Various Techniques Used For English Language Speech Recognition: A Review

Kanchan Naithani, Ashish Semwal

Abstrac: The Recognition of speech is a process, which can be defined as understanding of human speech, processing it into a machine-readable format and utilizing it for real time applications. English (the international language), comprises of the largest vocabulary among the languages, and is mostly used for giving commands and for speech recognition in various areas. This paper descirbes a review of varoius techniques that can be used for Speech Recognition at Feature Extraction and Classification level.

Index Terms: Speech Recognition, Linear Predictive analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive Coefficients (PLP), Mel Scale Cepstral Analysis (MEL), Power Spectral Analysis (FFT), Relative Spectra (RASTA) Filtering of log domain coefficients, Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), Hidden Markov Model (HMM), Vector Quantization (VQ), Artificial Neural Networks (ANN)..

1 INTRODUCTION

The Speech Recognition is the process of converting the speech i.e. The words that are said by the speaker into appropriate text. English is used as the Standard Language at international level and is best suited to give commands or to understand words (isolated or in continuous manner) while working on an artificial intelligence system such as a speech recognition system. Hence, English Language Speech Recognition System is implemented to achieve the knowledge of the content that is said by the speaker so that work can be done in various areas like dictation, system control and navigation, commercial industrial applications, voice dialing, etc. Speech Signals are naturally occurring and hence they are random signals. Speech recognition is a very challenging problem on which a lot of work has been done [1]. The main objective of a speech recognition system is to possess capability to pay attention, perceive so act on the spoken data [2].The process comprises of three main steps, preprocessing, feature extraction and recognition. Various technologies can be used for feature extraction and recognition. However, Recognition by Hidden Markov Model (HMM) andfeature extractionmel frequency Cepstral Coefficients (MFCC) are mostly used because of their higher reliability factors [3].

2 HISTORICALSURVEY

- Detailed The primary Speech Recognition System developed at Bell Laboratories in 1952, was called AUDREY. It recognized numbers spoken by one person. Later on, IBM incontestable at the 1962 World's honest its "Shoebox" machine, that might perceive sixteen words spoken in English[4]

- In 1970's,Carnegie MellonUniversitycame out with HARPY system, which could recognize 1011 words with different pronunciation.It was vital as a result of it introduced a a lot of economical search approach, referred to as beam search, to "prove the finite-state network of possible sentences," according to Readings in Speech Recognition by Alex Waibel and Kai-Fu Lee.[4][5]

- In 1980s, new systems based on Hidden Markov Model were introducedthat were more robust than the earliertechnology.The vocabulary of Speech Recognition has been magnified from a few few hundred words to many thousand words, and it had the potential to acknowledge an infinite variety of words. [4][5]

- In 1990s,speech recognition came to masses whenDragon launched the first product for consumer speech recognition named DRAGON DICTATE. Seven years later, the much-improved Dragon Naturally Speaking was launched. the appliance recognized continuous speech, thus you may speak naturally, at concerning one hundred words per minute. However, you had to coach the program for forty five minutes.[4]
- In 2000, CMU Sphinx was developed at Carnegie financier University, which incorporates variety of speech recognizers like Sphinx two, Sphinx three and Sphinx four. They additionally created use of Hidden Markoff Model for recognition purpose.

- Later in 2000, RWTH ASR (also referred to as RASR) was launched, that contains a decoder for speech recognition. this method contained the tools needed for acoustic models development that may be utilized in recognition systems.

- In 2003, Dragon NaturallySpeaking seven Medical – by ScanSoft lowered the value of Health-care by correct speech recognition.

- In 2007, Ting applied HMM for recognizing the Malay digits also. This study conducts speech recognition experiments between separate and Continuous Density Hidden Markoff Model (DHMM and CDHMM respectively). each are

- *Kanchan Naithani, Ashish Semwal*
- *Kanchan Naithani, Phd Scholar, Department Of C.S.E, H.N.B.G.U, Srinagar Garhwal,Uttarakhand, India, Email: kanchannaithani696@gmail.com*
- *Ashish Semwal, Faculty, Department Of C.S.E, H.N.B.G.U, Srinagar Garhwal,Uttarakhand, India, Email: ash.semwal@gmail.com*

trained with totally different coaching algorithms. In DHMM, speech samples are trained by mistreatment Baum-Welch parameter re-estimation, whereas HMM is trained by segmental k-mean. DHMM returns ninety six.62% word accuracy whereas HMM achieved ninety eight.85% word accuracy.

- In 2008, Siri Inc. was supported, and Google Voice Search in iPhone was developed. Siri Inc. had noninheritable a voice with that to talk its answers, whereas earlier it had offered solely written responses, and it absolutely was deeply integrated into the iPhone so it might faucet into a couple of dozen of Apple's own tools to handle easy tasks like planning a gathering, replying to emails or checking the weather. [7]

- In 2010, Google Voice Search was launched in humanoid. Google supplemental “personalized recognition” to Voice Search on humanoid phones, so the package might record users’ voice searches and manufacture a additional correct speech model. It additionally supplemental Voice Search to its Chrome Browser in mid-2011. [6]

- In 2012, Siri launched 4s in Apple iPhone, that worked as Associate in Nursing intelligent personal agent. The assistant uses voice queries and a language computer program to try to answer queries, build recommendations, and perform actions by delegation requests to a collection of net services. The package adapts to users' individual language usages, searches, and preferences with continued use and so the came results are personalized.[5][6][7]

