Comparative Analysis of Speech Processing Techniques at Different Stages

Sameer Dass[1]
Dr. Suresh kumar[2]

Ambedkar Institute Of Advanced Communication Technologies & Research
Geeta Colony, Delhi,India
[1]dass.sameer2@gmail.com
[2]sureshpoonia@aiactr.ac.in

Abstract: The Speech Processing is turned up as one of the essential software region in virtual signal processing. There are many fields wherein speech decoding ,synthesis of speech, reputation of speech, speech recognition, speech processing and so on may be used. Speech has ability of being crucial mode of interplay with personal computer. Speech reputation is the manner of mechanically recognizing the spoken words of individual based on facts content material in speech signal. This research offers a defined and critical technological perspective and appreciation of the fundamental improvement of speech recognition. This research permits in choosing the approach along their makes use of, application and implementations. A comparative have a look at of various techniques is accomplished as in keeping with stages.

Keywords*:* Speech analysis, Speech processing, optimization, Acoustic, phonetics, Signal processing .

(Received June 1st, 2020/ Accepted June 11st, 2020)

1 INTRODUCTION

Speech recognition is substitute to conventional strategies of interacting with a computer, which includes textual enter through a keyboard. An effective machine can replace, or reduce the reliability on, well known keyboard and mouse. Each spoken line is created by using of the phonetic mixture of a set of vowel, semivowel, consonant and many other syllables. Speech recognition is the capability of a application for find out the phrases in verbal languages and convert them right into a binary signal layout. Rudimentary speech recognition software application has a restrained vocabulary of terms and terms, and it could great turn out to be privy to the ones if they're spoken very in truth. After the development of computers many scientist tried make a system which can understand the human language and gave output according to it.

In 80's the first computer arrive in the development. But there is difficulty in sampling of the voice signal. So they want to develop that system which can gave answer in spoken form. Machine recognition of speech entails producing a chain of words exceptional matches the given speech signal. There are many acknowledged applications such as translators ,virtual reality, automobile-attendants, Multimedia searches ,journey Information and reservation, native language understanding and in other Applications.

The immaturity of existing technology is related mainly with the troubles of reputation of noisy and non-stop speech.

Basic Recognition system model [1].

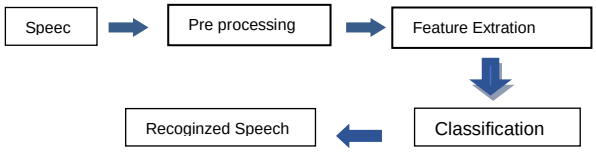

Figure 1. Recognition system model

Different languages include specific types of phoneme sets. Syllables contain one or extra phonemes, at the same time as words are fashioned with one or more syllables, concatenated to form terms and sentences.

Types of Speech Recognition system [2].

Speaking Mode: This will explain the how the word are spoken. Speaker spoke the words according to their use in sentence. They use spoke the words in continuous manner or spoke the words by taking the pause. This totally depends on the silent in between the sentences.

Speaking Style: Speech is either continuous in nature. Continuous speech is more fluent than the spontaneous one. Spontaneous speech has more flaws

with sentence formation, incorrect insertions and is like extempore as compared to well written, planned and smartly scripted speech.

Vocabulary: Vocabulary also plays an important role in speech where it is directly proportional to the error rate. Larger the vocab more is the error rate. But exceptions also exist where smaller sets of words are also prone to error.

Enrolment: There are two types of enrolments, one is speaker dependent and other is speaker independent. As the name suggests, the speaker dependent speech is of single speaker use and speaker has to provide his or her speech before using them.

The speech reputation devices are easily classified by means of the type of speech. They are continuous speech, remote phrase, linked phrase and spontaneous speech as we discussed above.

2. SPEECH EXTRACTION AND PROCESSING

There are three basic steps for speech. The purpose of speaker recognition system is to analyze, extract signify and understand statistics about the speaker context and signal [1].

2.1. Preprocessing: When we record some speech or audio there may chances of noise. The performance of a system may be reduced specially due to noise. Before supplying the speech signal into feature extraction module, the noise contained in speech signal must be removed. This task is done by the preprocessing. It removes the noise primarily based on 0-crossing charge and energy. The separation of voiced and voiceless speech based on both electricity and zero-crossing price offers the excellent result. By the help of this we can get better output

2.2 Feature Extraction Techniques: This is the segment where we extract the principle a part of speech is recognized from speech signal. This can be done by this module by lowering the Extensity of the input vector of input signal and keeping the power of signal.

