International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

A Review on Speech Feature Techniques and Classification Techniques

Ms. R.D. Bodke, Prof. Dr. M. P. Satone
Department of Electronic and Telecommunication
K. K. Wagh Institute of Engineering Education and Research, Nashik, Maharashtra, India

ABSTRACT
Speech Processing method is one of the important method used in application area of digital and analog signal processing. It is used in real world speech processing of human language such as human computer interface system for home, industry and medical field. It is the most common means of the communication because the information contains the fundamental role in conversation. From the speech or conversation, it converts an acoustic signal that is captured by a microphone or a telephone, to a set of words. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Speech recognition is the process of automatically recognizing the spoken words of person based on information content in speech signal. The introduces a brief detail study on Automatic Speech Recognition and discusses the various classification techniques that have been accomplished in this wide area of speech processing. The objective of this paper is to study some of the well known methods that are widely used in several stages of speech recognition system.

Keywords: *MFCC, LPC, Autocorrelation, SVM, ANN etc.*

I. INTRODUCTION
Speech recognition system is the system which taking the spoken word as an input to a computer program and converted into a text word and text format. It is the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned. Speech Recognition Systems that use training are called "speaker-dependent" systems. The speech signal is captured through microphone and processed by software running on a PC .Speech recognition technology is used in the field of robotics, automation and human computer interface applications.

There are main three Stages includes in speech recognition there is pre-processing, Feature Extraction And Classification. In this papers we studied each stage and techniques needs to apply for speech processing.

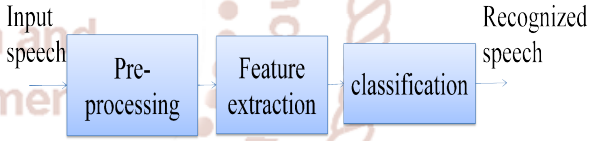

Fig1: Block Diagram Speech Recognition

This paper presents a survey on Speech Recognition and then describe some feature extraction and classifier technique. [1]. This paper presents the speech recognition is the process of automatically recognizing the spoken words of person based [2]. Efficiently tested the performance of MFCC, LPC and PLP feature extraction technique in voice[3]. Modular Neural Network (MNN) for speech recognition is presented with speaker dependent single word recognition and classification[4].The Automatic speech recognition has made great strides with the development of digital signal processing hardware and software specifically in case of speaker independent speech recognition so considering the speech feature extraction such as PLP[5].The helps in choosing the technique along with their relative merits demerits. A comparative study of different technique is done [6]. This paper is concludes with the decision on feature direction for developing technique in

human computer interface system[6].this papers give the deep study of speech recognition system using support vector machine for speech classification process[7,8].This study investigated the potential of using noisy speech training in MFCC-based speech recognition system with noise suppression robot-assisted autism therapy. [9].The ASR systems are developed employing the Gaussian mixture model-based hidden Markov model (GMM-HMM) for acoustic modeling [10].Emotion recognition from speech which is language independent. The emotional speech samples database is used for feature extraction. For feature extraction MFCC and DWT these two different algorithms are used [11].

II. Feature extraction

feature extraction is the component of a speech recognition systems. This component should derive descriptive features from the windowed and enhanced speech signal to enable a classification of sounds. The feature extraction is needed because the raw speech signal contains information besides the linguistic message and has a high dimensionality Units.

A. Mel Frequency Cepstral Coefficient (MFCC)

MFCCs are useful feature extraction techniques used in speech recognition system based on frequency plot using the Mel scale which is basically implemented the human ear scale act as filter. It concentrates on only certain Frequency component. These filters are non-uniformly space on frequency.

1.Pre-emphasis filter: Pre-emphasis filter is a very simple signal processing filter which increases the amplitude of high frequency bands and decrease the amplitudes of lower bands. In simple form it can be implemented as:

$$Y(n) = X(n) - \alpha X(n-1) \quad (1)$$

Where, $Y(n)$ is output of filter and $X(n)$ is input sampled speech signal. α the value commonly takes between 0.1 to 0.9 is filter coefficient to make stability.

2.Frame Blocking: By filtering the speech signal so we need to segment speech for further processing. So, for this we divides speech signal into sub-frame.

$$, x_2, x_3, \dots\dots\dots\dots\dots x_i \quad (2)$$

3.Windowing: In MFCC most of the Hamming window is used because is improve the Quality of speech help to increase the width of main lobe and decrease the width of side lobe. The hamming is applied to each sub frame simply remove the noise and improve the strength of speech signal.

$$w(n) = 0.54 - 0.\mathbf{46}\cos\left(\frac{2\pi(n-1)}{K-1}\right), 0 \le n \le N-1 \quad (3)$$

Where, N is total number of sample and n is resent sample.

$$Z(n) = X(n) * w(n) \quad (4)$$

Where Z(n) output of total operation.