- The quality of speech recognition systems is additionally impromptu up, because the results are often seen within the case of Sensory'sTrulyhandsfree Voice management, which may hear and perceive the speaker even in clattering environments. [8]

3 CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

I. ISOLATED WORDS

If the words same by the speaker are unambiguously and are clearly known, they're classified underneath Isolate Words Systems. Generally, it doesn't mean that it accepts single word at a time, however one auditory communication at a time. [9][10] On one hand, this classification is appropriate for things wherever the user must provide just one word responses or commands, however on the opposite hand, it's terribly aberrant just in case of multiple words input [9][10]

II. CONNECTED WORDS

The isolated words recognition system with a distinction of permitting separate utterances to be run along with the smallest amount potential pause amongst them are referred to as Connected Words Systems [9] [10]. These systems enable dividing the sound created by the speaker into components so extract data from them by enjoying them in proximity with the minimum pause in between. [10]

III. CONTINUOUS SPEECH

Continuous speech acknowledgers recognize the content within the most natural manner potential. They embrace an excellent deal of co-articulation [9] [10]. With these systems, adjacent words are compete along with absent pauses or the other apparent division between words. because the size of vocabulary will increase, confusion between totally different words sequences additionally grows [10].

III. SPONTANEOUS SPEECH

A Speech Recognition System with spontaneous speech is in a position to handle a large style of natural speech options like words being compete along and even slight stutters. In easy words, spontaneous speech isn't practiced however is natural and offhand [9] [10]. Spontaneous speech can also embrace mispronunciations, false starts, and non-words. Systems having spontaneous speech ability ought to be ready to handle totally different words and style of natural speech feature like variable accent or words that conclude different which means once used along, and so on.

4 PROCESS OF SPEECH RECOGNITION

The Speech Recognition is additionally called Automatic Speech Recognition (ASR). it's a troublesome procedure, that contains of pattern recognition exploiting the options of the speech for higher extraction and understanding of the spoken words. The process of speech recognition begins with a speaker manufacturing associate degree auditory communication within the sort of the soundwaves captured by a mike, that are reborn into electrical signals. Then these electrical wave signals are reborn into digital form to create them understood by the speech recognition system. Speech signal is then reborn into distinct sequence of feature vectors obtained once the feature extraction. These feature vectors are assumed to contain solely the relevant info regarding given auditory communication that's necessary for its correct recognition. within the feature extraction section, the data unsuitable for proper classification like harmonic and characteristic of a mike are suppressed or eliminated. Finally, recognition part finds the most effective match within the knowledge domain (database) for the incoming feature vectors.

4.1 STEPS OF SPEECH RECOGNITION

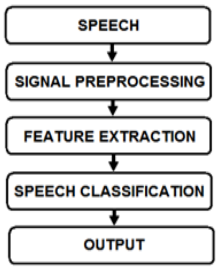


Fig. 1: *Steps of Speech Recognition*

I. SIGNAL PREPROCESSING

The first stage of the speech recognition method is pre-processing. In order for any speech recognition system to operate at a reasonable speed, the amount of data used as input must be kept to a minimum. The fundamental challenge is to remove the "bad" data, such as noise, without losing or altering, the critical data needed to identify that has been said. [13][14] Speech Analysis is done during the preprocessing phase. It is used to characterize the spectral information of an input speech signal. The speech analysis results in suitable frame size for segmenting speech signal for further analysis and extracting [12]. The speech analysis technique is done with the following three techniques: [11] [12][13].

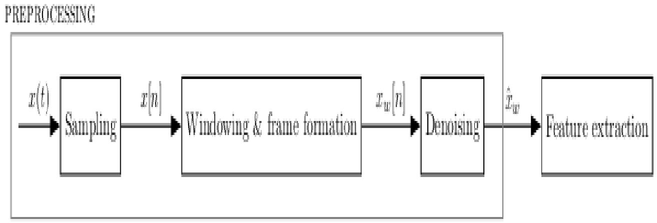
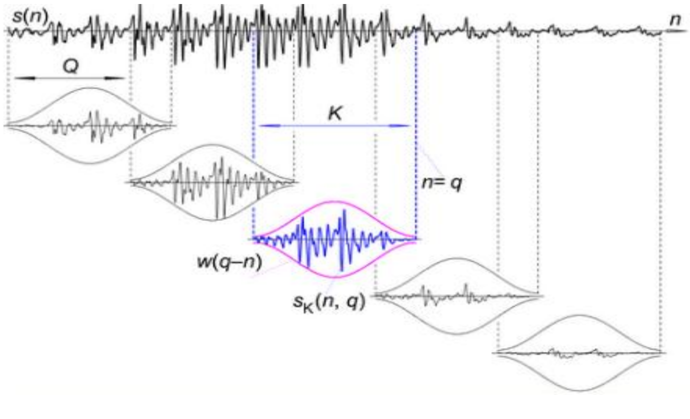


Fig.2 *Techniques of Preprocessing*

A) Sampling: Digitized signals are often processed simply by the pc so the time-continuous speech signal is required to be sampled and quantal. The output leads to a discrete-time and -value signal. o As per Nyquist-Shannon sampling theorem, a continuous-time signal x(t) with a band limit of a particular finite frequency fmax has to be sampled with a oftenness of a minimum of 2fmax. during this method, it are often reconstructed by its discrete-time signal x[n]. [13]o Studies of Sanderson, et al. have shown that the oftenness, together with the feature vector size, encompasses a direct impact on recognition accuracy.[13]