There are many methods for featuring extraction. Such as

1. Mel Frequency Coefficient(MFCC).
2. Linear predictive Coding .
3. Linear Prediction Cepstral coefficients(LPCC).
4. Relative Spectral.
5. Principal Component analysis.(PCA)

2.2.1. Mel Frequency Coefficient

MFCC are the coefficients which are made up of cepstral part of the audio clips [3]. There are the frequencies of audio signals are measure on the mel scale. MFCC finds maximum elements of the audio signal of human and speech. MFCC finds the logarithmic perception of loudness and pitch of human auditory organ and attempts to get rid of speaker hooked up trends by manner of aside from the essential frequency and its tones. To represent the dynamic nature of signal in human speech the MFCC moreover consists of the exchange of a characteristic vector over time as a part of the function vector.
The Vector is calculated by functions [3]

$$CT = [\ C(4)t,j\ \ ,\Delta\ C(4)t,j\ \ ,\ \Delta\Delta\ C(4)t,j].$$

A normal MFCC function vector is probably calculated from a window with 512 pattern points and includes 13 cepstral coefficients, 13 first and 13 2nd order derivatives.
A normal MFCC function vector is probably calculated from a window with 512 pattern points and includes 13 cepstral coefficients, 13 first and 13 2nd order derivatives.

2.2.1 Linear predictive Coding [1]: LPC is an algorithm which is used in signal processing after which we can get the equations for speech processing for the representation of spectrum digital speech. We can get speech signal in a compressed form .They provide the signal in very low bits and with very accuracy.

LPC makes use of the autocorrelation approach of autoregressive (AR) modeling to find the straight coefficients. The generated clear outputs may not satisfy the process exactly even supposing the facts theoretically but series is truly verified by an autoregressive (AR) manner of the ideal order. This is due to the fact the autocorrelation method implicitly home windows the statistics, that is, it assumes that sign samples past the duration of x are 0.

2.2.3. Linear Prediction Cepstral coefficients [4] :
The cepstral coefficients obtain from both linear prediction (LP) evaluation and a filter financial institution technique which are nearly dealt with as fashionable the first signal point capabilities [1, 2]. Speech structures developed based totally on those capabilities have done a very high stage of accuracy, for speech recording it must be done in smooth surroundings. Basically, spectral capabilities represent phonetic statistics, as they're derived directly from spectra of speech signals. The capabilities extracted from spectra, the usage of the electricity values of linearly organized clear out small bump in signals, in addition emphasize the contribution of all frequency additives of a speech signal. Cepstrum can be formed from the usage of linear prediction analysis of a speech signals. The simple idea behind linear predictive evaluation is that the nth speech sample may be anticipated by means of a linear combination of its previous p samples as shown by this equation.

$$p(n) \approx a1p(n-1) + a2p(n-2) + a3p(n-3) + \bullet\bullet\bullet + aqp(n-q).$$

2.2.4. Relative Spectral: A specific band-pass filter out became brought to each frequency sub spectral in conventional PLP set of guidelines to be able to easy out quick-term noise versions and to do away with any consistent offset in the speech spectrum . The following discern shows the maximum strategies worried in RASTA-PLP it include calculating the important-band strength spectrum as in PLP, reworking of spectral amplitude by compressing a static and fixed nonlinear transformation , filtering the time orientation of every and transformed channel difficulty by way of a band skip clean out the use of equation , remodeling the filtered speech through increasing static and nonlinear changes, replicate the electricity regulation of listening to, and sooner or later calculating an version of the signal spectrum.[5].

Equation: $H(y) = (0.1)*((2+y^{-1} - y^{-3} - 2y^{-4})/(y^{-4(1-0.98y^{-1})}))$

2.2.5. Principal Component Analysis.: It is a beneficial statistical technique that has observed application in fields together with facial recognition and reputation techniques and sound compression techniques, and is a common approach for locating styles in records of excessive dimension. It is a manner of identifying patterns in information, and expressing the facts in this sort of manner as to spotlight their similarities and variations. Since patterns in records may be hard to find in information or data of high size, in which didn't have the luxury of graphical representation of signal, PCA is a powerful tool for analyzing data. The different foremost benefit of PCA is that once you've got discovered these patterns in the facts, and also you compress the records. By means of reducing the quantity of dimensions in a signal and without much loss of originality of a signal. This method used in speech compression.