4.FFT(Fast Fourier Transform): It is simple process in which to convert the each frame of N samples time domain to frequency plot.

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-j2\pi}{N}} \qquad k = 0, \dots\dots, N-1 \quad (5)$$

Where X_k is FFT of each frame x_n is input frame of speech.

5.Mel Filter:

and logarithmic at greater frequencies. The relation between frequency of speech and Mel scale can be established as:

$$melf = 2595 * \log(1 + \frac{f}{700}) \quad (6)$$

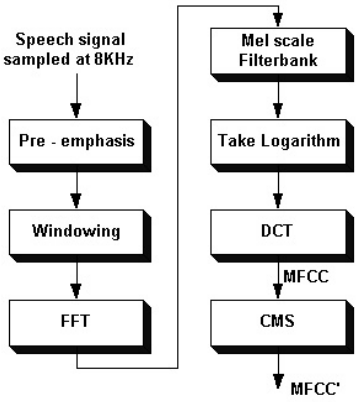

Figure 3.1. MFCC Flow

A. Pitch Detection Algorithm

A pitch detection algorithm is an algorithm designed to estimate the pitch or fundamental frequency of a quasi periodic or oscillating signal. The fundamental frequency of speech can vary from 44 Hz for low pitched male voices to 670 Hz for children or high pitched female voices .Autocorrelation methods need at least two pitch periods to detect pitch. When calculating the autocorrelation we get a lot of information of the speech signal. One information is

the pitch period. To make the speech signal closely approximate a periodic impulse train we must use some kind of spectrum flattening. To do this we have chosen to use "Center clipping spectrum flattener". After the Center clipping the autocorrelation is calculated and the fundamental frequency is extracted. An overview of the system is shown in figure 3.2.

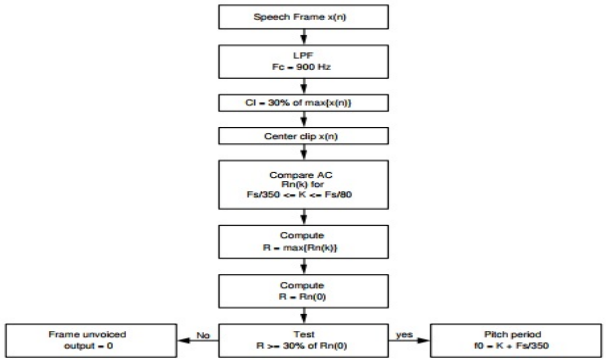

Figure 3.2 Overview of the autocorrelation pitch detector

The autocorrelation is a correlation of a variable with itself over time. The auto-correlation can be calculated using the equation.7

$$R(k) = \sum_{m=0}^{N,k-1} x(m)x(m+k) \qquad (7)$$

B. Linear Prediction Coding Method (LPC)

LP is based on speech production and synthesis models. speech can be modeled as the output of a linear, time-varying system, excited by either quasi-periodic pulse. LP provides a for estimating the parameters of the linear system. This methods have been used in control and information theory called methods of system estimation and system identification.

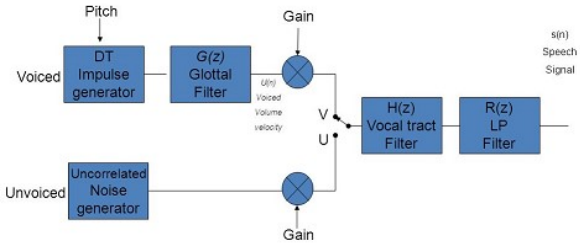

Figure 3.3 Block diagram of Linear Prediction Method

The time-varying digital filter represents the effects of the glottal pulse shape, the vocal tract IR, and radiation at the lips. The system is excited by an impulse train for voiced speech, or a random noise sequence for unvoiced speech. This all-pole model is a natural representation for non-nasal voiced speech but it also works reasonably well for nasals and unvoiced sounds.

The Basic LP Equation:

A p^{th} order linear predictor is a system of the form

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (8)$$

$$p(z) = \sum_{k=1}^{p} a_k z^{-k} \qquad (9)$$

The prediction error $e(n)$ is

$$e(n) = s(n) - \sum_{k=1}^{p} a_k z^{-k} \qquad (10)$$

The prediction error is the output of a system with transfer function:

$$A(z) = \frac{E(z)}{S(z)} = 1 - p(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = \frac{1}{H(z)} \qquad (11)$$

III. Classification

Classification technique involves studying the distinguished features and creating them in different groups depending on their feature set. Classification is the property assigned to new data set which is observed on basis of known training data set.

Artificial Neural Network

Artificial neural networks (ANNs) are statistical learning algorithms that are inspired by properties of the biological neural networks. It gives the knowledge about the behavior of dataset by increasing and decreasing with respect to vertical and horizontal computational time. In this network we need to find weight of each linkage. Let consider in figure 4.1 the inputs as $a1, a2$ and $a3$ and Hidden states as $h1$, $h2$, $h2$ And $h4$ output $o1$, $o2$. The weight between link can be denoted as,

$W(a1h1)$ is the weight link between $a1$ and $h1$ nodes.