B) Windowing and Frame Formation: This stage slices the signal into separate time segments, that is completed by employing a window of n milliseconds wide and at offsets of M milliseconds long. A overacting window is often accustomed stop edge effects related to the sharp changes during a rectangular window.[14]



The Fig. 3 *Principle of Frame Formation with a windowing function*

where

- S(n) = Sampled Speech Signal
- Q = Frame Length
- K = Window Length
- q = Sampled point at the window is applied.
- and $s_k(n,q)$ = resulting short time signal with $s_k(n,q) = s(n)w(q-n)$

Windows' overlapping is seen in Fig. three wherever the principle of frame formation with a windowing perform is determined. The frame length and window length depends on the scope of application and algorithms used. In speech process, the frame length sometimes varies from five to twenty fivems and therefore the window length from twenty to 25 ms. Smaller overlapping suggests that larger time shift within the signal, so lower processor demand, however the distinction of parameter values (e.g. feature vectors) of neighboring frames is higher. On the opposite hand, larger overlapping may end up during a sander amendment of the parameter values of the frames, though higher process power is required. [13][14] C). Denoising and Speech Enhancement: This step aims to spice up the quality of the speech signals. the target is to boost the Speech Intelligence by measure the speech understandability. Noise corrupting speech signals is mike connected noise, Electrical noise, as an example, electromagnetically evoked or radiated noise, or Environmental Noise. the basic drawback of noise reduction is to cut back the external noise while not distressing the unvoiced and low-intensity noise-like elements of the speech signal itself. [14][15]

Following Algorithms is employed in order to realize Noise reduction:

•Filtering Techniques:Adaptive Wiener filtering and also the spectral subtraction strategies are the algorithms conspicuously supported filtering techniques. reconciling Wiener filtering depends on the adaption of the filter transfer operate from sample to sample supported the speech signal statistics (mean and variance). Spectral subtraction strategies estimate the spectrum of the clean signal by subtracting the calculable noise magnitude spectrum from the noisy signal magnitude spectrum whereas keeping the section spectrum of the noisy signal. [15] [16]

- Spectral Restoration: The causation of missing spectral parts of nonverbal sounds by adding noise to extend intelligence is referred as Spectral Restoration. [15] [17]

- Speech Model Based: A denoising technique that uses a harmonic noise model of the speech, assumptive that the speech signal consists of a periodic/voiced and random/unvoiced half is referred as Harmonic Decomposition. The parts are processed severally and recombined later, so the speech signal are often improved. The weights are then wont to acquire Associate in Nursing estimate of the clean speech. [15] [18]

Feature extraction technique could use Non Negative Matrix factorisation in speech process.[15]

II. FEATURE EXTRACTION

- Descriptive options are derived from the windowed and increased speech signal throughout Feature Extraction section, to alter a classification of sounds. The feature extraction is required as a result of the raw speech signal contains info together with the linguistic message and exhibits high spatial property. These characteristics of the raw speech signal wouldn't be viable to differentiate the vocalization and would end in a high error rate of word recognition. Therefore, a characteristic feature vector with a lower spatial property ought to be derived as a results of feature extraction rule, that can be used for the identification of the sounds [18].

  Following are the assorted Feature Extraction Techniques:

  • Linear Predictive analysis (LPC)
  • Linear Predictive Cepstral Coefficients (LPCC)
  • Perceptual Linear Predictive coefficients (PLP)
  • Mel scale cepstral analysis (MEL)
  • Power Spectral Analysis (FFT)
  • Relative spectra (RASTA)
  • Mel-Frequency Cepstral Coefficients (MFCC)

1. LINEAR PREDICTIVE CODING (LPC)

LPC technique of feature extraction is based on an idea that, in a series of speech samples, we make annth sample prediction, which can be denoted by the sum of the previous samples (k) target signal. The production of an inverse filter is calculated so that it corresponds to the formant regions of the speech sample [2]. LP analysis exploits the redundancy in the speech signal. The prediction of current sample as a linear combination of past p samples forms the basis of linear prediction analysis where p is the order of prediction. The predicted sample$\hat{s}$(n) can be denoted as follows:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k \cdot s\,(n-k)$$

where a$_k$s = linear prediction coefficients
s(n) = windowed speech sequence obtained by multiplying short time speech frame with a hamming or similar type of window which is given by,

$$s(n) = x(n) \cdot w(n)$$

The prediction error e(n) can be calculated by achieving the difference between actual sample$\hat{s}$(n) and the predicted sample$\hat{s}(n)$ which is given as follows:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} (a_k \cdot s(n-k))$$

The fundamental objective of LP analysis is to obtain the LP coefficients, which help in minimizing the prediction error e(n). The commonly used method for computing the LP coefficients using least squares auto correlation method. This achieved by minimizing the total prediction error. The total prediction error can be represented as follows:

$$E = \mathrm{e}^2(n)$$

The prediction The fundamental objective of LP analysis is to obtain the LP coefficients, which help in minimizing the prediction error e(n). The commonly used method for computing the LP coefficients using least squares auto correlation method. This achieved by minimizing the total prediction error. The total prediction error can be represented as follows:

$$E = \mathrm{e}^2(n)$$

2. LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS (LPCC)

Speech systems developed based on these features have achieved a very high level of accuracy, for speeches recorded in noise-free surroundings. Basic LPCC technique displays spectral features,which represent phonetic information, as they are derived directly from spectra. The energy values of linearly arranged filter banks are used by the features extracted from spectra, that follows equally emphasizing the contribution of all frequency components of the given speech signal [19].

