ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 3 Issue 6; November-December-2016; Page No. 01-06

Speech Recognition Challenges by using Neural Network Approaches

[1]Dr. Kavita, [2]Dr. Akash Saxena, [3]Jitendra Joshi

[1]Research Supervisor, Jayoti Vidyapeeth Women's University, Jaipur, India.

[2]Research Co-Supervisor, *Compucom Institute* of Technology & Management, *Jaipur*, India.

[3]Research Scholar Jayoti Vidyapeeth Women's University, Jaipur, India.

ARTICLE INFO

Received: 03 June 2016
Accepted: 26 September 2016

Corresponding Author:

Jitendra Joshi

E-mail: Lect.jitendra29@gmail.com

Keywords- Speech Recognition; Feature Extraction; MFCC; LPC; Hidden Markov Model; Neural Network; Dynamic Time Warping.

ABSTRACT

Speech technology and frameworks in user computer dealings have seen a steady and notable progress over the past twenty years. Presently, speech technologies are accessible on the market for an infinite though appealing set of functions. Such technologies allow devices to answer accurately and dependably to people's voices, and give helpful and important services. Current studies focus on exploiting frameworks that will be significantly stronger against changes in surroundings, user and language. Thus current studies mostly concentrate on ASR frameworks having a consequent glossary that enable speaker autonomous process comprising fluid speech in dissimilar tongues. This article provides a summary of the speech identification framework and its current development. The basic aim of this article is to contrast and recapitulate some of the popular techniques employed in different levels of speech identification framework.

1. INTRODUCTION

Speech is the very first, universal and proficient means of communication technique for humans to deal with one another. Humans are at ease with speaking, so people will further prefer to deal with computers through speech, instead of utilizing basic platforms like keyboards and pointing gadgets. This may be enabled by exploiting an Automatic Speech Recognition or ASR framework which permits a computer to recognize the expressions that a user says into a microphone or phone and transform it into a lettered document. Consequently, it has the possibility of being a vital means of communication between people and computers [1]. Though, any function that implicates collaboration with a computer may possibly utilize ASR. The ASR framework can sustain numerous important functions such as transcription, order and control, integrated functions, phone index help, spoken database questioning, health functions, office transcription gadgets, and mechanical sound conversion into other languages, and so on. In the present Indian situation, such machine-adapted interfaces limit the computer utilization to small portion of the populace, who are computer knowledgeable as well as familiar with written English. Interaction between people is monopolized by spoken language. So, it is normal for humans to anticipate spoken communications with computers that may talk and identify speech in the mother tongue. It will allow even a simple person to garner the advantage of information technology. India has a semantically rich region having 18 legal languages, which are inscribed in 10 dissimilar writings [2]. Thus there is a particular requirement for the ASR framework to be exploited in their mother tongues. This article gives a summary of speech identification framework and the survey of methods accessible at different levels of speech identification.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS

Speech identification frameworks may be split in many dissimilar categories by depicting the kind of speech discourse, kind of speaker mode, kind of tunnel, and the kind of glossary that they possess the means to identify. Speech identification is turning out to be increasingly intricate and a difficult job owing to this changeability in the signal. Such difficulties are shortly described as follows.

A. Types of Speech Utterance

A discourse is the saying (speaking) of an expression or expressions that denote one signification to the computer. Utterances may be a solitary expression, a phrase, a sentence, or possibly several sentences. The kinds of speech discourse are:

1) Isolated Words

Isolated word detectors typically necessitate every discourse to have silence at the two ends of the model opening. It does not signify that it admits solitary expressions, though does necessitate one solitary discourse at a time. This is good for occasions where the person needs to provide just one expression answers or orders, but is really twisted for plural expression feeds. It is relatively basic and most facile to enforce as word limitations are clear and the expressions have the tendency to be unambiguously enunciated, which is the main benefit of this kind. The drawback of this kind is selecting various limitations impacts the outcomes.

2) Connected Words

Associated word frameworks (or more accurately 'associated discourses') are just like solitary expressions, though enable different discourses to be 'run-together' with the least gap between them.

