International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Overview: Speech Recognition Technology, Mel-frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN)

Divyesh S.Mistry[#1], Prof.A.V.Kulkarni[*2]
[#]Department of Electronics and Communication,
Pad. Dr. D. Y. Patil Institute of Engineering & Technology,
Pimpri, Pune, Maharashtra India.

Abstract*—Speech recognition allows the machine to turn the speech signal into text or commands through the process of identification and understanding, and also makes the function of natural voice communication. Speech recognition involves many fields of physiology, psychology, linguistics, computer science and signal processing, and is even related to the person's body language, and its ultimate goal is to achieve natural language communication between man and machine. The speech recognition technology is gradually becoming the key technology of the IT man machine interface. The paper describes speech recognition technology for The Mel-frequency Cepstral Coefficients (MFCC) and Artificial Neural Network (ANN) and its basic model, approach, application and reviewed the classification of speech recognition systems and voice recognition technology.

***Keywords*- Speech Recognition, MFCC, ANN, Basic Model, Approach , Application.**

I. INTRODUCTION

A. Definition of speech recognition:

Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

B. Basic Model of Speech Recognition:

Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years.Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Avoinics, Automobile portal, speech transcription, Handicapped people (blind people) supermarket, railway reservations etc. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services Fig.1 shows a mathematical representation of speech recognition system in simple equations which contain front end unit, model unit, language model unit, and search unit. The recognition process is shown below (Fig .1).

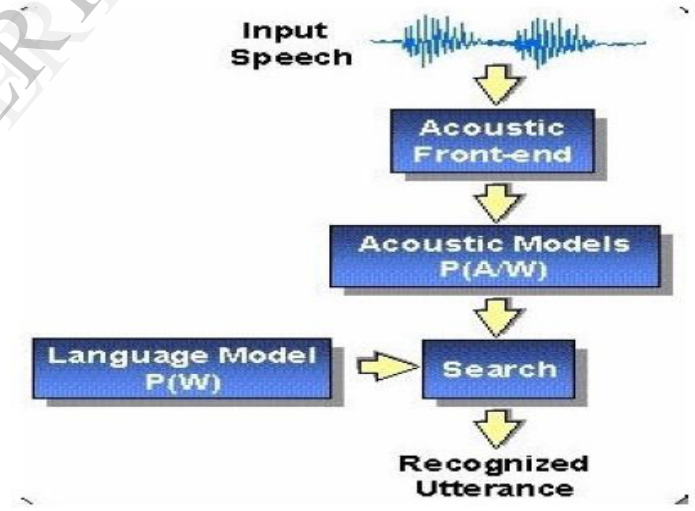


Fig.1 Basic model of speech recognition

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produces an acoustic observation sequence Y, with probability P(W,Y). The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability.

IJERT

$$P(W/A)= \arg\max_{W} P(W/A) \quad \ldots\ldots(1)$$

Using Baye s rule, equation (1) can be written as

$$P(W/A)=\frac{P(A/W)P(W)}{P(A)} \quad \ldots\ldots(2)$$

Since P(A) is independent of W, the MAP decoding rule of equation(1) is

$$W=\arg\max_{w} P(A/W)P(W) \quad \ldots\ldots(3)$$

The first term in equation (3) P(A/W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(A/W) is computed. For large vocabulary speech recognition systems, it is necessary to build statistical models for sub word speech units, build up word models from these sub word speech unit models (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods. The second term in equation (3) P(W), is called the language model. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

C. **Types of Speech Recognition:**

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following:

1. **Isolated Words.**
2. **Connected Words.**
3. **Continuous Speech.**
4. **Spontaneous Speech.**

1. **Isolated Words.**

ISOLATED word recognition is based on the premise that the signal in a prescribed recording interval consists of an isolated word, preceded and followed by silence or other background noise. Thus, when a word is actually spoken, it is assumed that the speech segments can be reliably separated from the nonspeech segments. (Clearly, in the case when there is no speech in the recording interval, a request to repeat the spoken word must be made.) The process of separating the speech segments of an utterance from the background, i.e., the nonspeech segments obtained during the recording process, is called endpoint detection. In isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely:

1) Reliable word recognition is critically dependent on accurate endpoint detection.
2) The computation for processing the speech is minimum when the endpoints are accurately located.