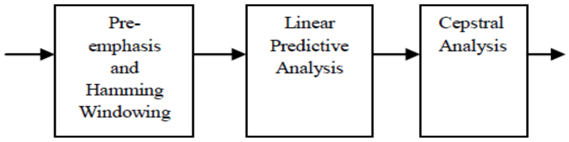

Pre-emphasis and Hamming Windowing
Linear Predictive Analysis
Cepstral Analysis

Fig. 4 *Stepsof LPCC*

Linear prediction analysis of a speech signal may provide Cepstrum. The basic idea behind linear predictive analysis is that the nth speech sample can be anticipated by a linear combination of its previous p samples as shown in the following equation: [20][21]

$$s(n) \approx a_1\, s(n-1) + a_2\, s(n-2) + a_3 s(n-3) + \cdots + a_p s(n - p)$$

Where a1, a2, a3... are assumed to be constants over a speech analysis frame and are known as predictor coefficients or linear predictive coefficients that are used to predict the speech samples. The difference of actual and predicted speech samples is known as an error and can be obtained as: [21]

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

where e(n) =is the error in prediction,
s(n) = original speech signal,
$\hat{s}(n)$ = predicted speech signal,
a$_k$s= the predictor coefficients.[21]

Later a unique set of predictor coefficients is calculated as the sum of squared differences between the actual and predicted speech samples has been minimized (process known as error minimization) is shown in the equation below: [21]

$$E_n = \left[\sum_m s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$

where m is the number of samples in an analysis frame. To solve the above equation for LP coefficients, $E_n$has to be differentiated with respect to each $a_k$and the result is equated to zero as shown below:

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad k = 1, 2, 3, \ldots, p$$

After finding the a_ks, cepstral coefficients are found using the followingrecursion.

$$C_0 = \log_e p$$

$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \text{for } 1 < m < p \text{ and}$$

$$C_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k} \text{ for } m > p$$

3. PERCEPTUAL LINEAR PREDICTIVE COEFFICIENTS (PLP)

Hermansky developed The sensory activity Linear Prediction (PLP) model. It models the human speech supported the idea of psychonomics of hearing. [22] PLP castoffs orthogonal data of the speech signals and so enhances speech recognition rate. PLP is clone of LPC except that its spectral characteristics are remodeled to match characteristics of human sensory system. [21][22]

Fig.5 *Steps of PLP*

The Power Spectrum of windowed signal is calculated as
$P(w) = \text{Re}\,(S(w))^2 + \text{Im}(S(w)^{)2}$

The PLP analysis of speech involves:

(i) Convolving short term power spectrum of speech with a simulated critical-band masking pattern

(ii) Resampling the critical-band spectrum at more or less 1-Bark intervals,

(iii) Pre-emphasis by a simulated fastened equal loudness curve,

(iv) Compression of resampled and pre-emphasized spectrum through the cubic-root nonlinearity and

(v) Approximating the compressed spectrum by the all-pole model [19]

The all-pole model coefficients are any remodeled into Cepstral Coefficients (CC). Speech signals are processed in frames of twenty fivems with a shift of ten ms. for every frame of twenty fivems, 13 PLPCC, thirteen ΔPLPCC and 13 ΔΔPLPCC are extracted by PLP analysis. [19][21][22]

4. MEL SCALE CEPSTRAL ANALYSIS (MEL)

The log spectrum on the Mel frequency scale (the Mel log spectrum) is taken into account a good illustration of the spectral envelope of speech. [23]Cepstral analysis is taken as a Cepstrum-based modulation of speech. it's outlined because the inverse Fourier remodel of the exponent of the Fourier transform module, descibed as follows: [23]

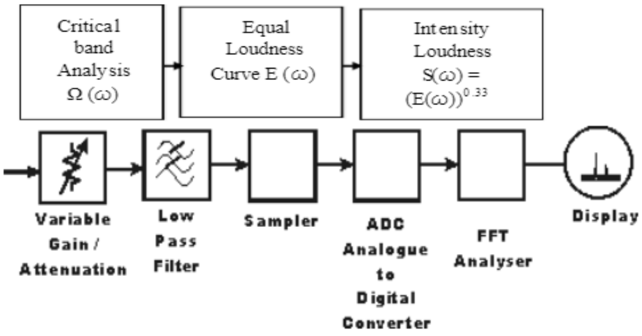
Cepstrum of signal
$= F\,\{\log[F^{-1}(\text{signal}) + j\cdot 2\pi\cdot m]\}$[24]

The log magnitude approximation (LMA) filter is applicable for the /cepstral synthesis. The LMA filter are often obtained while not remodeling the Cepstrum to associate degree impulse response, and it provides a right away synthesis from the Cepstral parameter. The cepstral vocoder mistreatment LMA filter produces a prime quality speech at a rate of three kbps. The cepstral synthesizer mistreatment the LMA filter is applicable to synthesis by rule. [23][24]

