE in Proc. COLING , 2012.

[4] Rongfeng Su,Xunying Liu,Lan Wang "Automatic Complexity Control of Generalized Variable Parameter HMMs for Noise Robust Speech Recognition," IEEE Transactions on Audio, speech and language processing, 2015.

[5] Sajeer Karattil,"A,"A novel approach of implementation of speech recognition using neural networks for information retrieval", International journal in Science and Technology vol 8 issue 33 Dec 2015