3) Continuous Speech

Non-stop speech detectors enable people to talk more or less normally, while the computer deduces the material. Essentially, it is computer transcription. It counts a significant amount of "co-articulation", where neighboring expressions are processed at the same time with no gaps or any other obvious spacing between expressions. Non-stop speech identification frameworks are toughest to design as they have to use particular techniques to deduce discourse limits. As vocabulary increases, muddling between various word arrangements increases.

4) Spontaneous Speech

This kind of speech is normal and not practiced. An ASR framework equipped with instantaneous speech must be capable of handling a range of normal speech characteristics like words being spoken at the same time, and possibly minor stammering. Instantaneous (not practiced) speech can count wrong-beginnings, mispronunciations, and nonwords.

B. Types of Speaker Model

Every speaker has his own particular voices, owing to his exclusive physical body and character. Speech identification framework is generally categorized into two major classifications according to speaker models, that is speaker reliant and speaker autonomous.

1) Speaker dependent models

Speaker reliant frameworks are created for a particular user. They are typically more correct for the specific user, though significantly less correct for other users. Such frameworks are normally simpler to exploit, less expensive, and more correct, though not as supple as speaker adjustable or speaker autonomous frameworks.

2) Speaker independent models

Speaker autonomous frameworks are devised for a range of users. It detects the speech designs of a large grouping of people. This framework is toughest to exploit, most costly, and provides reduced correctness compared to speaker reliant frameworks. Though, they are more adaptable.

C. Types of Vocabulary

The volume of the glossary that a speech identification framework has impacts the intricacy, operating needs, and the correctness of the framework. Some functions just need a few expressions (for instance just numbers), others necessitate really large glossaries (for instance transcription devices). In ASR frameworks, the kinds of vocabularies may be categorized as below.

Small dictionary - tens of expressions
Medium dictionary - hundreds of expressions
Large dictionary - thousands of expressions
Very-large dictionary - tens of thousands of expressions
Out-of-Vocabulary – Superposing an expression from the dictionary into the unidentified expression.

Besides the features above, the situation changes, tunnel changes, speech model, age, sex, and rapidity of speech additionally render the ASR framework more intricate. Though, the proficient ASR frameworks should manage with the changeability in the signal.

3. GROWTH OF ASR SYSTEMS

Constructing a speech identification framework turns out to be really intricate owing to the conditions stated in the last section. Even if speech identification technology has progressed to the stage where it is utilized by millions of people for utilization in a range of functions, the study is currently concentrating on ASR frameworks that include three characteristics, namely: extensive glossaries, non-stop speech capacities, and speaker autonomy. Presently, there are different frameworks which include such mixtures. Yet, with so many technological limitations in exploiting the ASR framework, it has nonetheless achieved the greatest progress. The ASR framework's milestones are provided in Table 1 below.

Year	Progress of ASR System
1952	Digit Recognizer
1976	1000 word connected recognizer with constrained grammar
1980	1000 word LSM recognizer (separate words w/o grammar)
1988	Phonetic typewriter
1993	Read texts (WSJ news)
1998	Broadcast news, telephone conversations
1998	Speech retrieval from broadcast news
2002	Rich transcription of meetings, Very Large Vocabulary, Limited Tasks, Controlled Environment
2004	Finnish online dictation, almost unlimited vocabulary based on morphemes
2006	Machine translation of broadcast speech
2008	Very Large Vocabulary, Limited Tasks, Arbitrary Environment
2009	Quick adaptation of synthesized voice by speech recognition (in a project where TKK participates in)
2011	Unlimited Vocabulary, Unlimited Tasks, Many Languages, Multilingual Systems for Multimodal Speech Enabled Devices
Future Direction	Real time recognition with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent.

Table 1: Growth of ASR System

4. AUTOMATIC SPEECH RECOGNITION (ASR) SYSTEM

The function of ASR is to use a sound wave as a contribution and generate a list of expressions as result. Essentially, the issue of speech identification may be described as follows. When provided with sound surveillance X = X1,X2…Xn, the aim is to determine the equivalent arrangement of expressions W = W1,W2…Wm that contains the most subsequent possibility P(W|X) denoted by utilizing the Bayes theorem as illustrated in equation (1). Figure 1 below depicts a summary of the ASR framework.