The log spectrum is taken into account an inexpensive illustration of the spectral envelope of speech. [24]

5. POWER SPECTRAL ANALYSIS (FFT)

The power spectrum of a speech signal describes the frequency content of the signal over time. acting a distinct Fourier remodel (DFT) is that the start towards computing the ability spectrum of the speech signal. It calculates the frequency data of the equivalent time-domain signal. A real-point quick Fourier remodel (FFT) are often accustomed increase potency, as a speech signal contains solely real purpose values. The computed output includes of each the magnitude and part data of the initial time domain signal. [25][26]


Critical band Analysis Ω (ω)
Equal Loudness Curve E (ω)
Intensity Loudness S(ω) = (E(ω))^0.33
Variable Gain / Attenuation
Low Pass Filter
Sampler
ADC Analogue to Digital Converter
FFT Analyser
Display

:

Fig.6 *Block Diagram of Power Spectral Analysis*

The pure FFT spectrum carries a lot of data regarding the speech signal than different trategies wont to exploit information about human sensory system. However, abundant of the additional data is found within the comparatively higher frequency bands once victimisation high sampling rates (e.g., 44.1 kHz etc.), that aren't sometimes thought of to be salient in speech recognition. The frequent use of FFT spectrum models the loudness perception

6. RELATIVE (RASTA)

The analysis library provides the flexibility to perform adherent filtering to atone for linear channel distortions. The adherent filter is used either within the log spectral or in cepstral domains. In effect, every feature coefficients is gone by the adherent filter band. Linear channel distortions seem as AN additive constant in each the log spectral and therefore the cepstral domains. The high-pass portion of the equivalent band-pass filter alleviates the impact of convolutional noise that's introduced within the channel and therefore the low-pass filtering helps in smoothing frame-to-frame spectral changes. [25]It involves temporal process and this algorithmic program for speech sweetening is simulated in MATLAB and evaluated underneath totally different condition exploitation NOIZEUS information.The original filter is restructured to realize improved performance. Real time implementation faces challenge thanks to this algorithmic program, because it is nonlinear and non-causal. [26]

6.1 sense modality Masking options

The development of sense modality masking involves concealing of 1 sound by the presence of another sound part. There are 2 totally different psychoacoustic phenomena termed as frequency and temporal masking. analysis in psychoacoustic has conjointly shown that human ear could face struggle in hearing fragile signals that fall in frequency or time neck of the woods of durable signals (as well as those superimposed in time or frequency on the masking signal, as within the higher than 2 cases). This principle of masking is exploited for noise reduction in frequency domain whereas aiming for speech sweetening.

6.2 Frequency-Domain Masking Principles

Spectral analysis of force per unit area level showing at the myringa is performed by the psychologically based mostly filters. Maskee may be a tone at some intensity that human ear tries to understand. A second tone, adjacent in frequency, makes an attempt to make noise the presence of the maskee referred to as the masquer. If strength of the maskee (relative to absolutely the level of hearing) is determined, at that it's not sonic within the presence of the masquer, it's referred to as the masking threshold of the maskee. the overall form of the masking curve for a masking tone at frequency Ω0, with a selected sound-pressure level (SPL) in decibels is shown in figure five.1. Adjacent tones with the SPL below the solid lines aren't sonic within the presence of the tone at Ω0. it's discovered that there's a variety of frequencies accessible within the masquer whose perceptibility is influenced. Tones with intensity below the masking threshold curve are cloaked by the masking tone. uneven nature is shown by the curve around Ω0.[26].

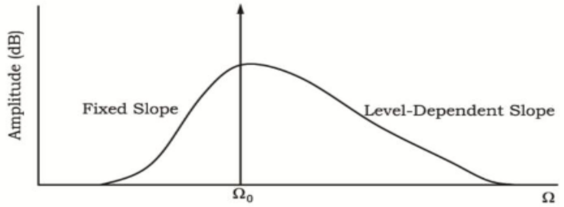


Fig. 7 *General shape of the masking threshold curve for a masking tone at frequency Ω_0*

7. MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

First mentioned by Bridle and Brown in 1974 and more developed by Mermelstein in 1976, this feature extraction methodology relies on four experiments of the human thought of words. Mel-Frequency Cepstral Coefficients (MFCC) could be a illustration of the $64000 cepstral of a windowed short-time signal derived from the quick Fourier rework (FFT) of that signal [28]. The reliable and sturdy coefficients are obtained consistent with multiple speakers and ranging recording conditions.

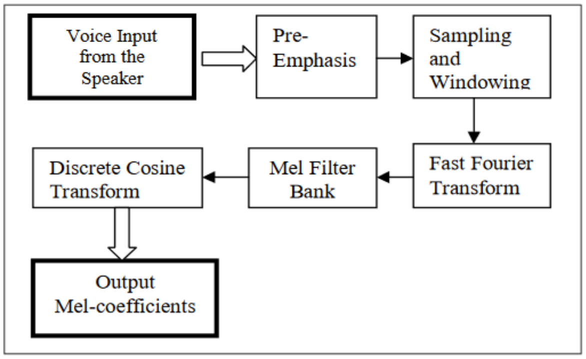


Fig. 8: *Block Diagram of MFC Coefficients*

The filter coefficients w(n) of a Hamming window of length n are computed using following formula:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1$$

= 0, otherwise

After the windowing every frame, quick Fourier Transformation (FFT) is calculated to extract frequency parts of a sign within the time-domain. [21] FFT is employed to hurry up the process. The power Mel-scaled filter bank is applied to the Fourier remodeled frame. This scale is about linear up to one kilocycle per second, and power at bigger frequencies. [29] The relation between frequency of speech and Mel scale are often established as:

Frequency (Mel Scaled) = [2595 log (1 + f (Hz)/700]

MFCCs use Mel-scale filter bank wherever the upper frequency filters have bigger information measure than the lower frequency filters, however their temporal resolutions are the identical.[21][22] The simplicity of the procedure for implementation of MFCC makes it the foremost most well-liked technique for speech recognition.

