State of the Art Review of Speech Recognition using Genetic Algorithm

Trishna Barman
Department of Information Technology
Gauhati University
Guwahati, India
trishna.guist@gmail.com

Nabamita Deb
Department of Information Technology
Gauhati University
Guwahati, India
deb.nabamita@gmail.com

***Abstract*—In recent years, speech recognition has become of the most emerging topic for many researchers. Speech is one of the most important tools for communication between human and his environment. People are so easeful with speech that they would also like to intercommunicate with computers via speech, rather than having to apply to ancient adapters. This paper covers mainly two topics firstly, what are the different speech recognition techniques and secondly, how genetic algorithm helps in speech recognition.**

***Keywords*—Speech Recognition, Analysis, Feature extraction, Modeling, Testing, Modeling Techniques, Genetic Algorithm**

I. Introduction

Speech Recognition is a process of converting an acoustic signal to a text form, in which the voice recognition program works to recognize the proper word resembling to each word of voice. Speech is one of the most significant tools for communication between humans and his environment, hence building of Automatic System Recognition is desire for him all the time [1]. In Speech Recognition, the training phase plays a very essential part. Design of a good training model for speech pattern enhances the characteristics of the overall performance in recognizing the speech utterance.

Genetic algorithm is inspired by the Darwinian Theory of "Survival of the Fittest". Genetic Algorithm is an optimization method based on the idea of natural selection. Genetic algorithms are frequently used to solve both constrained and unconstrained optimization problems in investigation and in artificial intelligence. Genetic algorithm relies on different biological operators viz. mutation, crossover and selection.

II. Speech Recognition Techniques

The Speech Recognition system may be classified into four sub systems which are as follows:

1) Analysis
2) Feature extraction
3) Modeling
4) Testing

A. Speech Analysis Techniques

Speech data contains various information which are specific to the speaker due to the vocal tract, excitation, behaviour, etc. Those information can be used to recognize the speaker. Speech analysis is one of the fundamental steps in speech processing. It deals with segmenting the speech signal with suitable frame size for further analysis and extraction [2]. Speech analysis technique is carried out using following techniques.

1) Segmentation Analysis: In this technique, speech is analyzed using the frame size and shift in a particular range of 10-30ms to extract speaker information.

2) Sub-segmental Analysis: In this technique, speech is analyzed using the frame size and shifting is done in a particular range of 3-5ms. This technique is used to mainly analyze and extract the features of the excitation state [3]

B. Feature Extraction Techniques

Feature extraction is the process of obtaining various features such as vocal tract, excitation, behaviour, etc. from the speech signal. This step should derive some descriptive features from the segmented speech signal to enable further classification of speech.

The different feature extraction technique are as follows:

- Spectral feature like band energies, formats, spectrum and Cepstral coefficient mainly speaker specific information due to vocal tract.
- Excitation source feature like pitch and variation in pitch.
- Long term feature like duration, information energy due to behavior feature.

C. Modeling Techniques

The aim of modeling technique is to generate speaker models using speaker specific feature vector. The modeling techniques are used for speaker recognition and speaker identification.

Following are some of the modeling which can be used in speech recognition process.

978-1-5386-0814-2/17/$31.00 ©

1) The Acoustic-Phonetic Approach: This approach is based upon theory of acoustic phonetics and postulates [4]. The earliest approaches to theory recognition were based on finding speech sounds and provide appropriate labels to these sounds. This is the basis of the acoustic phonetic search. Using IPA methods we can find similarities for probabilities of content dependant acoustic model for new language [5].

2) Pattern Recognition Approach: The pattern matching approach involves mainly two steps namely, pattern training and pattern comparison. The notable feature of this technique is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations from a set of training samples. The representation of this approach can be in the form of a speech template or a statistical model (Hidden Markov Model or HMM) and can be applied to a sound, a word, a phrase.

3) Dynamic Time Warping: Dynamic time warping or DTW is the oldest method and the simplest way to recognize an isolated word from a sentence. In this approach, the words are compared against a number of stored word templates and determine which word has the best match.

4) The Artificial Intelligence Approach: The artificial intelligence [6] is a hybrid of the acoustic phonetic approach and pattern recognition approach. This approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence. This form of knowledge based system improvement has enriched extensively to the modeling of all successful strategies.

5) Stochastic Approach: The stochastic modeling approach [7] encompasses the use of probabilistic models to handle with unpredictable information. The most popular stochastic approach today is Hidden Markov Model (or HMM).

D. Matching Techniques

Once the modeling step is performed, the next step in speech processing is to match a detected word to a known word. This can be achieved either by whole-word matching or sub-word matching [8].

1) Whole-word Matching: In this technique, the engine compares the incoming digital-audio signal against a pre-recorded template of the word. This technique takes much processing than sub-word matching. Whole-word templates requires large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed [9].

2) Sub-word Matching: In this approach, the engine looks for sub-words usually phonemes and then performs further pattern recognition on those. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand [10] [11].

III. Performance of Systems

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy is measured in terms of performance accuracy which is generally rated with word error rate (WER). On the other hand, speed is measured with the real time factor. Some of the other accuracy measures include Single Word Error Rate (SWER) and Command Success Rate (CSR) [12].

