Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020)
IEEE Xplore Part Number:CFP20F70-ART; ISBN:978-1-7281-4685-0

A Survey on Speech Feature Extraction and Classification Techniques

Nivetha S
PG Scholar - Department of Computer Science and Engineering
Dr. Mahalingam College of Engineering and Technology
Pollachi, Tamil Nadu, India
nivethashanmugam8@gmail.com

Abstract

Speech recognition is an approach of acknowledging human speech with the aid of the system and to produce string output in written format. A model is positioned from a crew of audio recordings whose corresponding transcripts are created with the resource of taking recordings of speech as audio and their textual content transcriptions, the use of software program to create statistical representations of the sounds that would make up every phrase famous via incorporating Language Processing (NLP) methods. For several decades, researchers are working in the field of speech recognition and communication. This paper describes some of the techniques and approaches that are developed by various researchers in the field of speech recognition.

Keywords — Automatic Speech Recognition (ASR), Classification techniques, Feature Extraction techniques.

I.INTRODUCTION

Speech recognition is the way of mapping a waveform into a content which ought to be comparable to the data to be passed on by the expressed word[1]. Significant uses of NLP are machine interpretation and programmed speech acknowledgment. A portion of the potential outcomes incorporates sub phoneme units, biphones, diphones, dyads or transemes, triphones, demisyllables, entire words and expressions.

ASR for Indian dialects is nonetheless at its earliest degrees whereas western dialects like English and Asian dialects like the Chinese are in a similar way very much developed [11]. Consequently progressing decades shows developing enthusiasm for this field and furthermore enormous extension for research work.

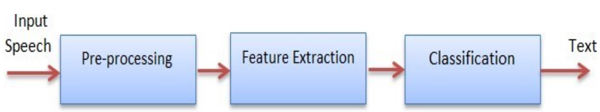

Fig 1:Basic block diagram of Speech Recognition

II.CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS

Speech recognition [2] can be grouped in a few unique classes dependent on the sort of speech expression, kind of speaker model, kind of channel and the sort of speech that they can perceive.

1. Based on Speech Expression

An expression is the vocalization (talking) of a word or phrases that speak to an outstanding value to the computer system. Expressions can be a word, a couple of words, a sentence, or even one of the kind sentences. Types of speech expression are:

A. Unique Words:

Unique word recognizers regularly require every utterance to have quiet on each aspect of the pattern window. It is comparatively less intricate and less complicated to put in force due to the real word boundaries are accessible and the phrases have a tendency to be clearly reported which is the most essential benefit of this type.

B. Joined Words:

Joined word frameworks (all the more effectively "associated articulations") are like confined words, alternatively enable separate expressions to be "run together" with a negligible extend between them.

C. Consecutive Speech:

Consecutive speech recognizers enable clients to talk normally, while the system essentially decides the substance. Fundamentally, it's system correspondence. It incorporates a lot of "co-articulation", where nearby words run together without any delays.

D. Unconstrained Speech:

This kind of speech isn't practised, however normal. Unconstrained (and unrehearsed) sounds may likewise incorporate errors, false-begins, and non-words.

2. Based on Speaker type

All speakers have their brilliant voices, because of their unique bodily physique and character. Speech awareness is comprehensively ordered into two predominant gatherings based on speaker fashions to be speaker dependent and speaker independent.

A. Speaker based models

Speaker based models are supposed for a precise speaker. They are oftentimes more specific for the specific speaker, however, considerably less genuine for one of a kind speakers.

B. Non-Speaker based models

It perceives the speech examples of a huge gathering of individuals. This sort of framework is hardest to create, most costliest and offers less precision than speaker dependent models.

