A study on Automated Speech Recognition Technique

Vijayalakshmi A, Midhun Jimmy, Moksha Nair

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.**

Index Terms—**Automated speech recognition, speech, speech to text.**

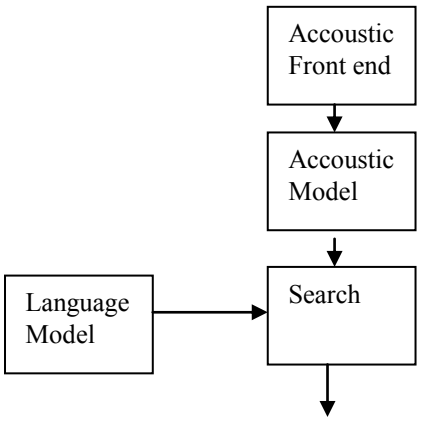
I. INTRODUCTION

Speech is the primary means of communication between humans. The natural ease with which we communicate through conversations masks the complexity of language [1][2]. The diversity in language arises from many factors:

- Geographical - there are over thousands of languages comprising various dialects.
- Cultural - the level of education has a strong influence on speaking style
- Physical - each individual's voice box have slightly different shapes hence differentiates among others in their speaking style and clarity.
- Psychological - each person has different speaking styles depending on their emotional state, intention, attitude etc.

Automated speech recognition is a procedure in which speech signals are converted into sequence of words. This is a enhancing research area as researchers are interested in emulating human behaviour. In earlier times automated speech recognition systems were responding to particular sounds alone. This has evolved to system that can recognize natural languages. The complexity of speech recognition has got the attention of researchers for centuries, and a number of aspects of language and speech have been explained. For reasons, ranging from scientific curiosity on the mechanisms for mechanical realization of human speech capabilities to, the desire to automate simple tasks which necessitate human-machine interactions. Speech recognition research work began in 50's. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [2].The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching string. The earliest attempts to devise ASR systems were made in 1950s and 1960s, when various researchers tried to exploit fundamental ideas of acoustic phonetics. But it really made substantial progress and as an important issue in conducting research in the late 60's the early 1970s.Further speech recognition in the 1980s the HMM model and the artificial neural network(ANN) are successfully used in speech recognition. Speech recognition technology converts human voice command to text form. This has a wide range of application like telephone network, voice calls etc.

Manuscript received Feb, 2013.
Vijayalakshmi A, *Department of Computer Science, Christ University, Bangalore, India.*
Midhun Jimmy, *Department of Computer Science, Christ University, Bangalore, India,*
Moksha Nair, *Department of Computer Science, Christ University, Bangalore, India,*



BASIC METHODS FOLLOWED IN AUTOMATED SPEECH RECOGNITION

The three basic approaches of speech recognition are

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 labelled to produce text form. The phonetic units in spoken language are broadly classified by set of acoustic properties that vary with respect to 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 followed the following sequence.

1. Spectral analysis
2. Feature detection
3. Segmentation and labelling
4. Recognising valid word

ISSN: 2278 – 1323

Spectral analysis performs the spectro temporal analysis of the signal. Describes the power spectrum of speech intervals. Feature detection involves the procedure of detecting features.

PATTERN RECOGNITION APPROACH

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

1. Pattern training
2. Pattern comparison

In this approach a direct comparison is made between the spoken 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

In [8] Giuseppe Riccardi and and Dilek Hakkani-Tür Proposed an active learning algorithm for ASR. By the newly proposed algorithm Automatic speech recognition systems are trained under human supervision to provide 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 compute the confidence score for every utterance by an on-line algorithm using the lattice output of a speech recognizer. The utterance scores are filtered through the in formativeness function and an optimal subset of training samples is selected.Active learning algorithm is an optimization algorithm that selects the training examples and optimize the test set word accuracy. Their proposed algorithm is a solution to LVCSRs two drawbacks.
It makes inefficient use of data which is expensive to transcribe.
It restricts 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) resulted in a higher accuracy (71.0%), when used 19 000 utterances as AL provides a faster learning rate for new words and new -grams.