3. CLASSIFICATION

The goal of modeling technique is to generate speaker signal characteristics by the use of speaker particular characteristic vector. The speaker modeling method divided into magnificence of speaker recognition and speaker identity. The speaker identity technique routinely end up aware about who's communicate me on devices or any digital medium in which the character data protected in speech signal The speaker popularity or classification is likewise divided into components which means that speaker mounted and speaker unbiased. Classification tool wants to extract speaker traits inside the acoustic signal. The critical reason of speaker identity is evaluating a speech signal from an unknown speaker to a database of analyzed signal of speaker. The machine can understand the speaker, which has been informed with a number of audio gadget. Speaker recognition also can be divide into two type techniques, 1) Text- dependent. 2) Text independent methods.

3.1. Acoustic-phonetic approach [6]: To improve the section which is based on totally speech reputation accuracy, we targeted on enhancing the section graph satisfactory by way of increasing the number of accurate segments in the segment graph and proposing extra standards for section scoring for the duration of the deciphering technique. Acoustic-phonetic information turned into applied to achieve those improvements. The older approach for speech classifications were based mostly on locating speech signals and supplying appropriate markings to the ones signal.

There are three strategies that have been carried out to the language identification. Problem telephone reputation, Gaussian combination modeling, and guide vector gadget type. Using IPA Methods we are able to find similarities for possibilities of content material dependent acoustic model for new language. There are some acoustic phonetic approach has now not extensively utilized in most business packages and software.

3.2 Pattern Recognition approach [6]: It is the approach of purely mathematic based and also represent for speech signal representation. A pattern recognition has been developed over two decade received an awful lot interest and implemented broadly too many sensible sample reputation trouble. It has two basic step pattern finding and the pattern testing. It helps in representing the signal into mathematical form. There are many examples like HMM, DTW, SVM etc. In the pattern comparison level of the technique, an immediate evaluation is made among the unknown speeches (the speech to be recognized) with every viable sample found out inside the training degree that allows you to decide the identification of the unknown signal in line with the sample signals which is already stored in system.

3.2.1 Dynamic Time Warping [7]: Dynamic time warping (DTW) algorithm is famous technique for finding a maximum suitable alignment among given signal sequences based on powerful policies or steps. DTW has been used to study considered one of a type speech patterns in computerized speech popularity. In fields together with records mining and records retrieval, DTW has been correctly carried out to routinely address time deformations and unique speeds related to time-based totally statistics. DTW became diagnosed as the maximum appropriate method for speech popularity and recognitions because of its capability to deal with extraordinary speaking speeds.

[7]An (R,S)-warping route is a series T = (t1, . . . , tL) with t = (r,s) ∈ [1 : R] × [1 : S] for ∈ [1 : L] gratifying the adjacent 3 conditions.

(i) Boundary circumstance:t1 = (1, 1) and tL = (R,S).

(ii) Monotonicity situation: n1 ≤ n2 ≤ . . . ≤ nL and m1 ≤ m2 <…..mL.

(iii) Step length condition : t_+1 − t_ ∈{(1, 0), (0, 1), (1, 1)} for _ ∈ [1 : L-1].

3.2.2 Hidden Makrov Model : Hidden Makrov Model is particular version for sequenced signal. A series model or series classifier is model whose undertaking is to assign a label or beauty to every unit in a chain. An HMM is a probabilistic collection version: given a sequence of devices . They compute a possibility distribution over feasible sequences of labels and select the top notch label collection. The semantic of the model is commonly encapsulated in the Hidden element as an instance in ASR, an HMM can be used to version a word inside the mission-based totally vocabulary, In which every nation of the hidden element represents a phoneme [8].

[9] An HMM is extraordinary with the aid of the following components: Q = p1p2 ...PN a fixed of N states A = b11 , b12, ..., bn1 ...Bnn a transition risk matrix A, every ai j representing the opportunity of transferring from united states of america i to united kingdom j, s, t, Pn j=1 ai j = 1 ∀i O = q1, q2 ,..... , qT is a chain of T observations, each one is drawn from a vocabulary V = w1, w2,..., wV, B = ci(ot) a series of assertion likelihoods, also known as emission opportunities, each expressing the opportunity of an statement ot being generated from a country i p0 ,pF a completely unique start kingdom and stop (very last) country that aren't related to observations, together with transition probabilities b01, b02,..... , b0n out of the begin state and a1F , a2F ..... AnF into the end state.