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

2. **Connected Words.**

In this technique, the sentence is decoded by patching together models built from discrete words and matching the complete utterance to these concatenated models. The system usually does not attempt to model word boundary allophonic effects, nor sloppy intra or inter-word articulation. There is an implicit assumption that, while distinct boundaries cannot be located among words, the words are reasonably well articulated. The accuracy of the system could be increased when probabilistic relationships among words (syntax) are known.

3. **Continuous Speech.**

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. (Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

4. **Spontaneous Speech.**

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

D. **Approaches to speech recognition:**

Basically there exist three approaches to speech recognition. They are

1. **Acoustic Phonetic Approach.**
2. **Pattern Recognition Approach.**
3. **Artificial Intelligence Approach.**

1. **Acoustic Phonetic Approach.**

The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach (Hemdal and Hughes 1967), which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighbouring sounds , it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine. The

IJERT

first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labelling.The acoustic phonetic approach has not been widely used in most commercial applications.The following table 1 broadly gives the different speech recognition techniques.

Table 1: Speech Recognition Techniques

Approach	Representation	Recognition Function	Typical Criterion
Acoustic phonetic approach	Phonemes/segmentation And labeling	Probabilistic lexical access procedure	Log likelihood ratio
Pattern recognition approach			
• Template	Speech samples, pixels & curves	Correlation, distance measure	Classification error
• DTW	Set of a sequences of spectral vectors	Dynamic warping	Dissimilarity measure
• VQ	Set of spectral vectors	optimal algorithm	Euclidian distance
• Statistical	Features	Clustering functions (code book) Discriminant functions	Classification error
Neural network	Speech features/perceptrons/ Rules/units/procedures	Network function	Mean square error
Support vector machine	Kernel based features	Maximal margin hyperplane, Radial basis function classifier(fitting functions)	Minimizing a bound on the Generalization error.
Artificial Intelligence approach	Knowledge based		Word error probability

2. Pattern Recognition approach:

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades. A block schematic diagram of pattern recognition is presented in fig.2 below.

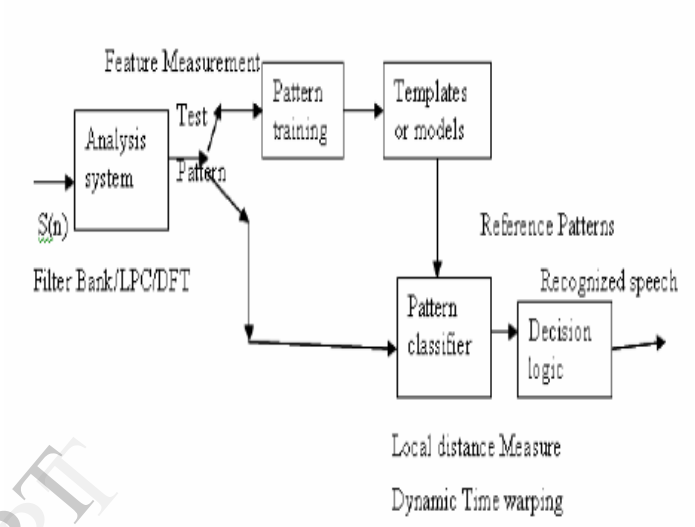


Fig:2. Block Diagram Of Pattern Recognition On Speech Recognizer.

3. Artificial Intelligence approach:

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert s speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert

IJERT

knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms Algorithms enable us to solve problems. Knowledge enable the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

E. Applications of Speech Recognition:

Various applications of speech recognition domain have been discussed in the following table 2.

Table 2: Applications of Speech Recognition

Problem Domain	Application
Speech/Telehphone/Communication Sector/Recognition.	Telephone directory enquiry without operator assistance.
Education Sector	Teaching students of foreign languages to pronounce vocabulary correctly. Teaching overseas students to pronounce English correctly.
Military sector	High performance fighter aircraft, Helicopters, Battle management, Training air traffic controllers, Telephony and other domains.
Medical sector	Health care, Medical Transcriptions (digital speech to text)
Physically Handicapped	Useful to the people with limited mobility in their arms and hands or for those with sight .
Translation	It is an advanced application which translates from one language to another.

