Journal of Advances and Scholarly Researches in Allied Education
Vol. 15, Issue No. 9, October-2018, ISSN 2230-7540

Study on Automated Speech Recognition is an Emerging Technique

Rajendra Kumar Mahto[1]* Dr. Shankar Rammoorthy[2]

[1] Research Scholar, Swami Vivekananda University, Sagar (MP)

[2] Professor, Computer Science Department, Swami Vivekananda University, Sagar (MP)

Abstract – Automated speech recognition is an emerging field in research. There are numerous research going on in this field where speech is identified and converted into text. This paper gives a basic idea on automated speech recognition and the techniques used behind this method.

----------X----------

I. INTRODUCTION

Speech is the primary means of communication between humans. The natural ease with which we communicate through conversations masks the intricacy of language. The decent variety in language arises from many factors:

- Geographical - there are more than thousands of languages comprising various dialects.
- Cultural - the dimension of education has a strong influence on speaking style
- Physical - each individual's voice box have marginally unique shapes subsequently differentiates among others in their speaking style and clarity.
- Psychological - each person has diverse speaking styles depending on their emotional state, intention, attitude and so on.

Automated speech recognition is a technique in which speech signals are converted into succession of words. This is an enhancing research area as researchers are interested in emulating human behavior. In earlier occasions automated speech recognition frameworks were responding to particular sounds alone. This has developed to framework that can perceive natural languages. The intricacy of speech recognition has got the attention of researchers for a considerable length of time, and various aspects of language and speech have been explained. For reasons, ranging from logical interest on the mechanisms for mechanical realization of human speech capabilities to, the craving to automate basic tasks which necessitate human-machine interactions. Speech recognition research work began in 50's. In 1959, at University College in England, Fry and Denes attempted to construct a phoneme recognizer to perceive four vowels and nine consonants [2].The Harpy framework was the first to take advantage of a finite state arrange (FSN) to lessen computation and productively determine the nearest matching string. The earliest attempts to devise ASR frameworks were made in 1950s and 1960s, when various researchers endeavored to abuse fundamental ideas of acoustic phonetics. In any case, it really made substantial advancement and as an important issue in conducting research in the late 60's the early 1970s.Further speech recognition in the 1980s the HMM demonstrate and the artificial neural network(ANN) are effectively used in speech recognition. Speech recognition innovation converts human voice command to text frame. This has an extensive variety of application like telephone organize, voice calls etc.

BASIC METHODS FOLLOWED IN AUTOMATED SPEECH RECOGNITION

The three basic approaches of speech recognition are

- Accoustic Front end
- Accoustic Model
- Search

Language Model

1. Acoustic phonetic approach
2. Pattern recognition approach
3. Artificial intelligence approach

ACCOUSTIC APPROACH:

In acoustic phonetic approach, speech sounds were found and these sounds are labeled to deliver text shape. The phonetic units in talked language are broadly classified by set of acoustic properties that vary as for time in speech signal. In this method, features are extracted from speech based on various classification like ratio of high and low frequencies, voiced and unvoiced classification, nasality. Acoustic phonetic approach pursued the following arrangement.

1. Spectral analysis
2. Feature detection
3. Segmentation and labeling
4. Recognizing valid word

Spectral analysis plays out the spectro temporal analysis of the signal. Depicts the power range of speech intervals. Feature detection involves the method of detecting features.

PATTERN RECOGNITION APPROACH

This method has gained its popularity in the ongoing years in speech recognition. Pattern recognition approach includes two basic advances namely

1. Pattern training
2. Pattern comparison

In this approach an immediate comparison is made between the talked words and the patterns learned in the training stage.

ARTIFICIAL INTELLIGENCE APPROACH

The artificial intelligence approach compared to other techniques concentrates in the similar fashion how a person applies his intelligence in visualizing and analyzing to make a decision on the acoustic features.