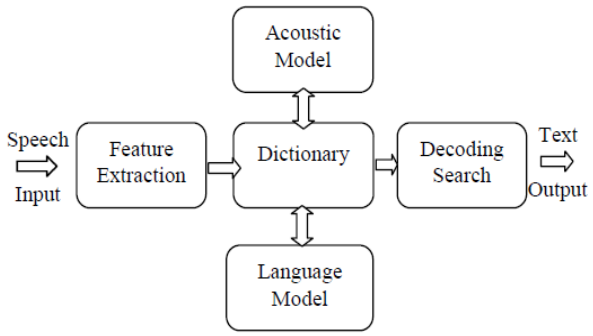

Figure 1: ASR System

For the purpose of identifying speech, the framework typically comprises of two phrases, termed pre-processing and post-processing. The former implicates characteristic retrieval and the latter stage involves constructing a speech identification machine. Speech identification machine generally comprises of data concerning constructing an acoustic system, vocabulary, and grammar. When every element is properly provided, the identification machine recognizes the most probable equivalent for the provided contribution, and it generated the identified expression.

A basic job of exploiting any ASR framework is to select the appropriate characteristic retrieval method and the identification technique. The appropriate characteristic retrieval and identification methods may generate fine correctness for the particular function. Thus, these two main elements are studied and contrasted according to their benefits and drawbacks to determine the most suitable method for speech identification framework. The different kinds of characteristic retrieval and speech identification methods are described in the next section.

V. SPEECH FEATURE EXTRACTION TECHNIQUES

Characteristic retrieval is the most significant aspect of speech identification as it has an essential function to differentiate one discourse from another. Since each discourse has dissimilar particular features integrated in the speech. Such feature may be retrieved from an extensive variety of characteristic retrieval methods suggested and effectively used for speech identification function. However, the retrieved characteristic must satisfy some conditions while handling the speech signal like:

- Simple to gauge retrieved speech characteristics.
- It must not be vulnerable to imitation.
- It must display slight variation from one communication situation to another.
- It must be steady with use.
- It must show up often and normally when speaking.

The most extensively employed characteristic retrieval methods are illustrated as follows.

A. Linear Predictive Coding (LPC)

Among the most influential signal evaluation methods is the technique of linear prediction. The LPC [3][4] of a discourse has evolved to be the principal method to deduce the essential boundaries of speech. It gives a correct approximation of the speech as well as being a proficient computational version of speech. The primary concept behind LPC is that a discourse example may be estimated to be a linear mixing of previous speech examples. By reducing the addition of squared differences (over a fixed range) between the real speech examples and expected values to a minimum, an exclusive range of boundaries or predictor coefficients may be deduced. Such coefficients form the foundation for LPC of speaking [10]. The evaluation gives the possibility for calculating the linear prediction design of speech with time. So, the predictor coefficients are converted to a stronger series of boundaries termed cepstral coefficients. Figure 2 illustrates the steps implicated in LPC characteristic retrieval.

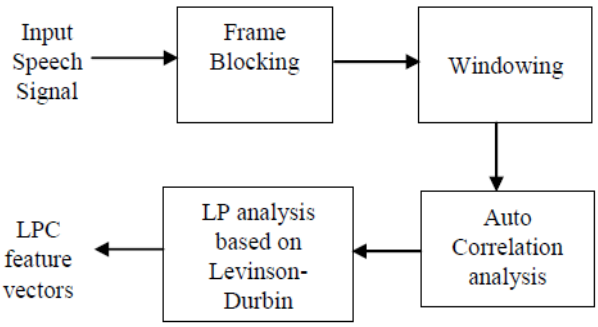

Figure 2: Steps involved in LPC Feature extraction

B. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [3][4] is clearest instance of a characteristic fixed that is widely utilized in speech identification. Since the frequency bands are placed logarithmically in MFCC [6], it estimates people's response more minutely compared to all other systems. The method of calculating MFCC is founded on the immediate evaluation, and hence from every frame an MFCC vector is calculated. For retrieving the coefficients, the speech example is used as the contribution and characterizing window is employed to reduce to a minimum the interruptions of a signal. Subsequently, DFT is utilized to create the Mel sorting bank. As per the Mel frequency distortion, the width of the triangular sieves changes and thus the registered summed energy in a crucial band about the middle frequency is counted. After distorting, the coefficient numbers are acquired. Figure 3 illustrates the steps implicated in MFCC characteristic retrieval.