F. Existing techniques for speech recognition have been represented diagrammatically in the following figure 3.

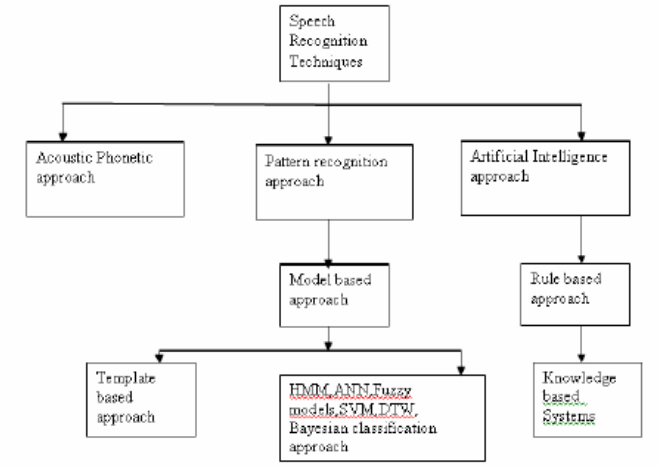


Fig: 3- Taxonomy Of Speech Recognition.

II. MFCC(The Mel-frequency Cepstral Coefficients):

A. Definition of MFCC in speech recognition:

The mel-frequency cepstrum (MFC) 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.

B. Basic Block Diagram of MFCC:

The Mel-frequency Cepstral Coefficients (MFCCs) introduced by Davis and Mermelstein is perhaps the most popular and common feature for SR systems. For speech recognition purposes and research, MFCC is widely used for speech parameterization and is accepted as the baseline. This may be attributed because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better. They are derived from a mel-frequency cepstrum (inimize-of-spectrum) where the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. The block diagram of MFCC as given in is shown in Fig.4.

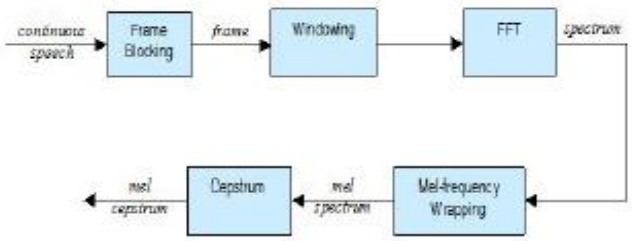


Fig. 4: Block diagram of MFCC

IJERT

C. **Feature Extraction (MFCC):**

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC is shown in Figure 5.

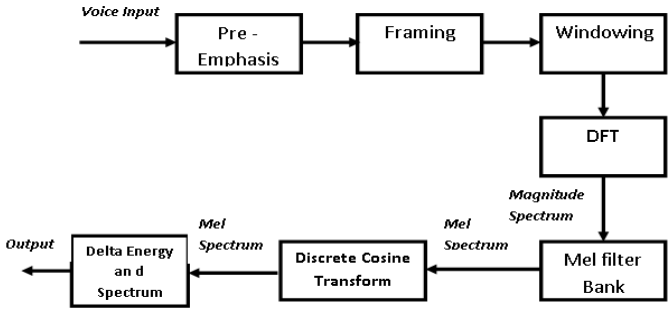


Fig. 5. MFCC Block Diagram.

As shown in Figure 5, MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis:

Pre-emphasis of the speech signal at higher frequencies has become a standard pre-processing step in many speech processing applications such as linear prediction (LP) analysis-synthesis and speech recognition. For LP analysis-synthesis systems, pre-emphasis serves a useful purpose because, at the analysis stage, it reduces the dynamic range of the speech spectrum and this helps in estimating the LP parameters more accurately while, at the synthesis stage, speech synthesised from the LP parameters representing the pre- emphasised speech is deemphasised. But, it is not clear how pre-emphasis helps in speech recognition systems.

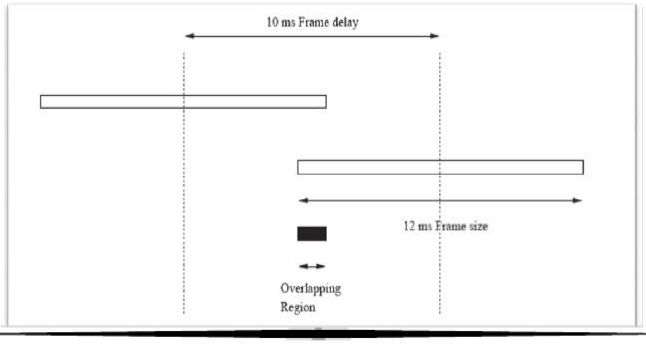
This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\,X[n-1] \quad \ldots\ldots(4)$$

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing:

Speech signal is a kind of unstable signal. But we can assume it as stable signal during 10--30ms. Framing is used to cut the long-time speech to the short-time speech signal in order to get relative stable frequency characteristics. Features get periodically extracted. The time for which the signal is considered for processing is called a window and the data acquired in a window is called as a frame. Typically features are extracted once every 10ms, which is called as frame rate. The window duration is typically 20ms. Thus two consecutive frames have overlapping areas show in fig.6.