III. SPEECH CLASSIFICATION (RECOGNITION)

Finally, within the Recognition stage, the pc should determine what has been same and recognition of phonemes, teams of phonemes, and words is achieved. The word recognizer used, affectively maps the sequences of spoken vectors more or less referred to as observation vectors with the wished image sequences that are to be recognized. Methods like dynamic time deformation (DTW), Hidden Andre Markoff Model (HMM), Vector quantisation (VQ), Artificial Neural Network (ANN) are utilized in case of Representative Speech.

1. DYNAMIC TIME deformation (DTW)

Dynamic Time deformation may be a model matching speech recognition algorithmic program, that follows 2 steps:

1. part of Training: the feature vector sequence of the speech like each word in vocabulary table was extracted as a result of the guide.

2. part of Recognition: thus on check feature, vector sequences of the speech to be recognized with each guide of model library by dynamic time distortion formula, and so the result with the perfect similarity would be taken as a result of the popularity outputIn this stage, the choices of word calculated inside the previous step are compared with reference templates. DTW formula is enforced to calculate least distance between choices of word spoken and reference templets love least value among calculated scores with each model, the word is detected. the most effective alignment between double series is found by DTW formula. If just one occasion series is also "warped" non-linearly by stretching or shrinking it on its time axis [31], the vary of matching between a pair of datum is measured in terms of distance issue. Dynamic time wrapping for two voice samples is illustratedin Fig. 9. [30][31]

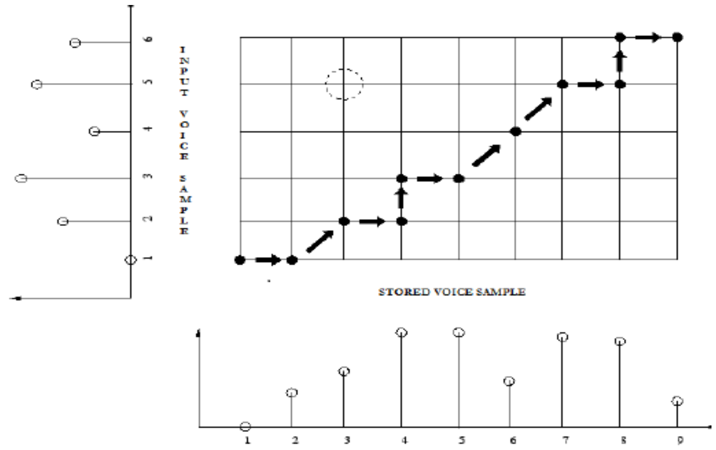


Fig. 9: *Dynamic Time Wrapping of two voice samples*

2. HIDDEN MARKOV MODEL

The Hidden Markov Model (HMM) analysis was found within the Nineteen Seventies. Later in Eighties, it had been technologically advanced and successfully applied to the acoustic signal modeling . In 1990s, HMM has in addition been the introduction of the laptop word recognition and mobile communication core technology of multi-user detection. The parameters of HMM represents the time variable characteristics of the speech signal.

HMM creates random models from glorious utterances, that were generated by every model and compares the chance that the unknown vocalization was generated by each model [32]. These random models incorporate info from many gradable data sources. the shape of the expressed word is chosen per the data of the applying domain to coach the parameters from glorious knowledge [32]. Hidden Markov's Model may be a Pattern Recognition approach that may be applied to a word, a sound or a phrase. it's assumed that the speech signal are often well characterised as a constant random process and also the parameters of the theoretical account can be determined during a precise, well-defined manner. Therefore, signal characteristics of a word can modification to a different basic speech unit as time will increase, and it indicates a transition to a different state with bound transition chance as outlined by HMM [20]. This discovered sequence of observation vectors O can be denoted by

$$\backslash O = o(1), o(2) \ldots o(T)$$

Where each observation of (t) is an m-dimensional vector, extracted at time t with

$$O(t) = [o_1(t), o_2(t) \ldots o_m(t)]$$

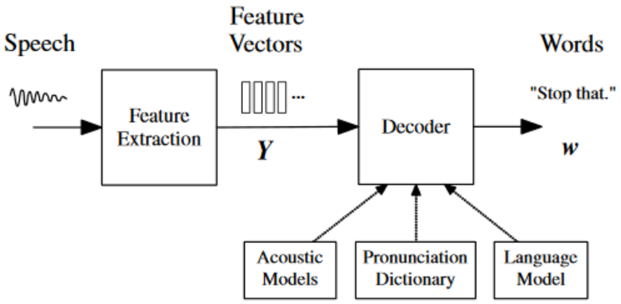


Fig. 10 *Architecture of HMM-based Recognizer*

Where each observation of (t) is an m-dimensional vector, extracted at time t with