Li Deng and Xiao Li [9] introduces a set of prominent ML paradigms that are motivated in the context of ASR technology and applications. Their New insight from modern ML methodology shows great promise to advance the state-of-the-art in ASR technology. They intended to foster further cross-pollination between the ML and ASR communities than has occurred in the past.

ML notion of structured classification as a fundamental problem in ASR—with respect to both the symbolic sequence as the ASR classifier's output and the continuous-valued vector feature sequence as the ASR classifier's input. By presenting each of the ML paradigms, they have highlighted the most relevant ML concepts to ASR, and have emphasized the kind of ML approaches that are effective in dealing with the special difficulties of ASR including deep/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 experiments. The main examples discussed in their research paper includes HMM-related and dynamics-oriented generative learning, discriminative learning for HMM-like generative models, complexity control (regularization) of ASR systems by principled parameter tying, adaptive and Bayesian learning for environment-robust and speaker-robust ASR, and hybrid supervised/unsupervised learning or hybrid generative/discriminative learning as exemplified in the more recent "deep learning" scheme involving DBN and DNN.
Nevertheless, they have also discussed a set of ASR models and methods that hadn't become mainstream but have a solid theoretical foundation.

Even though there are several methods for automatic classification of utterances into emotional states have been proposed. However, the reported error rates are rather high, far behind the word error rates in speech recognition. Their research has given way for performance optimization by the use of a self-adaptive genetic algorithm [10]. This Paper consist of self-adaptive genetic algorithms (GA)'s to increase the probability of correct classification in emotional speech recognition when the Bayes classifier with feature subset selection is used. It consists of two stages which are employed to search for the worst performing features with respect to the probability of correct classification achieved by the Bayes classifier in the first stage. That is, a genetic algorithm based implementation of backward feature selection (SBS) is proposed. These features are successively excluded from sequential floating feature selection using the probability of correct classification achieved by the Bayes classifier as criterion. In the second stage, self-adaptive genetic algorithms are employed to search for the worst performing utterances with respect to the same criterion.By the sequential application of both stages it is demonstrated that it improves speech emotion recognition.

In [11] a comparative study of past work in speech recognition and reviews by comparing modern speech recognition systems and humans in order to determine how far recent dramatic advances in technology have made progress towards the goal of human-like performance is performed. This paper measures how far existing researches have progressed towards this goal. Results from random studies which have compared human and machine speech recognition on similar tasks are being summarized to determine the degree to which speech recognizers must improve to match human performance. Speech corpora used in these comparisons don't represent day-to-day listening conditions, but span a band ranging from quiet read isolated words - to noisy read sentences - to spontaneous telephone speech. Results, demonstrate that the modern speech recognizers are still performing much worse than humans, both with wideband speech read in quiet, and with band-limited or noisy spontaneous speech. The results comparing humans to machines are presented with four important goals. These are

ISSN: 2278 – 1323

to motivate research in directions that will decrease the human–machine performance gap, to promote further human–machine comparisons, to promote 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 modern speech corpora with vocabularies ranging from 10 to more than 65,000 words and content ranging from read isolated words to spontaneous conversations. The error rates of machines are observe to be often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Moreover, machine performance degrades further below than that of humans in noise, with channel variability and for spontaneous speech. Humans can recognize quiet, clearly spoken nonsense syllables and nonsense sentences with little high-level grammatical information. These comparisons propose that the human–machine performance gap can be significantly reduced by basic research on improving low-level acoustic-phonetic modeling, aiming on improving robustness with noise and channel variability, and also on more accurately modeling spontaneous speech techniques.