III.RELATED WORKS

TABLE I: ANALYZING VARIOUS SPEECH RECOGNITION TECHNIQUES BASED ON DATASET, FEATURE EXTRACTION AND RECOGNITION APPROACH

S.No	ANALYZING VARIOUS SPEECH RECOGNITION TECHNIQUES BASED ON DATASET, FEATURE EXTRACTION AND RECOGNITION APPROACH				
	Research Work	**Dataset**	**Feature Extraction Technique**	**Recognition Technique**	**Accuracy**
1	Continuous Telugu Speech Recognition through Combined Feature Extraction by MFCC and DWPD Using HMM based DNN Techniques[8]	Speech database is created by recording the speech signal in a silent environment	Combination of MFCC and DWPD	DNN based HMM is Hybrid architecture of senones and can be effortlessly suitable for non-stop speech. Training with senones allows greater facts to be represented in the training network	91.89 %
2	Speech Recognition using SVM[13]	Experiments are conducted for speech recognition audio using Television broadcast speech data collected from Tamil news channels using a tuner card.	MFCC	SVM constructs linear model primarily based upon assist vectors in order to estimate decision function.	78%
3	Speaker Identification based on Hybrid Feature Extraction Techniques[22]	The dataset involves various sounds download from (http://www.voxforge.org)	1.Discrete wavelet transform + PCA, 2.DWT +curvlet+PCA	Neuralnetwork(Back propagation algorithm)	1.85% 2.87.6%
4	End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition[24]	American English vowels dataset	MFCC and PLP	Error rate reduction can be done by using Convolutional neural network	70% Up to a 15% relative error reduction over the HMM
5	Isolated Pali Word (IPW) Feature Extraction using MFCC & KNN Based on ASR[14]	The 'IPW' speech database was developed at Natural Sounding Speech Recognition and Speech Synthesis Lab	MFCC	KNN classifier	80.36%
6	Audio-visual feature fusion via deep neural networks for automatic speech recognition[15]	The CUAVE audio-visual database of isolated and connected digits	MFCC	Discriminatively-tuned bimodal deep autoencoder	80.1%.
7	Deep Neural Network based Place and Manner of Articulation Detection and Classification for Bengali Continuous Speech[16]	1. The corpus for continuous spoken Bengali speech is collected from CDAC speech corpus 2. A subset from TIMIT speech corpus is selected	MFCC	DNN (Deep neural network)	70%
8	Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach [19]	Small vocabulary which consists of 15 speakers voices.	MFCC	GMM (Gaussian mixture model)	85%
9	Spectral feature extraction techniques for speech recognition[5]	The database contains 2000 samples made up of utterances by two hundred speakers uttering ten digits	Combination of MFCC and LPC features	SVM constructs linear model primarily based upon assist vectors in order to estimate decision function.	94%

10	Speaker Identification & Verification Using MFCC & SVM[6]	TIMID database was used as a corpus of labeled speech data	MFCC	SVM constructs a linear model primarily based upon assist vectors in order to estimate decision function.	83%
11	Neural network based voiced and unvoiced classification using Egg and MFCC feature[7]	Small vocabulary Speaker independent Isolated word	MFCC	Neura lnetwork(Back propagation algorithm)	87.3%
12	Speech Recognition System with Different Methods of Feature Extraction[17]	Small vocabulary Speaker independent Continuous speech	1. MFCC 2. LPC	ANN is used for pattern matching.	1.80.3% 2.75.33%

IV. FINDINGS OBSERVED

1. FEATURE EXTRACTION

It is an enormous part of ASR systems that are used in analyzing the given sample and place the extracted information[10]. The spoken words are identified at once from the digitized waveform. As speech signals appear to be non-stationary in nature, some structure of statistical representations have to be carried out to limit speech signal variability and this can be completed by performing feature extraction. In addition, the feature extraction process becomes very difficult due to various constraints involved in the speech input. They are

1. Speech signal differs for a given word between speakers
2. Copy of words by same speaker
3. Intonation will vary between speakers
4. Changes in speech production will produce variability To solve the above constraints, a good feature

extraction technique should be capable of identifying specific properties that are more relevant to the linguistic content. Also, it should discard all other irrelevant information like background noise, channel distortion and emotion etc.

Various feature extraction techniques[18] have been developed to extract spectral features from speech and most commonly used techniques are

1. Mel Frequency Cepstral Coefficients (MFCC)
2. Linear Predictive Coefficients (LPC)
3. Perceptual Linear Predictive (PLP) Coefficients
4. Discrete wavelet transform (DWT)
5. Principal component analysis(PCA)

A. Mel Frequency Cepstral Coefficients (MFCC)

To derive a characteristic vector containing all information about the linguistic message, MFCC mimics some elements of human speech manufacturing and speech understanding [13]. As the frequency bands are placed exponentially in MFCC, it can approximate human auditory device response more strongly than other feature extraction techniques. MFCC [15] feature extraction is built on large word analyzes and thus from each frame, MFCC feature vector is extracted. Then the spectrum of the speech signal is calculated for each frame, with the aid of using Discrete Cosine Transform (DCT).