A. Word Error Rate (WER)

Word error rate is a common metric of the performance of a speech recognition system. The common challenges of measuring performance lies in the fact that the recognized word sequence may have a different length from the reference word sequence. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [11] [13]. Word error rate can then be computed as:

$$WER = S + D + I/N \tag{1}$$

Where,

- S is the number of substitutions
- D is the number of the deletions
- I is the number of the insertions
- N is the number of words in the reference

IV. Speech recognition using Genetic Algorithm

Genetic algorithm is a method for solving optimization problems that is based on the process of biological evolution. Genetic algorithm solves a variety of optimization problems which are not well suited for standard optimization algorithms.

A. How Genetic Algorithm Works

A genetic algorithm has a pool or a population of the possible solutions to the given problem. These solutions then undergo recombination and mutation, producing new children, and the process is repeated over various generations. Each individual is assigned a fitness value and the fitter individuals are given higher chance to mate and produce more fitter individuals. In this way a genetic algorithm keeps yielding fitter individuals or solutions over generations, till it reaches a stopping criterion.

B. Related Works

To identify speaker by analyzing the sound signal using artificial neural network and fuzzy system, Melin et. al. [14] (in 2006) described a method of voice recognition based on a monolithic neural network. They have implemented tests with 20 different words recorded from three different speakers and achieved a very good recognition results by the monolithic neural network based recognition system with considered it can be achieving about 96% recognition rate when increasing the database of words over the 100 words.

Another voice recognition system has been proposed by Wroniszewska et. al. [15] (in 2010) that hybrid the genetic algorithm with a classifier of K-nearest neighbor. They have achieved a satisfactory construction of the model and determined the influence of simulation parameters on the classification score. The results show the overall system accuracy has been got 94.2% of correctly classified patterns in 26 seconds.

To solve nonlinear, discrete and constrained problems for dynamic time warping, Benkhellat et. al. [16] (in 2012) have used a genetic algorithm and their works have shown that the important contribution of the genetic algorithms in temporal alignment through increasingly small factor of distortion.

In order to achieve better result for speech recognition, Gupta et. al. [17] (in 2014) proposed genetic algorithm for optimization. They have found that the level of accuracy using HMM was strongly influenced by the optimization of extraction process and modeling methods. On the other hand, they have shown better results can be achieved with the help of genetic algorithm. They have found that recognition accuracy for feature extraction with Fourier-Bessel cepstral coefficients (FBCC) in comparison with Mel-frequency cepstral coefficients (MFCC) is better.

V. Conclusion

In this review, we have presented a study of various steps as well as methods involved in a speech recognition system. In our consideration, MFCC is used widely for feature extraction of speech and HMM is best among all modeling. Moreover, we have also studied some of the works related to speech recognition using genetic algorithm and genetic algorithm gives better results for speech recognition. HMM was the widely used method in speech recognition but the level of accuracy using HMM was strongly influenced by the optimization of extraction process and modeling methods while on the other hand, better results can be achieved with the help of genetic algorithm.

References

[1] F. F. Meysam and F. Fardad, "An advanced method for speech recognition," *World Academy of Science, Engineering and Technology*, 2009.

[2] G.-D. Wu and Y. Lei, "A register array based low power fft processor for speech recognition." *Journal of Information Science & Engineering*, vol. 24, no. 3, 2008.

[3] N. Morales, J. H. Hansen, and D. T. Toledano, "Mfcc compensation for improved recognition of filtered and bandlimited speech," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–521.

[4] (2017) Ibm research - home. [Online]. Available: http://www.research.ibm.com

[5] C. Myers and L. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 284–297, 1981.

[6] R. K. Moore, "Twenty things we still dont know about speech," in *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology*, 1994.

[7] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 845–848.

[8] N. J. Ibrahim, Z. Razak, M. Yakub, Z. M. Yusoff, M. Y. I. Idris, and E. M. Tamil, "Quranic verse recitation feature extraction using mel-frequency cepstral coefficients (mfcc)," in *Proc. of the 4th IEEE Int. Colloquium on Signal Processing and its Application (CSPA), Kuala Lumpur, Malaysia*, 2008.

[9] S. Katagiri, "Speech pattern recognition using neural networks," *Pattern Recognition in Speech and Language Processing*, pp. 115–147, 2003.

[10] D. Raj Reddy, "An approach to computer speech recognition by direct analysis of the speech wave," Tech. Rept. CS 49. Stanford: Comp. Sci. Dept., Stanford Univ, Tech. Rep., 1966.

[11] L. R. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[12] K. Nagata, "Spoken digit recognizer for japanese language." *NEC research & development*, no. 6, 1963.

[13] D. T. Tran, "Fuzzy approaches to speech and speaker recognition," Ph.D. dissertation, university of Canberra, 2000.

[14] P. Melin and O. Castillo, "Voice recognition with neural networks, fuzzy logic and genetic algorithms," in *Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing*. Springer, 2005, pp. 223–240.

[15] M. WRONISZEWSKA and J. DZIEDZIC, "Voice command recognition using hybrid genetic algorithm," *TASK QUARTERLY*, vol. 14, no. 4, pp. 377–396, 2010.

[16] Z. Benkhellat and A. Belmehdi, "Genetic algorithms in speech recognition systems," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2012, pp. 853–858.

[17] H. Gupta and D. S. Wadhwa, "Speech feature extraction and recognition using genetic algorithm," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 1, pp. 363–369, 2014.