3.VECTOR QUANTIZATION (VQ)

A classical quantity technique comes from signal method observed as Vector quantization was originally used for information compression. [35] It works by dividing Associate in Nursing large set of points (vectors) into groups having concerning the identical sort of points nighest to them. The centre of mass purpose every} cluster is used to represent each cluster, as in K-means and some totally different agglomeration algorithms. Vector quantization includes of a density matching property, that's extraordinarily powerful for large and high-dimensioned information. VQ is acceptable for lossy information compression. It will even be used for lossy information correction and density estimation. In VQ, degree ordered set of signal samples or parameters could also be expeditiously coded by matching the input vector to the identical pattern or codevector (codeword) in Associate in Nursing extremely predefined codebook [Tzu-Chuen metal, et al., 2010] [34]. The main objective of data compression achieved from vector division is to chop back the bit rate for transmission or data storage whereas maintaining the obligatory fidelity of the knowledge. The feature vector may represent sort of numerous realizable speech committal to writing parameters likewise as linear prognostic secret writing (LPC) coefficients, cepstrum coefficients.

4.ARTIFICIAL NEURAL NETWORK (ANN):

Artificial Neural Networks (ANNs) are the crude electronic models supported structure of brain. ANNs are computers having their style modelled once the brain. [36] System architectures World Health Organization vogue machines for arts psychological feature functions like speech and image pattern recognition, have developed mathematical models to imitate the style our brain performs these functions. [37] the substitute neural network models are designed as replicates of human brain organization that forever include a stratified design that encompass nodes love neurons and weights corresponding to connections between neurons. [36][37] additionally, most neural network models have some kind of "learning" rule whereby the weights are adjusted supported a series of coaching patterns. the fundamental structure of artificial neural network includes of varied inputs, that are increased by connecting weights. The add of the product of inputs and weights is fed to the transfer operate to come up with the output results. This structure will be delineated with the assistance of Fig. eleven wherever inputs are symbolized with graphic symbol i(n) and weights with w(n). The network is trained by display information ofclassification identified, that uses the information of coaching to develop to perform distribution. This successively perform is employed to estimate the chance of associate degree input pattern accessible inside many given classes. Ideally, the method may be combined with a priori likelihood of every class to work out the foremost seemingly category for a givenpattern of input. [37]

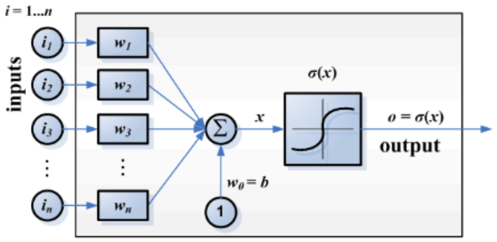


Fig. 11 *Architecture of Basic Artificial Neuron*

5 CONCLUSION

If Accurate speech recognition is achieved usingmodels which will account for the next degree of variability within the speech signals. Various experiments are carried out on various techniques in order to achieve higher variability and results have been encouraging. The results indicated that the performance of various techniques in collaboration might prove to be very useful in several real-life applications. A small value of smoothing parameter obtained after the feature extraction step is recognized and trained by the recognition system thus yielding the best possible result. The decisive aim is to progress a system which will method speeches, i.e. to grasp speakers' demands and to hold out the actions. solely a extremely correct, reliable and strong speech recognition procedure will cause the winning development of a secure voice authentication system for voice net applications.

REFERENCES

[1] Koustav Chakraborty, AsmitaTalele and Savitha Upadhya, A Research Paper on Voice Recognition Voice Recognition Using MFCC Algorithm,International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN 2349-2163, Volume 1 Issue 10, November 2014.

[2] Divya Gupta, Speech Feature Extraction Techniques A Review byShreya Narang, International Journal of Computer Science and Mobile Computing(IJCSMC), Vol. 4, Issue. 3, pp.107–114, March 2015.

[3] BhadragiriJagan Mohan, Ramesh Babu, Speech Recognition using MFCC and DTW, Advances in Electrical Engineering (ICAEE), 2014 International Conference, ISBN 978-1-4799-3543-7, DOI 10.1109/ ICAEE.2014.6838564, 19 June 2014.

[4] PC World, Speech recognition through the decades: How we ended up with Siri https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html

[5] Siri Wikipedia

[6] Wikipedia, https://en.wikipedia.org/wiki/Siri

[7] Literature Survey on Speech Recognition http://shodhganga.inflibnet.ac.in/bitstream/10603/40667/8/08_chapter3.pdf

[8] Siri Rising : The Inside Story of Siri's Origion

[9] http://www.huffingtonpost.in/entry/siri-do-engine-apple-iphone_n_249916

[10] Speech recognition through the decades https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html?page=2

[11] Prachi Khilari and Bhope V.P., A review on speech to text conversion methods,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4, Issue 7, July 2015

[12] Aman Ankit, Sonu Kumar Mishra, Rinaz Shaikh, Chandraketu Kumar Gupta, Prakhar Mathur and SoudaminiPawar,A Survey Paper on Acoustic Speech Recognition Techniques,International Journal of Recent Advances in Engineering & Technology (IJRAET),ISSN (Online): 2347-2812, Volume-4, Issue-7, 2016.

[13] Preeti Saini, ParneetKaur,Automatic Speech Recognition: A Review by International

Journal of Engineering Trends and Technology- Volume4 Issue 2, 2013.