II. RELATED WORKS

The recently proposed algorithm Automatic speech recognition frameworks are trained under human supervision to give transcripts of speech utterances. Their main goal by proposing Active Learning was to minimize the human supervision for preparing acoustic and dialect . They portray how to register the confidence score for each utterance by an on-line algorithm using the lattice yield of a speech recognizer. The utterance scores are sifted through the in formativeness function and an optimal subset of training samples is selected.Active learning algorithm is an optimization algorithm that chooses the training examples and advance the test set word accuracy. Their proposed algorithm is a solution to LVCSRs two drawbacks. It makes inefficient utilization of data which is costly to transcribe. It confines the machines behavior to adapt dynamically to non-stationary input channels. Their approach to train adaptive LVCSRs based on the concept of active learning(AL) brought about a higher accuracy (71.0%), when used 19 000 utterances as AL gives a faster learning rate to new words and new - grams. Li Deng and Xiao Li introduces an arrangement of prominent ML paradigms that are motivated in the context of ASR innovation and applications. Their New insight from current ML methodology demonstrates great promise to advance the state-of-the-art in ASR innovation. They intended to encourage further cross-pollination between the ML and ASR people group than has happened in the past.

ML notion of organized classification as a fundamental issue in ASR—regarding both the emblematic succession as the ASR classifier's yield and the continuous-valued vector feature grouping as the ASR classifier's input. By presenting each of the ML paradigms, they have featured the most relevant ML concepts to ASR, and have emphasized the kind of ML approaches that are powerful in dealing with the special challenges of ASR including profound/dynamic structure in human speech and strong variability in the observations. They have also paid special attention to discussing and analyzing the major ML paradigms and results that have been confirmed by ASR tests. The main examples discussed in their research paper includes HMM-related and dynamics-situated generative learning, discriminative learning for HMM-like generative models, multifaceted nature control (regularization) of ASR frameworks by principled parameter tying, adaptive and Bayesian learning for environment-powerful and speaker-vigorous ASR, and half breed supervised/unsupervised learning or crossover generative/discriminative learning as exemplified in the later ―deep learning‖ plot involving DBN and DNN. Nevertheless, they have also discussed an arrangement of ASR models and methods that hadn't progressed toward becoming mainstream yet have a strong theoretical foundation. Despite the fact that there are several methods for automatic classification of utterances into emotional states have been proposed. Be that as it may, the detailed mistake rates are rather high, far behind the word blunder rates in speech recognition. Their research has given way for performance optimization by the utilization of a self-adaptive hereditary algorithm. This Paper consist of self-adaptive hereditary algorithms (GA's) to increase the probability of right classification in emotional speech recognition when the Bayes classifier with feature subset selection is used. It consists of two stages which are utilized to search for the most exceedingly awful performing features as for the probability of right classification achieved

by the Bayes classifier in the principal stage. That is, a hereditary algorithm based implementation of backward feature selection (SBS) is proposed. These features are progressively rejected from sequential floating feature selection using the probability of right classification achieved by the Bayes classifier as criterion. In the second stage, self-adaptive hereditary algorithms are utilized to search for the most exceedingly terrible performing utterances as for the same criterion.By the sequential application of the two stages it is demonstrated that it enhances speech emotion recognition.

In a comparative investigation of past work in speech recognition and surveys by comparing present day speech recognition frameworks and humans in request to determine how far later dramatic advances in innovation have made advancement towards the goal of human-like performance is performed. This paper measures how far existing researches have advanced towards this goal. Results from random investigations which have compared human and machine speech recognition on similar tasks are being summarized to determine how much speech recognizers must enhance to match human performance. Speech corpora used in these comparisons don't speak to day-to-day listening conditions, however span a band ranging from calm read isolated words - to noisy read sentences - to spontaneous telephone speech. Results, demonstrate that the advanced speech recognizers are as yet performing much more terrible than humans, both with wideband speech read in calm, and with band-constrained or noisy spontaneous speech. The outcomes comparing humans to machines are given four important goals.