The Following stages are follow to implement ANN Algorithm:

1. Assign a random weight to all link to start.
2. Using input node $a1$ and input hidden node $a1h2$ their linkage to find Activation rate of Hidden Nodes.
3. Using Activation Rate of Hidden node and like to Output, Find the Activation rate of Output Nodes.
4. Find the error rate of the Output Node and recalibrate the link between Hidden Nodes and Output Nodes.
5. Using the Weights and error found at Output node, cascade down the error to Hidden nodes.
6. Recalibrate weight between hidden node and the input node.
7. Repeat the process till convergence criterion is met.
8. Using the final link weight score the activation rate of the output nodes.

Let's find out the activation rate for hidden node $h1$is given by,

$$\log h1 = W(a1h1) * a1 + W(a2h1) * a2 + W(a3h1) * a3 \quad (6)$$

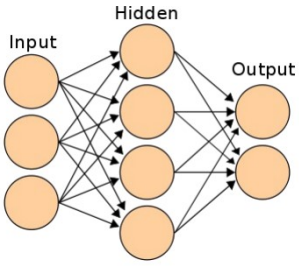

Figure 4.1 A simple Artificial Neural Network hidden layer

Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector net-works) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Let X be a test point. The Support Vector Machine will predict the classification of the test point X using the following formula:

$$f(X) = sign((w, \phi(X)) - b) \quad (4)$$

IV. Advantages and Limitations

Table 5.1 A Comparison of Feature Technique

Feature Technique	Advantages	Limitations
MFCC	Identical to the human auditory perception system. Support to the non- linear system. more accurate response	Give the weak power spectrum. Little noise present
LPC	Easy to calculate coefficient. less coefficient required to give efficient result. Easy to separate voice and unvoiced speech	Not get the linear scale of coefficient. Not give the more information.

Table 5.2 A Comparison of Classification Technique

Classification	Advantages	Limitation
ANN	Learn itself network. Easy to add new data into it. Most of used for huge data operation	More complex to implement and complicated for huge dataset.
SVM	Easy to classify the data. Get more accurate result. Mostly used for pattern recognition	It required fixed length of coefficients.

V. Summary

The performance of MFCC technique is much more efficient as compared to LPC Technique . Feature extraction methods and their advantages and limitation .LPC and MFCC are the most frequently used features extraction techniques in the fields of speech recognition and speaker verification applications. SVM and ANN are considered as the

most dominant pattern recognition techniques used in the field of speech recognition

References

1) Namrata Dave “Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition “ , International Research Journal of Engineering and Technol-ogy,2013
2) Nidhi Desai,Prof.Kinnal Dhameliya,Prof.Vijayendra Desai “Feature Extraction and Classification Techniques for Speech Recognition”,IEEE Conference Pub-lications,Depok, Indonesia, Feb. 2002.
3) Akansha Madan ,Divya Gupta “Speech Feature Extraction and Classification", International Journal,Mar 2014.
4) P Vanajakshi, M. Mathivanan ,“A Detailed Survey on Large Vocabulary Contin-uous Speech Recognition Techniques", International Conference on Computer Communication and Informatics,Coimbatore, INDIA jan-2017
5) Kennedy Okokpujie, Etinosa Noma-Osaghae, Samuel John, Prince C. Jumbo,“Automatic Home Appliance Switching Using Speech Recognition Software andEmbedded System", IEEE International Conference on Embedded sytem ,2017
6) Konrad Kowalczyk, Stanisaw Kacprzak, Mariusz Zioko,“On the Extraction of Early Reflection Signals for Automatic Speech Recognition", IEEE 2nd Inter-national Conference on Signal and Image Processing,2017
7) Abhishek Dhankar,“Study of Deep Learning and CMU Sphinx in AutomaticSpeech Recognition",IEEE Treansaction on specch processing,2017.
8) Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Speech Recogni-Tion Features Based On Deep Latent Gaussian Models",Ieee International On Machine Learning For Signal Pro-Cessing,2017
9) Yinyin Zhao ,Lei Zhu, “Speaker-dependent Isolated-Word Speech Recognition System Based on Vector Quantization ",IEEE Treansaction on specch process-ing,2017
10) Murman Dwi Prasetio,Tomohiro Hayashida ,Ichiro Nishizaki ,Shinya Sekizaki, “Structural Optimization of Deep Belief Network Theorem for Classification in Speech Recognition ",IEEE 10th International Workshop on Computational Intelligence and Applications November 11-12, 2017, Hiroshima, Japan
11) G. Dahl, D. Yu, L. Deng and A. Acero, “Context-dependent pre-trained deepneuralnetworks for large vocabulary speech recognition ", IEEE Transactionson Audio, Speech, and Language Processing, 2011.