Subsequently, Mel scaling is performed on the obtained spectrum by filtering out through the filter bank. The

MFCC computation is calculated using the equation below:

$$Mel(f) = 2595*\log10(1+f/700)$$

B. Linear Predictive Coefficients (LPC)

The critical idea at the back of the Linear Predictive Coding (LPC) evaluation is that a speech sample can be accurate as a continuous combination of formerly speech samples [5]. The LPC offers a strong, dependable and perfect technique to approximate the parameters that signify the vocal tract system. The autocorrelation analysis is done. The LPC[17] gives excellent effects for the speaker recognition as an alternative than speech recognition. LPC is an effective speech recognition method and it has acquired regard as a formant estimation technique.

C. Perceptual Linear Predictive (PLP) Coefficients

This model was developed by Hermansky. The target of PLP model [24]is to depict the psychonomics of human being precisely in the feature extraction method. In contrast to linear predictive estimation of speech, sensory activity LP alter the rapid time duration spectrum of the speech via many psychonomics based transformations. PLP in exact following main sensory activity aspects namely

- Power spectrum compared to windowed signal exploitation FFT.
- Bark scale is applied to it, as it refers to another variety of sensory activity

D. Discrete wavelet transform (DWT)

Wavelet Transform (WT) [22] is a current parameterization approach effectively used for some signal handling activities..It is frequently worked as a substitute of the Fourier Transform (FT) because of its capacity to indicate the signal in each time frequency domains. Parameterizations are built on Fourier Transform which is mostly used in speech recognition works. Due to the fact speech signal differs slowly and it could be consequently regarded as quasi stationary. However , this belief is kind of ease of the reality and it is accordingly appropriate to symbolize each speech sample more accurately. Therefore, modern-day attempts of researchers focus on the aspects of Wavelet Transform in countless fields of automated speech recognition.

E. Principal component analysis (PCA)

Principal aspect analysis (PCA) is mostly used as a method for data depletion besides any dropping of data. It is a method of changing one set of a variable into other sets, where the newly created one is difficult to elucidate. In various systems, PCA works to grant data on the actual measurement of a recordset. If the information set consists of S variables, they do not signal the required information. PCA converts a set of correlated variables into a new one that is known as principal components. Beside the interrelated data, the most important elements are extraneous and are arranged in words.PCA can be used for speech data containing any number of variables.

Table II: Comparison of Feature Extraction Techniques

S.No	COMPARISON OF FEATURE EXTRACTION TECHNIQUES		
	Feature Extraction Techniques	Advantages	Limitation
1	MFCC	Accuracy is high, low complexity	Background noise
2	LPC	Reliable ,accurate and robust technique, high speed ,low bit rate	Does not distinguish similar vowels, Degradation
3	PLP	More receptive to human aural faculty	Resultant feature vectors are dependent
4	DWT	Ability to flatten a signal without major degradation	Not flexible enough
5	PCA	Robust in nature	For high dimension data, PCA is expensive

2. CLASSIFICATION

Several speech recognition techniques [12] have been developed successfully and used in many applications. They are divided into three broad categories,

1. Acoustic Phonetic Approach
2. Pattern Recognition Approach
3. Machine learning Approach

A. Acoustic Phonetic Approach

Acoustic [27] offers with the study of special sounds and phonetics is the learn about the phonemes in the language. It is primarily based on the concept that describes there exist finite, unique vocal unit and those units are commonly distinguished with a set of properties that represent the speech signal. Although the acoustic properties of vocal units are incredibly a variant, each with audio sample and with nearby vocal units, sometimes it is also known as "co-articulation" of sounds.

Following are two steps taken in this approach,

•Fractionation and Categorizing section :

In first step, fractionation is finished alongside with categorizing section as it entails the speech signal into the distinct location where acoustic properties are represented as one vocal unit.

• Determination of valid words from Fractionation: The second steps tries to determine a legit phrase from the order of vocal labels generated in the first process.

B. Pattern Recognition Approach

Actuarial pattern matching is used efficiently in a large variety of industrial speech processing systems. In this approach, a sample is denoted as a set of features that are seen as an extensional feature vector. Familiar concepts are used to set up resolution border between samples. Here two modes are designed: training and classification.

a) HIDDEN MARKOV MODEL

It is a mathematical approach to recognize speech and a doubly embedded stochastic device with an underlying stochastic method that is now no longer immediately observable (it is hidden), however, can be determined completely by some different set of stochastic strategies that produce the sequence of observations. This modeling requires the use of anticipation fashions to proceed with inadequate information. In speech processing, unpredictability and inadequate information occur from many scenarios; for example, unclear sounds, copy of words and homophone words. Thus, stochastic fashions are a particularly appropriate strategy for speech recognition.