3.3 Artificial Intelligence Approach : AI method is automatically method for understand the pattern in signal. By applying this we are able to visualize the sign and examine it Knowledge based totally technique uses the data according to language, phonetics and spectrograms. While template primarily based processes were very effective inside the design of a ramification of a speech recognition structures .They furnished perception approximately human speech processing, therefore making errors evaluation or know-how-primarily based device enhancement tough. There are many examples however we have discussed most effective Time Delay Neural Network (TDNNs) and Multilayer Perceptron (MLP).

3.3.1. Time Delay Neural Networks (TDNNs) [1]: TDNN define the narrow context and close context of the deeper working of data distribution. Hence the higher layers have the functionality to study wider relationships. Each layer in a TDNN operates at a extraordinary temporal resolution, which will increase as we go to higher layers of the community.

Further, all through lower back-propagation, the lower layers of the network are updated via a gradient gathered over all of the time steps of the enter temporal context. Thus the decrease layers of the network are pressured to study translation invariant feature transforms. The input layer has been enlarged to just accept as many enter styles as the constant collection period to be processed at whenever step. There are three layers input, hidden and output layers. The enter vector enters the community from the leftmost set of enter neurons. At whenever step the inputs are shifted to right through the unit time put off. The outputs of the enter layers are feed to the right maximum of the hidden layer and process for all of the subsequent layers. Sequentially we get the result form output layer.

We explain it by figure below.

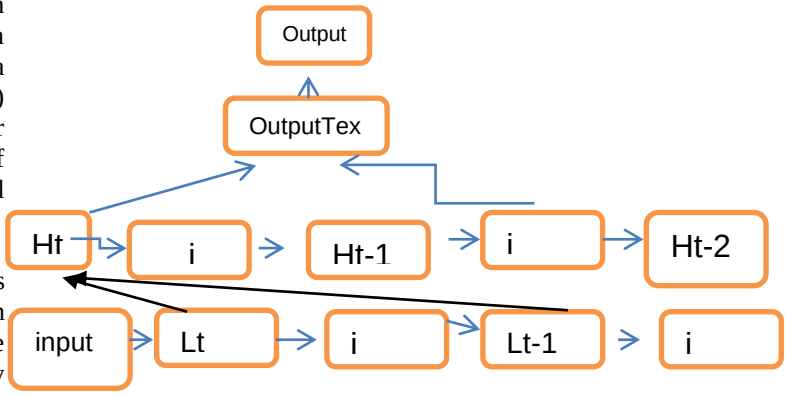

Figure2: Time Delay Neural Network [8].

3.3.2. Multilayer Perceptron (MLP)[10]: An MLP is a community of easy and disbursed neurons called perceptron. The easy idea of a character perceptron modified into delivered through the usage of manner of Rosenblatt in 1958. The perceptron finds unlinked output of a couple of actual-valued inputs with the aid of forming a linear mixture in step with its input weights after which in all likelihood placing the output through some curved activation function.
[10] Equation is given:

$$y = \varphi\left(\sum_{i=1}^{n} \omega ixi + b\right) = \varphi(w^{Tx} + b)$$

Where W denotes the vector of weights, x is the vector of inputs; b is the prejudice and is the activation feature. The center is given from left aspect of enter neurons. The output of enter neurons are extended with weights and feed as input to hidden neurons. Same step of operation is finished for hidden and output layer.

After that we have to design such speech engine that we can recognize the speech by whole sentence matching or sub word matching. We should take care of the languages, speech style and many other factor .So that we should train our machine by this algorithm as we discussed above according to our use.