$0.54 + 0.46cos(n\pi/m)$

Fig:6:overlapping frames.

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N).
Typical values used are M = 100 and N= 256.

Step 3: windowing:

Windowing is mainly to reduce the aliasing effect, when cut the long signal to a short-time signal in frequency domain.
There are different types of windows, there are:
• Rectangular window
• Bartlett window
• Hamming window

Out of these, the most widely used window is Hamming window. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

IJERT

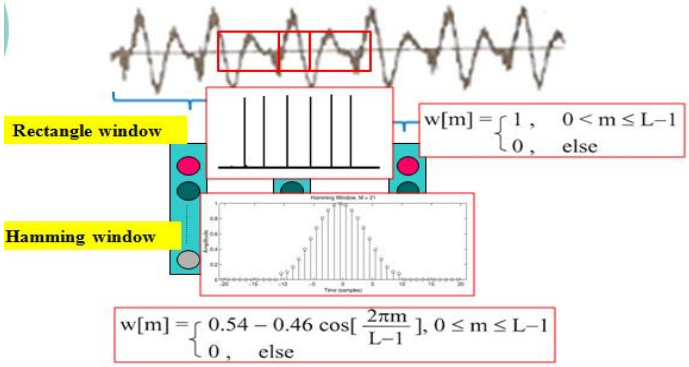


Fig: 7 :Windowing

The Hamming window equation is given as:
If the window is defined as W (n), $0 \le n \le N-1$ where

L = number of samples in each frame
Y[m] = Output signal
X (m) = input signal
W (m) = Hamming window,
Then the result of windowing signal is shown below:

$$Y(m) = X(m) * W(m) \quad \text{.....(5)}$$

$$w[m] = \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi m}{L-1}\right], & 0 \le m \le L-1 \\ 0, & \text{else} \end{cases} \quad \text{.....(6)}$$

Step 4: Fast Fourier Transform:

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(w) = FFT\ [h(t) * X(t)] = H(w) * X(w) \quad \text{.....(7)}$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 5: Mel Filter Bank Processing:

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure-7 is then performed.

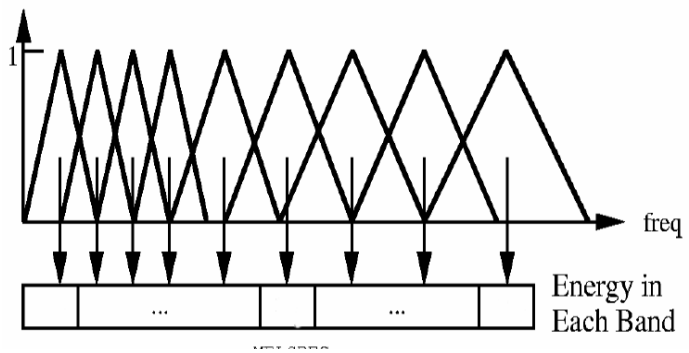


Fig. 8. Mel scale filter bank

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency F in HZ:

$$F(Mel) = [2595 * \log 10[1 + f]700] \quad \text{.....(8)}$$

Step 6: Discrete Cosine Transform:

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Step 7: Delta Energy and Delta Spectrum:

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added.

IJERT

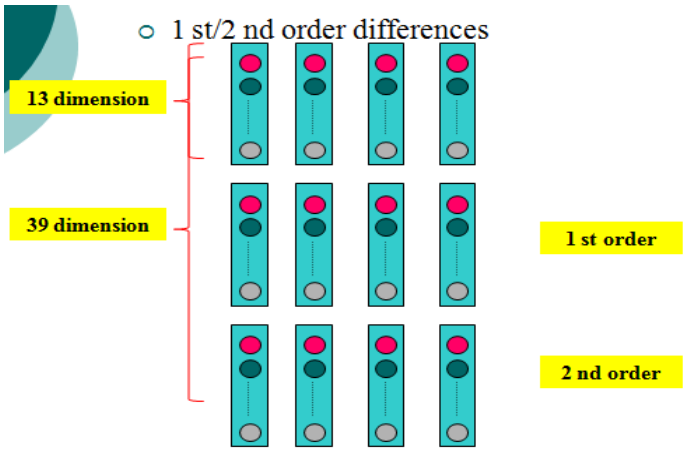


Fig: 9 :Delta coefficient

The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:

$$Energy = \sum X^2[t] \quad \text{.....(9)}$$

Each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \quad \text{.....(10)}$$

III. Artificial Neural Network (ANN):

A. Definition of Artificial Neural Network:

The simplest definition of a neural network, more properly referred to as an 'artificial' neural network (ANN), is provided by the inventor of one of the first neuron computers, Dr. Robert Hecht-Nielsen. He defines a neural network as:
"a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs."