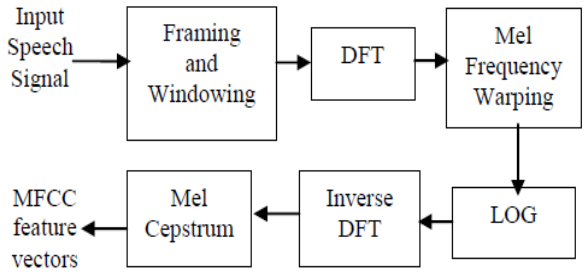

Figure 3: Steps involved in MFCC feature extraction

5. SPEECH RECOGNITION APPROACHES

In precedent years, active programming methods were exploited to resolve the pattern-identification issue [12]. Later studies were founded on Artificial Neural Network or ANN methods, where the analogous computing discovered in biological neural systems is imitated. Of late, stochastic simulation methods have been integrated to resolve the speech identification issue, like the Hidden Markov Modeling or HMM technique. Currently, a good portion of the new investigations on speech identification implicate identifying non-stop speech from an extensive dictionary employing HMMs, ANNs, or a mixed form [12]. Such methods are shortly described as follows.

A. Template-Based Approaches

Systems based on template for speech identification have given a family of methods that have significantly developed the domain during the past twenty years. The concept behind this system is basic. It is an operation of associating unfamiliar speech is contrasted against a series of pre-registered words (templates) so as to obtain the most likely equivalent (Rabiner *et al.*, 1979). The benefit of this is that it uses hundred percent correct word models; though it has as demerit that the pre-registered templates are set, therefore changes in speech may just be simulated by utilizing multiple templates per word, which gradually turns out to be unrealistic. Template readying and associating become excessively costly or unrealistic as dictionary volume grows more than a few hundred words. This approach was somewhat unproductive with regards to both necessary storage and operating power required to execute the association. Template association was significantly user reliant as well, and non-stop speech identification was not viable.

B. Knowledge-Based Approaches

The application of knowledge/condition founded method to speech identification has been recommended by numerous scientists and used on speech identification (De Mori and Lam, 1986; Alikawa, 1986; Bulot and Nocera, 1989), speech comprehension frameworks (De Mori and Kuhn, 1992). The "proficient" awareness on changes in speech is manually edited into the source code of a system. It utilizes series of characteristics from the speech, then the preparation system produces a series of manufacturing laws robotically from the examples. Such conditions are deduces from the boundaries that give the maximum data concerning a categorization. The identification is executed at the frame stage, employing an inference device (Hom, 1991) to implement the decision tree and categorize the execution of the conditions. This has as benefit the clear simulation of changes in speech; however, this proficient awareness is tough to acquire and utilize effectively, therefore this method was deemed to be unworkable, and robotic learning processes were rather studied.

C. Neural Network-Based Approaches

One more method in vocal simulation is the utilization of neural networks. These are able to resolve significantly more complex identification jobs, though do not rate as outstanding as Hidden Markov Model or HMM when it involves large dictionaries. Instead of being utilized in common speech identification uses, they may deal with inferior quality, noisy

information, and user autonomy [7] [11]. These networks may realize more consequent correctness then HMM founded systems, so long as there is preparation information and the dictionary is constrained. A more common method utilizing neural networks is phoneme identification. This is a dynamic area of study, though normally the outcomes are superior to HMMS [7] [9]. There are the blended NN-HMM systems as well that utilize the neural network portion for phoneme identification and the HMM portion of language simulation.

D. Dynamic Time Warping (DTW)-Based Approaches

Active Time Warping is a calculation for gauging likeness between two chains which can change with time or speed [8]. A popular use has been ASR, to manage with various speech paces. Generally, it is a technique that enables a computer to obtain the best equivalence between two known chains (such as time series) with some constraints, that is the chains are "warped" non-straightly to correspond to each other. This chain configuration approach is usually utilized in the HMM situation. This method is rather proficient for singular word identification and may be changed to identify related word as well [8].