4 ANALYSIS OF FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES

We have gone through the literature from various methods as explained in section 2 to 3 .We find various techniques like Mel Frequency Coefficient, Linear predictive Coding, Linear Prediction Cepstral coefficients and Relative Spectral Principal Component analysis and for classification technique we discuss Acoustic-phonetic approach, Pattern Recognition approach and Artificial Intelligence Approach . Based on the available method and literature we have found that there are some differences between them, we analyzed the some factors which is explained in Table 1 and Table 2

Table 1. Featuring's Extraction Technique's

Sr. No.	*Method of Feature Extraction*	*Uses*	*Properties*	*Implementation*
1	Mel Frequency Coefficient	Speech Recognition and Audio Retrieval.	Fourier remodel calculations	Subjective frequency scale referred to as Mel Scale
2	Linear Predictive Coding	Encoding for good quality speech at a low bit rate, Used in Shorten, MPEG , FLAC,ALS, SILK formats.	Formants and poles required , Spectral envelopes are form	Vocal Tract Components signal Represented By Lp coded Filter and Residual
3	Linear Prediction Cepstral coefficients	Noise cancellation, Reverberation suppression and echo cancellation.	Separation of the excitation signal	By the characteristic extraction of excited sign
4	Relative Spectral	Noise cancellations.	Filtering the signals	By the characteristic extraction of excited sign
5	Principal Component analysis.	Speech covariance analysis and neuroscience	Exploratory Data Analysis and Eigen vector calculation need	By calculation of eigen values and Gaussian vector

TABLE2. CLASSIFICATION TABLE.

Sr.no	*Method of classification*	*Techniques*	*Uses*	*Properties*
1	Acoustic Based		Multilingual Sounds ,Phonetics ,Speech Activity	Easily Classify Vowels, Differentiate between Phonetics
2	Pattern recognition Approach	Hidden Makarov Model	Speech Analysis, Speech Tagging, Time series Analysis	Allow Continuous Space, Generative model,
		Dynamic Time Wrapping	Spoken word Recognition, Signature Recognition	Series Alignment, Supervised learning
3	Artificial Intelligence Based Approach	Time Delay Neural Networks	Speech Recognition Using phoneme Detection	Define Signal with time, Function approximation
		Multilayer Perceptron	Solve problems stochastically, Fitness Approximation	Deploy arbitrary Functions, Binary classifications

5. PERFORMANCE

For [11] The Performance of system is defined by the word error rate .The WER is subsitute levenstein distance is improvised version of phoneme degree.It formula is
WER = Substitution + Deletion + cost of insertion / Counts of references in the phrase,
Or

WER = S + D + I / N ;
Lower the WER is much we expected from our Speech Recognizing system.

6. CONCLUSION

This Research discussed about the various Technique for speech processing . There are many which can easily classify the signal according to the environment of speech in which it is spoken. But we have to achieve the accuracy in every environment .So that we conclude that HM model and MFCC algorithm is best for speech processing.

REFERENCES

[1.] Mayur R Gamit, Prof. Kinnal Dhameliya, Dr. Ninad S. Bhatt3 , *Classification Techniques for Speech Recognition: A Review, International Journal of Emerging Technology and Advanced Engineering, page 55-63, 2015.*

[2.] Parneet Kaur, Parminder Singh, Vidushi Garg, *Speech Recognition System; Challenges and Techniques. International Journal of Computer Science and Information Technologies , Volume (3), pages 3989-3992, 2012.*

[3.] Lehrstuhl fur Inf. VI, Rheinisch-Westfalische Tech. Hochschule Aachen, *Computing Mel-frequency cepstral coefficients on the power spectrum, International Conference on Acoustics, 2001.*

[4.] E. Wong, *Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification, Proceedings of 2001 International Symposium on Intelligent Multimedia,2002.*

[5.] P. Prithvi1, Dr. T. Kishore Kumar , *Comparative Analysis of MFCC, LFCC, RASTA – PLP, International Journal of Scientific Engineering and Research (IJSER),Page 1-7,2015.*

[6.] Bharti W.Gawali , *A Review on Speech Recognition Technique. International Journal of Computer Applications, pages 16–24, November 2010.*

[7.] Chotirat Ann Ratanamahatana, *Everything you know about Dynamic Time Warping is Wrong,2014.*

[8.] Xian Tang, *Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition, IEEE Pacific-Asia Conference on Circuits, Communications and Systems, 2009.*

[9.] *Https:// Web.stanford.edu/~jurafsky/slp3/9.pdf.*

[10.] *Https://www.hiit.fi/u/ahonkela/dippa/node41.htm*

[11.] Dat Tat Tran, *Fuzzy Approaches to Speech and Speaker Recognition, 2001.*