[14] Speech Analysis and synthesis http://shodhganga.inflibnet.ac.in/bitstream/10603/25341/7/07_chapter%202.pdf

[15] Speech Recognition : Preprocessing

[16] http://recognize-speech.com/preprocessing

[17] Karpagavalli S. and Chandra E., A Review on Automatic Speech Recognition Architecture and Approaches,International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.9, No.4, 2016, pp.393-404,http://dx.doi.org/10.14257/ijsip.2016.9.4.34

[18] A.G. Maher, R.W. Kind, J.G. Rthmell, A Comparison of Noise Reduction Techniques for Speech Recognition in Telecommunications Environments. The Institution of Engineers Australia Communications Conference, Sydney, October 1992.

[19] M.A. Abd El-Fattah, M.I. Dessouky, S.M. Diab, F.E. Abd El-samie, Adaptive Wiener Filtering Approach for Speech Enhancement. In Ubiquitous Computing and Communication Journal, Vol. 3, No. 2, pp. 1-8. April 2008.

[20] R.M. Warren, K.R. Hainsworth, B.S. Brubaker, J.A. Bashford, E.W. Healy. Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps,Perception and psychophysics, 59 (2) (1997), pp. 275-283.

[21] Speech Recognition : Feature Extraction

[22] http://recognize-speech.com/feature-extraction

[23] Short time Spectral and Cepstral analysis

[24] http://cdn.iiit.ac.in/cdn/wissap.iiit.ac.in/proceedings/SCS_SRM_T1.pdf

[25] Ibrahim Patel and Y. Srinivasa Rao, Speech recognition using Hidden Markov Model with MFCC-sub band technique Recent Trends in Information, Telecommunication and Computing (ITC), 2010 InternationalConference on Date of Conference: 12-13 March 2010,Electronic ISBN: 978-1-4244-5957-5 Print ISBN: 978-1-4244-5956-8 CD-ROM ISBN: 978-0-7695-3975-1.

[26] MFCC Features

[27] https://link.springer.com/content/pdf/bbm%3A978-3-319-17163-0%2F1.pdf

[28] Feature extraction methods LPC, PLP and MFCC in speech recognition https://www.researchgate.net/publication/261914482_Feature_extraction_methods_LPC_PLP_and_MFCC_in_speech_recognition

[29] Stochi Imai, Cepstral Analysis Synthesis on the Mel Frequency Scale,Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 83.

[30] DOI:10.1109/ICASSP.1983.1172250,Conference LocationBoston, Massachusetts, USA, USA.

[31] Cepstral Analysis synthesis on the Mel frequency scale,and an adaptative algorithm for it

[32] https://pdfs.semanticscholar.org/58b0/3641b91b997ee8982f9b2cae9a15e08ff65a.pdf

[33] Techniques for feature extraction in speech recognition system : a comparative study

[34] https://arxiv.org/ftp/arxiv/papers/1305/1305.1145.pdf

[35] Feature Extraction Methods LPC, PLP and MFCC in speech Recognition https://pdfs.semanticscholar.org/0b44/265790c6008622c0c3de2aa1aea3ca2e7762.pdf

[36] Relative Spectral Analysis - RASTA http://shodhganga.inflibnet.ac.in/bitstream/10603/7432/16/16_chapter%205.pdf

[37] Namrata Daval, Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition,International Journal For Advance Research In Engineering and Technology Volume 1, Issue VI, July 2013

[38] R.V.Pawar, P.P.Kajave, S.N.Mali, “Speaker Identification using Neural Networks,” Proceeding of world Academy of Science, Engineering and Technology, Vol. 7, ISSN 1307-6884, August-2005.

[39] BhadragiriJagan Mohan, Ramesh Babu N., Speech Recognition using MFCC and DTW,Advances in Electrical Engineering (ICAEE), DOI:10.1109/ICAEE.2014.6838564 International Conference on 9-11 Jan. 2014

[40] Abdelmajid H. Mansour, Gafar Zen AlabdeenSalh, Khalid A. Mohammed, Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients AlgorithmsbyInternational Journal of Computer Applications (0975 – 8887) Volume 116 – No. 2, April 2015

[41] Speech Recognition using Hidden Markov's Model https:/www.ll.mit.edu/publications/journal/pdf/vol03_no1/3.1.3.speechrecognition.pdf

[42] Application of Hidden Markov's Model in Speech Recognition https://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf

[43] Speech Recognition using Vector Quantization through modified K-means LBG Algorithm https://pdfs.semanticscholar.org/907b/2aa3add619b0d357a9295325e3e783d91a31.pdf

[44] Mr. Amit Pathak, Mr. AchalA study on speech recognition system: a literature review Shikha Gupta, Saraf published in International Journal of Science, Engineering and Technology Research (IJSETR),ISSN: 2278 – 7798 Volume 3, Issue 8, August 2014

[45] Speech Recognition using Artificial Neural Network: A Review http://iieng.org/images/proceedings_pdf/U01160026.pdf

[46] Chee Peng Lim, Siew Chan Woo, Aun Sim Lohand Rohaizan Osman, Speech Recognition Using Artificial Neural Networks, Web Information Systems Engineering, 2000. Proceedings of the First International Conference, DOI 10.1109/WISE.2000.882421 Print ISBN 0-7695-0577-5 19-21 June 2000