In a Markov method, the inspection is attached to the releasing state. In HMM, the remark is an anticipation characteristic of the state. Every state has a connected probability density of released symbols. Suppose when the process is in the actual state, output symbols are released according to the probability density

B. TEMPLATE MATCHING APPROACH

a) DYNAMIC TIME WARPING

DTW algorithm is primarily based on Dynamic programming. This algorithm [28] is used for analyzing equality among sequence which additionally differ in the domain of time and speed. Therefore the method is also used to find the perfect arrangement among a collection of series, if one of the series may additionally be covered non-linearly through extending it alongside its time domain. This can be further used to locate respective regions among the collection to determine the similarity between the two-time series. DTW provides a manner to align with the test and reference pattern to give the common distance related to the most effective wrapping direction.

C. MACHINE LEARNING APPROACH

The ability of machine learning is to code the computers in order to solve a given problem for given data. The method combines the study of pattern recognition with the machine's ability to analyze, research and make a decision accordingly. Several methods exist for this task such as Artificial Neural Networks, SVM, Decision Trees and the combination of methods.

a) NEURAL NETWORK RECOGNITION APPROACH

During the last two decades, some choice techniques to HMMs and GMMs have been proposed which are in general based on ANN. Generally, ANN [4] is represented as an important class of discriminating techniques, which are very well suited for classification problems.

b) SUPPORT VECTOR MACHINE (SVM)

SVMs are developed from Statistical Learning

Theory. The objective of SVM [13] is to generate a model which is built on the training data to foresee the target values of the test data using the Kernel Adatron algorithm. The SVMs are high-quality discriminative classifiers with various incredible characteristics, namely: their answer is that with most margin; they are successful to deal with samples of a very greater dimensionality, and their convergence to the minimum of the related price feature is guaranteed [6]. These characteristics have made SVMs very popular and successful.

V.CONCLUSION

This paper presents various feature extraction and classification techniques in the field of the speech recognition system. The most commonly used feature extraction method is MFCC and is considered to balance between enhancing accuracy and reducing computational complexity in speech systems. A neural network is widely used as a classification technique to improve accuracy for a medium set of words. Artificial Intelligence emerges as a recent trend in achieving successful output for a large database of words and are carried out by researchers.

ACKNOWLEDGEMENT

This work is performed at Dr.Mahalingam College of Engineering and Technology as a part of project work supported by Department of Computer Science and Engineering.

REFERENCES

[1] Feras E. Abualadas1 , Akram M. Zeki , Muzhir Shaban Al-Ani3, Az-Eddine Messikh4 (2019), "Speaker Identification based on Hybrid Feature Extraction Techniques", International Journal of Advanced Computer Science and Applications (IJACSA), Vol: 10, Issue: 3, 2019

[2] Turgut Ozseven (2019), "A novel feature selection method for speech emotion recognition", Applied Acoustics,Dec2019

[3] Dimitri Palaz, Mathew Magimai-Dossb, Ronan Collobert (2019), "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM- based Automatic Speech Recognition", IOSR Journal of Computer Engineering (IOSR-JCE), March 2019

[4] Rajeev Ranjan, Abhishek Thakur(2019), "Analysis of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol:8, Issue:7C2, May 2019

[5] Gulbakshee J Dharmale, Dipti D Patil (2019), "Evaluation of Phonetic System for Speech Recognition on Smartphone", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol:8 Issue:10, Aug 2019

[6] Chatterjee, Akshay, and Ghazaala Yasmin, "Human Emotion Recognition from Speech in Audio Physical Features",In Applications of Computing, Automation and Wireless Systems in Electrical Engineering, pp. 817-824. Springer, Singapore, 2019.

[7] Patil, Nilesh M., and Milind U. Nemade, "Content- Based Audio Classification and Retrieval Using Segmentation, Feature Extraction and Neural Network Approach", In Advances in Computer Communication and Computational Sciences, pp. 263-281. Springer, Singapore, 2019

[8] Bhoomika Dav, Prof. D. S. Pipalia (2014), "Speech recognition: A Review", International Journal of Advance Engineering and Research Development, Vol: 1,Issue :12, Dec-2014

[9] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur(2016), "Automatic Speech Recognition: A Review", International Journal of Computer Science Trends and Technology, Vol: 7, Issue: 4, Dec 2016