They are to motivate research in directions that will decrease the human– machine performance gap, to advance further human– machine comparisons, to advance further experimental work with human listeners to understand how humans adapt to talker and environmental variability, and to encourage a multi-disciplinary dialog between machine recognition and speech perception researchers. This research comprises of comparisons in six current speech corpora with vocabularies ranging from 10 to in excess of 65,000 words and content ranging from read isolated words to spontaneous conversations. The mistake rates of machines are seem to be regularly in excess of a request of magnitude greater than those of humans for peaceful, wideband, read speech. In addition, machine performance degrades further beneath than that of humans in noise, with channel variability and for spontaneous speech. Humans can perceive calm, clearly talked nonsense syllables and nonsense sentences with minimal abnormal state grammatical information. These comparisons recommend that the human– machine performance gap can be significantly lessened by basic research on improving low-level acoustic-phonetic modeling, aiming on improving strength with noise and channel variability, and also on more accurately modeling spontaneous speech techniques. In In this paper the researchers with the aim to distinguish the sexual orientation of a speaker based on the voice of the speaker using by applying various speech processing techniques and algorithms, two models were made, one for generating Formant values of the voice sample and the other for generating pitch value of the voice sample using Lab VIEW. These two models were used for extracting sexual orientation biased features, i.e. Formant 1 and Pitch Value of a speaker. A preprocessing model was readied for filtering out the noise components in the voice sample and to raise the high recurrence formants in the voice sample[6]. In request to calculate the mean of formants and pitch of all the samples of a speaker, a model containing circle and counters were applied which generated a mean of Formant 1 and Pitch value of the speaker. By utilizing nearest neighbor method, calculating Euclidean distance the Mean estimation of Males and Females of the generated mean values of Formant 1 and Pitch, the speaker .For finding the sexual orientation of a speaker they have used acoustic measures from both the voice source and the vocal tract, the fundamental recurrence (F0) or pitch and the primary formant recurrence (F1) separately. It is outstanding that F0 values for male speakers are bring down because of longer and thicker vocal folds. F0 for adult males is typically around 120 Hz, while F0 for adult females is around 200 Hz.The algorithm was executed in real time using NI Lab VIEW.

From the outcomes obtained its proficiency is satisfying it was concluded that the algorithm actualized in Lab View is working effectively. Be that as it may, since the algorithm does not extract the vowels from the speech, the value obtained for Formant 1 weren't totally right as they were obtained by processing all the samples of the speech. It was also seen from trials that by increasing the unvoiced part in the speech, similar to the sound of ‗s', the value of pitch increases consequently hampering the sex detection in case of Male samples. Likewise by increasing the voiced, similar to the sound of ‗a', decreases the value of pitch yet the framework takes care of such plunge in value and results were not affected by the same. Also, extraordinary speech by the same speaker talked in the near to identical conditions generated the same pitch value establishing that the framework can be used for identification of speaker after further work.

III. WORKING OF AUTOMATED SPEECH RECOGNITION

Speech Recognition is the major type of communication among human beings. Speech

recognition is the way toward converting the speech signals created by human being to machine recognizable frame by means of the algorithm created by the client. There can be distinctive sorts of speeches.

ISOLATED WORD In the case of isolated word, the utterances are very on the two sides of the sample window. This type of speech recognition accepts words or single utterances at once

CONNECTED WORD In this type of speech recognition, words are separated by pauses. Like isolated word speech recognition, the basic speech recognition unit is the word.

CONTINUOUS SPEECH In continuous speech recognition, words are connected together instead of being separated by pauses. Continuous speech recognizers allows client to speak almost naturally, while the algorithm determine the continuity. Automatic speech recognizer with continuous speech capabilities are probably the most hard to create because they use special method to determine utterance boundaries. Subsequently in this method, boundary information about words, surrounding phonemes and rate of speech impact the performance.

SPONTANEOUS SPEECH

At a basic dimension, it tends to be thought of as speech that is natural sounding and not rehearsed. This sort of framework ought to be able to handle a variety of natural speech feature, for example, words being run together. Automatic speech recognition is gaining importance these days as the vast majority of the cell phones are worked with this application that make the client easy to make a call or type a message Automatic speech recognition framework contains the following modules

1. Speech signal acquisition
2. Feature extraction
3. Acoustic modeling
4. Language modeling

SPEECH SIGNAL ACQUISITION

In this module, sound recording is done. The motivation behind this module is to capture the most ideal signal.