B. Basic Model of Artificial Neural Network:

Artificial neural networks (ANNs) are inspired by the human nervous system. The human nervous system consists of approximately 1011 nerve cells, or neurons, each of which has 104 connections with other neurons . A simplified model of a biological neuron is shown in Figure-10.

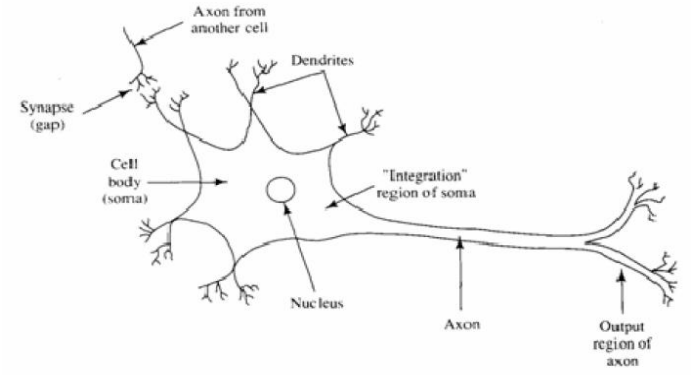


Figure-10. A simplified model of a biological neuron.

There are three main components in a nerve cell: the dendrites, the cell body (or the soma), and the axon. The dendrites are the receptive nerve fibers that carry the input signals into the cell body. The cell body sums and thresholds the received signals through the dendrites. The axon is a long transmission line that carries the signals from one cell body to others. The synapse is the connection point between an axon of a cell and a dendrite of another. The nervous system is a highly parallel structure, which is a combination of the nerve cells described above .

ANNs, which are inspired by the biological neural system introduced above, are the simplified version of the complex human nervous system, although the exact mathematical behavior of the nervous system is unknown [Hagan, Demuth, Beale, 1996]. An artificial neuron accepts signals from other neurons or from its inputs, integrates or sums the incoming signals, and then the output is determined according to some sort of threshold function. A typical artificial neuron structure is illustrated in Figure 11.

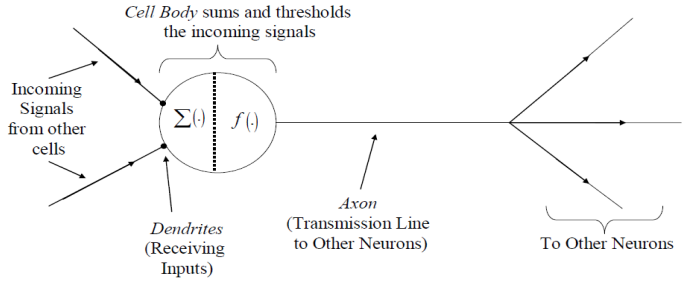


Figure-11. Artificial neuron model.

Neural network model that more commonly used and has the potentiating of speech recognition mainly include single layer perception model, multi-layer perception model, Kohonen self-organizing feature map model , radial basis function neural network , predictive neural network etc. In addition, in order to make the neural network reflects the dynamic of the speech signal time-varying characteristics, delay neural network, recurrent neural network and so on.

IJERT

C. Artificial neural network speech recognition process:

Speech recognition using artificial neural network technology, including e-learning process and the speech recognition process, shown in Figure 12. The network learning process is to known speech signal as a learning sample, self-learning neural network, and ultimately a set of connection weights and bias. The speech recognition process is to test the voice signal as network input, the recognition results obtained through the network of associations. The key of tthese two processes is to strike a speech characteristic parameters and neural network learning.