E. Statistical-Based Approaches

In this technique, changes in speech are simulated mathematically (such as HMM), utilizing robotic studying processes. This method signifies the actual state of the art. Contemporary all-purpose speech identification frameworks are founded on mathematical sound and language examples. Useful sound and language examples for ASR in unlimited area necessitate significant quantity of vocal and language information for boundary approximation. Treatment of great quantities of preparation information is a major component in the progress of a successful ASR technology these days. The major drawback of mathematical examples is that they have to first make simulation suppositions, which are likely to be not precise, affecting the system's execution.

Hidden Markov Model (HMM)-Based Speech Recognition

The cause for HMM's popularity is that they may be prepared robotically and are easy and computationally practical to utilize [2] [5]. HMMs to denote full expressions may be simply built (employing the pronunciation glossary) from mobile HMMs and word arrangement possibilities included and full network researched for optimum way matching to the best word arrangement. HMMs are basic networks that may produce speech (chains of cepstral vectors) utilizing a quantity of states for every example and simulating the immediate spectra related with every state with, normally, combinations of multivariate Gaussian distributions (the state result distributions). The boundaries of the example are the state changeover possibilities and the averages, variances, and blended weights that define the state result distributions. Every word, or every phoneme, will have a dissimilar result distribution; a HMM for an arrangement of words or phonemes is generated by concatenating the specifically prepared HMM [12] for the different expressions and phonemes.

Modern HMM-founded large dictionary speech identification systems are usually practiced on hundreds of hours of vocal information. The word arrangement and a pronunciation glossary and the HMM [6] [12] preparation operation may robotically deduce word and phone parameter data during practice. This implies that it is comparatively direct to utilize large preparation compilation. It is the main merit of HMM which will tremendously diminish the time and intricacy of identification procedure for preparing consequent dictionary.

6. CONCLUSION

Speech identification has been under progress for over 50 years, and has been considered as a substitute access technique for people having handicaps for nearly as long. This article debates the basics of speech identification and researched its current development. The different methods accessible for exploiting an ASR framework are plainly described with its advantages and disadvantages. This article also contrasts the execution of the ASR method founded on the accepted characteristic retrieval scheme and the speech identification method for the specific language. Lately, the necessity for speech identification studies founded on large dictionary user autonomous non-stop speech has greatly risen. As per the survey, the compelling benefit of HMM method together with MFCC characteristics is more appropriate for these necessities and provides good identification outcome. These methods will allow us to generate more and more powerful frameworks that may prospectively be set up on a global basis.

REFERENCES

1. Bassam A. Q. Al-Qatab , Raja N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)", 978-1-4244-6716-711 0/$26.00 ©2010 IEEE.
2. M. Chandrasekar, M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal «Technical Acoustics» 2008, 20.
3. Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods

Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

4. DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03$17.00 © 2003 IEEE.
5. Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda , "Automatic Speech Recognition for Bangia Digits", Proceedings of 2009 12th International Conference on Computer and Information Technology (ICCIT2009) 21-23 December, 2009, Dhaka, Bangladesh, 978-1-4244-6284-1/09/$26.00 ©2009 IEEE.
6. A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.
7. Meysam Mohamad pour, Fardad Farokhi, "An Advanced Method for Speech Recognition", World Academy of Science, Engineering and Technology 49 2009.
8. Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
9. Raji Sukumar.A, Firoz Shah.A and Babu Anto.P, "Isolated question words recognition from speech queries by Using artificial neural networks", 2010 Second International conference on Computing, Communication and Networking Technologies, 978-1-4244-6589-7/10/$26.00 ©2010 IEEE.
10. N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.
11. Vimal Krishnan V. R, Athulya Jayakumar and Babu Anto.P, "Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network", 4th IEEE International Symposium on Electronic Design, Test & Applications, 0-7695-3110-5/08 $25.00 © 2008 IEEE
12. Zhao Lishuang , Han Zhiyan, "Speech Recognition System Based on Integrating feature and HMM", 2010 International Conference on Measuring Technology and Mechatronics Automation, 978-0-7695-3962-1/10 $26.00 © 2010 IEEE.