In [12] In this paper the researchers with the aim to identify the gender 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 LabVIEW. These two models were used for extracting gender 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 frequency formants in the voice sample[6]. In order to calculate the mean of formants and pitch of all the samples of a speaker, a model containing loop 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 gender of a speaker they have used acoustic measures from both the voice source and the vocal tract, the fundamental frequency (F0) or pitch and the first formant frequency (F1) respectively. It is well-known that F0 values for male speakers are lower due to 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 implemented in real time using NI Lab VIEW. From the results obtained its efficiency is satisfying it was concluded that the algorithm implemented in Lab View is working successfully. But, since the algorithm does not extract the vowels from the speech, the value obtained for Formant 1 weren't completely correct as they were obtained by processing all the samples of the speech. It was also observed from experiments that by increasing the unvoiced part in the speech, like the sound of 's', the value of pitch increases thus hampering the gender detection in case of Male samples. Likewise by increasing the voiced, like the sound of 'a', decreases the value of pitch but the system takes care of such dip in value and results were not affected by the same. Also, different speech by the same speaker spoken in the near to identical conditions generated the same pitch value establishing that the system can be used for identification of speaker after further work.

III. WORKING OF AUTOMATED SPEECH RECOGNITION

Speech Recognition is the major form of communication among human beings. Speech recognition is the process of converting the speech signals produced by human being to machine recognizable form by means of the algorithm developed by the user. There can be different types of speeches.

ISOLATED WORD

In the case of isolated word, the utterances are quite on both sides of the sample window. This form of speech recognition accepts words or single utterances at a time

CONNECTED WORD

In this form 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 user to speak almost naturally, while the algorithm determine the continuity. Automatic speech recognizer with continuous speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries. Hence in this method, boundary information about words, surrounding phonemes and rate of speech effect the performance.

SPONTANEOUS SPEECH

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. This type of system should be bale to handle a variety of natural speech feature such as words being run together [7].

Automatic speech recognition is gaining importance these days as most of the mobile phones are built with this application that make the user easy to make a call or type a message

Automatic speech recognition system 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 purpose of this module is to capture the best possible signal.

FEATURE EXTRACTION

Feature extraction is the most important step in automated speech recognition. The performance of the recognition of the speech highly depends on the feature extraction phase. The speech feature extraction in a categorization problem is about

ISSN: 2278 – 1323

reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need feature extraction of speech signal. Researchers have used many feature extraction techniques like PCA, LDA, ICA, linear predictive scoring and so on [3].

ACOUSTIC MODELING

This is the main component of an Automatic Speech Recognition system. This model takes care of the performance of the system. This particular module takes care of the spoken phonetics. This module in specific uses the audio recordings of the speech and use the text scripts to compile 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 designed 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 developed so that the connection between the words in a sentence is designed with the help of pronunciation dictionary.

IV. CONCLUSION

Automated speech recognition is an emerging technique 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 types and methods followed.

REFERENCES

[1] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," Int. J. Man-Machine Studies, 2, pp. 223, 1970.
[2] D. B. Fry, "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, 1959.
[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. "Fiction database for emotion detection in abnormal situations.",In: Proc. Int. Conf. Spoken Language Process.(ICSLP '04). Korea, pp. 2277–2280, 2005.
[6] Batliner, A., Hacker, C., Steidl, S., N¨ oth, E., D' Archy,S., Russell, M., Wong, M. ,"You stupid tin box- children interacting with the AIBO robot: A crosslinguistic emotional speech ", In: Proc. Language Resources and Evaluation (LREC '04). Lisbon, 2004.
[7] Santhosh K, Bharthi W, Yannawar Pravin , "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. "Active learning: Theory and applications to automatic speech recognition." Speech and Audio Processing, IEEE Transactions on 13.4 (2005): 504-511.
[9] Deng, Li, and Xiao Li. "Machine learning paradigms for speech recognition: An overview." IEEE Transactions on Audio, Speech and Language Processing21.5 (2013): 1060-1089.
[10] Sedaaghi, Mohammad Hossein, Dimitrios Ververidis, and Constantine Kotropoulos. "Improving speech emotion recognition using adaptive genetic algorithms." Proc. European Signal Processing Conference (EUSIPCO), Polland. 2007.
[11] Lippmann, Richard P. "Speech recognition by machines and humans." Speech communication 22.1 (1997): 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. WellekensMultitel, Parc Initialis, Avenue Copernic, B-7000 Mons, Belgium,6 February 2007.
[14] "Literature Review on Automatic Speech Recognition", Wiqas Ghai, Navdeep Singh,International Journal of Computer Application, *Volume 41– No.8, March 2012.*

ISSN: 2278 – 1323