FEATURE EXTRACTION

Feature extraction is the most important advance in automated speech recognition. The performance of the recognition of the speech very relies upon the feature extraction phase. The speech feature extraction in a categorization issue is about reducing the dimensionality of the input vector while maintaining the discriminating intensity of the signal. As we probably am aware from fundamental formation of speaker identification and verification framework, that the quantity of training and test vector required for the classification issue develops with the dimension of the given input so we require feature extraction of speech signal. Researchers have used many feature extraction techniques like PCA, LDA, ICA, linear prescient scoring and so on.

ACOUSTIC MODELING

This is the main component of an Automatic Speech Recognition framework. This model takes care of the performance of the framework. This particular module takes care of the talked phonetics. This module in particular uses the audio recordings of the speech and utilize the text contents to gather them into a statistical representation of the sounds that creates the word.

LEXICAL MODELING

Lexicon is a module in which pronunciation of each module is structured according to the given language. Various combinations of speeches are defined to give valid words for the recognition.

LANGUAGE MODELS

This module is trained on many words. This module is produced so the connection between the words in a sentence is planned with the assistance of pronunciation dictionary.

IV. CONCLUSION

Automated speech recognition is an emerging strategy that helps in recognizing the human speech by the machine. There are numerous research going on in building a model for recognizing speech and converting into text. The paper summarizes the various kinds and methods pursued.

REFERENCES

1. V. M. Velichko and N. G. Zagoruyko (1970). Automatic recognition of 200 words,‖ Int. J. Man-Machine Studies, 2, pp. 223.
2. D. B. Fry (1959). Theoretical aspects of mechanical speech recognition‖; and P. Denes, —The design and operation of the mechanical speech recognizer at University College London,‖ J. British Inst. Radio Engr., 19, 4, pp. 211-229.
3. Automatic speech recognition: the development of the SPHINX system‖,Kai-

Fu Lee; Boston; London: Kluwer Academic, c1989.

4. Review of Neural Networks for Speech Recognition, R. P. Lippmann in Neural Computation‖, v1(1), pp. 1-38, 1989.

5. Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T. (2005). Fiction database for emotion detection in abnormal situations.‖,In: Proc. Int. Conf. Spoken Language Process.(ICSLP '04). Korea, pp. 2277–2280.

6. Batliner, A., Hacker, C., Steidl, S., N¨oth, E., D'Archy, S., Russell, M., Wong, M. (2004). You stupid tin box- children interacting with the AIBO robot: A cross linguistic emotional speech—, In: Proc. Language Resources and Evaluation (LREC '04). Lisbon, 2004.

7. Santhosh K., Bharthi W. & Yannawar Pravin (2010). A review of Speech Recognition Technique‖, International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

8. Riccardi, Giuseppe, and Dilek Hakkani-Tur (2005). "Active learning: Theory and applications to automatic speech recognition." Speech and Audio Processing, IEEE Transactions on 13.4: pp. 504-511.

9. Deng, Li, and Xiao Li (2013). "Machine learning paradigms for speech recognition: An overview." IEEE Transactions on Audio, Speech and Language Processing 21.5: pp. 1060-1089.

10. Sedaaghi, Mohammad Hossein, Dimitrios Ververidis, and Constantine Kotropoulos (2007). "Improving speech emotion recognition using adaptive genetic algorithms." Proc. European Signal Processing Conference (EUSIPCO), Polland. 2007.

11. Lippmann, Richard P. (1997). "Speech recognition by machines and humans." Speech communication 22.1: pp. 1-15.

12. Rakesh, Kumar, Subhangi Dutta, and Kumara Shama. "Gender Recognition using speech processing techniques in LABVIEW." International Journal of Advances in Engineering & Technology 1.2 (2011): 51-63.

13. Automatic speech recognition and speech variability: A review‖, M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont *, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens Multitel, ParcInitialis, Avenue Copernic, B-7000 Mons, Belgium,6 February 2007.

14. Literature Review on Automatic Speech Recognition‖, Wiqas Ghai, Navdeep Singh (2012). International Journal of Computer Application, *Volume 41– No.8, March 2012.*

Corresponding Author

Rajendra Kumar Mahto*

Research Scholar, Swami Vivekananda University, Sagar, (MP)