[10] Dr.V.Ajantha Devi , Ms.V.Suganya (2016), "An Analysis on Types of Speech Recognition and Algorithms", International Journal of Computer Science Trends and Technology, Vol: 4 ,Issue: 2, Apr 2016

[11] Bhushan C. Kamble1 (2016), "Speech Recognition Using Artificial Neural Network – A Review", International Journal of Computing, Communications & Instrumentation Engineering. (IJCCIE), Vol:3, Issue :1

[12] Dinesh Sheoran, Pardeep Sangwan, Manoj Khanna, (2017), "Spectral feature extraction techniques for speech recognition", International Journal of Multidisciplinary Research and Development, Vol: 4 Issue: 6, June 2017

[13] Ahmed Sajjad, Ayesha Shirazi, Nagma Tabassum, Mohd Saquib, Naushad Sheikh (2017), "Speaker Identification & Verification Using MFCC & SVM", International Research Journal of Engineering and Technology (IRJET) ,Vol: 04 ,Issue: 02 ,Feb -2017

[14] S.Bagavathi, S.I.Padma(2017), "Neural network based voiced and unvoiced classification using egg and MFCC feature", International Research Journal of Engineering and Technology (IRJET), Vol: 04, Issue: 04,Apr -2017

[15] Archek Praveen Kumar, Ratnadeep Roy, Sanyog Rawat and Prathibha Sudhakaran(2017), "Continuous Telugu Speech Recognition through Combined Feature Extraction by MFCC and DWPD Using HMM based DNN Techniques", International Journal of Pure and Applied Mathematics, Vol: 114, Issue: 11,2017

[16] Gurpreet Kaur, Mohit Srivastava, Amod Kumar (2017), "Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition", International Journal of Engineering and Technology Innovation, Vol: 7,Issue:2

[17] Ms. R.D. Bodke, Prof.Dr.M.p.Satone (2018), "A Review on Speech Feature Techniques and Classification Techniques", International Journal of Trend in Scientific research and Development, Vol: 2, Issue: 4, June 2018

[18] Atma Prakash Singh , Ravindra Nath , Santosh Kumar(2018), "A Survey: Speech Recognition Approaches and Techniques", International Conference on Electrical, Electronics and Computer Engineering

[19] Khin May Yee, Moh Moh Khaing, Thu Zar Aung(2018), "Classification of Language Speech Recognition System", International Journal of Trend in Scientific research and Development, Vol: 3,Issue: 5, August 2018

[20] R. Thiruvengatanadhan (2018), "Speech Recognition using SVM ", International Research Journal of Engineering and Technology (IRJET), Vol: 05 ,Issue: 09, Sep 2018

[21] Siddharth S More, Prashantkumar L. Borde, Sunil S Nimbhore(2018),"Isolated Pali Word (IPW) Feature Extraction using MFCC & KNN Based on ASR", IOSR Journal of Computer Engineering (IOSR-JCE), Vol: 20, Issue: 06,Dec -2018

[22] Mohammad Hasan Rahmani, Farshad Almasganj, Seyyed Ali Seyyedsalehi(2018),"Audio-visual feature fusion via deep neural networks for automatic speech recognition ", Digital Signal Processing, July 2018

[23] Tanmay Bhowmika, Amitava Chowdhury, Shyamal Kumar Das Mandala (2018), "Deep Neural Network based Place and Manner of Articulation Detection and Classification for Bengali Continuous Speech", International Conference on Smart Computing and Communications, Dec 2018

[24] H.M.Mohammed , M.S. Alkassab , H.R. Mohammed , H.Abdulaziz , Ahmed S. Jagmagj(2018), "Speech Recognition System with Different Methods of Feature Extraction", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 3, March 2018

[25] Bhuvaneshwari Jolad& Dr. Rajashri Khanai(2018), "Different feature extraction techniques for Automatic speech recognition: a review" International Journal of Engineering Sciences & Research Technology,Dec2018

[26] Ankur Mauryaa, Divya Kumara, R.K.Agarwal(2018), "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach", International Conference on Smart Computing and Communications (ICSCC),Dec 2018

[27] Ali Bou Bassif , Ismail shahin , Imtinan Attili, Mohammad Azzeh, and Khaled Shaala (2018), "Speech Recognition Using Deep Neural Netwoks: A Systematic Review", IEEE Access,Vol:7,Feb 2018

[28] C.B.Kare, Mrs.V.S.Navale2(2015), "Speech recognition by Dynamic Time Warping", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) ,Vol:8 Issue:16