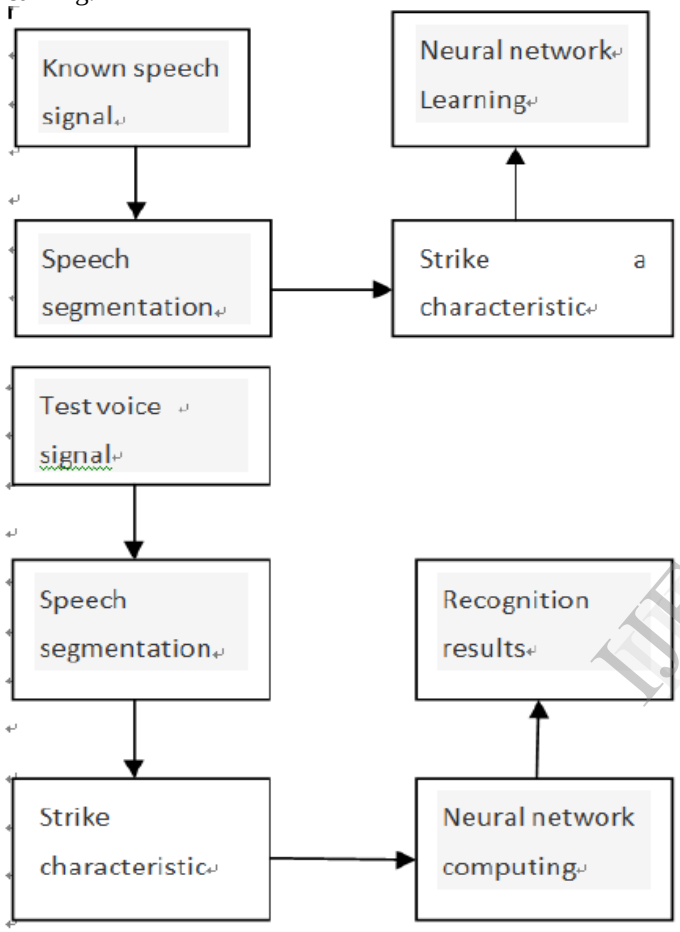


Figure12-Artificial neural network speech recognition process

The application of artificial neural networks in the field of speech recognition has been greatly developed in recent years, artificial neural networks in speech signal processing can be divided into the following areas: firstly, improve the performance of artificial neural networks. Secondly, artificial neural network has been developed method combines a hybrid system. Thirdly, explore the use of newly emerging or widespread concern mathematical methods constitute the unique nature of the neural network, and applied to the field of speech signal processing.

The application of artificial neural networks in speech recognition has become a new hotspot. Artificial neural network technology has been successfully applied to solve pattern classification problems, and was shown to have enormous energy, we can predict that in the last decade, artificial neural network-based speech recognition system products will appear in the market, people will adjust their own way of speaking to accommodate a variety of recognition system.

IV. CONCLUSION:

In this paper, we have used MFCC and Neural Network for speech recognition. The whole paper demonstrates how to use the mel-frequency cepsral coefficients and the neural network in speech recognition technology. And also demonstrates approach , application for speech recognition.

ACKNOWLEDGMENT:

I am really thankful to my guide without which the accomplishment of the task would have never been possible. I am also thankful to all other helpful people for providing me relevant information and necessary clarifications.

REFERENCES:

1) David Dean “Synchronous HMMs for Audio-Visual Speech Processing” PhD thesis, Queensland University of Technology, July 2008.

2) R. P. Lippmann, “Speech recognition by machines and humans” Speech Commun., vol. 22, pp. 1–15, 1997.

3) Vimala.C,, Dr.V.Radha “A Review on Speech Recognition Challenges and Approaches” WCSIT Vol. 2, No. 1, pp 1-7, 2012.

4) Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques”, Journal Of Computing, Volume 2, Issue 3, March 2010.

5) Alex Weibel and Kai-Fu Lee, Reading in Speech Recognition ,Morgan Kaufmann Publishers, Inc. San Mateo, California,1990.

6) B.H.Juang and S.Furui, Automatic speech recognition and understanding: A first step toward natural human machine communication , Proc.IEEE,88,8,pp.1142-1165,2000.

7) D.R.Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave , Tech.Report No.C549, Computer Science Dept., Stanford Univ., September 1966.

IJERT

8) M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," IEEE Transactions on Signal Processing, vol. 45, pp. 2673–2681, 1997.

9) F. Gers, N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," Journal of Machine Learning Research, vol. 3, pp. 115–143, 2002.

10) Oriol Vinyals, Suman Ravuri, and Daniel Povey, "Revisiting Recurrent Neural Networks for Robust ASR," in ICASSP, 2012.

IJERT