Survey paper on Different Speech Recognition Algorithm: Challenges and Techniques

Ayushi Y. Vadwala
B.Tech Student
Madhuben & Bhanubhai Patel Women's Institute Of Engineering, New Vallabh Vidyanagar
Gujarat, India

Krina A. Suthar
B.Tech Student
Madhuben & Bhanubhai Patel Women's Institute Of Engineering, New Vallabh Vidyanagar
Gujarat, India

Yesha A. Karmakar
B.Tech Student
Madhuben & Bhanubhai Patel Women's Institute Of Engineering, New Vallabh Vidyanagar
Gujarat, India

Nirali Pandya
Assistant Professor
Madhuben & Bhanubhai Patel Women's Institute Of Engineering, New Vallabh Vidyanagar
Gujarat, India

ABSTRACT

The Speech is most major & prime mode of Communication among human beings. The communication among human and computer is referred as human computer interface. Speech can be used to commune with computer. The speech recognition research is becoming more and more determined. Today, researchers are trying to making an effort to extend the capabilities of what computers can do with the spoken words. This paper consists of the classification of algorithms through which an uttered word can be converted to computer intelligible form. The challenges in speech recognition will be enumerated and analyzed for the most popular recognition techniques used today. The analysis ends with a brief description of some of the applications of speech recognition.

General Terms

Algorithms of Speech Recognition.

Keywords

Speech Recognition, Hidden Markov Model, Artificial Intelligence, Pattern Recognition, Neural Network.

1. INTRODUCTION

Speech is the most crucial, widespread and proficient form of communication method for people to commune with each other. Human are comfortable with speech hence persons would also like to interact with computers via speech, rather than via primitive interfaces such as keyboards and pointing devices. Speech Recognition is the inter-disciplinary sub-field of computational linguistics that build up techniques and technologies that facilitates the recognition and translation of spoken words into textual format by computers. It is also known as "speech to text" (STT). It includes knowledge and research in the linguistics, computer science, and electrical engineering fields. The objective of speech recognition is for a computer to be capable to "perceive speech", "recognize" and "take action upon" spoken words [2][3][4][5].

This linguistic techniques and approaches can be used to develop different type of applications based on speech recognition. The applications with speech recognition feature will also make life easier for those who are physically disabled and every common user who is fascinated by voice recognition. Best example of speech recognition based application is Intelligent voice assistant which takes speech as an input and performs actions consequently on different platforms [1]. Google provides API of speech recognition for android application developers to make their programming easier which uses various techniques which are described in this paper. [22].

2. CHALLENGES OF SPEECH RECOGNITION SYSTEM

An utterance is the speaking of a word. Utterances can be a single word, a few words, a sentence, or even multiple sentences. The types of speech utterance are:

2.1 Utterance approach

It implies that how the words are spoken either in isolated or in connected manner.

2.1.1 Isolated words

An isolated word speech recognition system necessitates that the speaker provides a brief intermission between words. It doesn't denote that it recognize single words, but does entail a single utterance at a time. This is fine for conditions where the user is obligatory to give only one word responses or commands, however it is extremely aberrant for multiple word inputs.

2.1.2 Connected words

Connected word systems (or more correctly 'connected utterances') are analogous to isolated words, nevertheless it allows separate utterances to be 'run-together' with a nominal pause among them.

2.2 Utterance style

All humans converse differently, it is a means of expressing their personality. Not only do they use personal terminologies, they have an unique way to articulate and emphasize. The speaking style also shows a discrepancy in different situations. Humans also communicate their emotions via speech. A person converse differently when he or she is happy, sad, irritated, anxious, upset, self-protective etc. It is majorly divided in two parts that is whether the speech is continuous or spontaneous.

2.2.1 Continuous Speech

Continuous speech recognizers permit users to speak roughly naturally, while the computer concludes the content. It includes an immense pact of "co articulation", where adjoining words are spoken together without temporary halts or any other noticeable division between words. Continuous speech recognition systems are extremely complicated to create because they must use extraordinary means to determine utterance boundaries. As the list of vocabulary increases, confusability between different word sequences increases.

2.2.2 Spontaneous Speech

This type of speech is natural and not rehearsed. Spontaneous or extemporaneously produced, speech contains disfluencies, and is much more complicated to recognize than speech read from script. It is immensely difficult because it tends to be peppered with disfluencies like “uh” and “um”, bogus starts, imperfect sentences, coughing, laughter and moreover vocabulary is essentially limitless, so the system ought to be able to deal cleverly with unknown words.

2.3 Types of Speaker Model

All speakers have their individual voices, due to their inimitable physical body and personality. Speech recognition system is largely classified into two main classes based on speaker models namely speaker dependent and speaker independent.

2.3.1 Speaker dependent models

Speaker dependent systems are designed for a specific speaker. In speaker dependent a user must provide samples of his or her speech before using them, a speaker dependent system is supposed for the use of a single speaker.

2.3.2 Speaker independent models

Speaker independent systems are planned for range of speakers having different articulations. It identifies the speech patterns of a huge collection of people. This system is most tricky to develop, most steep and provides a reduced amount of accuracy than speaker dependent systems. Though, they are more supple.

2.4 Vocabulary

The sizes of vocabulary of a speech recognition system have an effect on the intricacy, processing requirements and the exactness of the system. A few applications solitary require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation machines). In Speech Recognition Systems the categories of vocabularies can be classified as below:

- Small vocabulary - tens of words
- Medium vocabulary - hundreds of words
- Large vocabulary - thousands of words
- Very-large vocabulary - tens of thousands of words
- Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word

2.5 Channel variability

One phase of variability is the perspectives were the sound wave is uttered. In this, the difficulty with noise that alter over time, and various kinds of microphones and everything else that effects the content of the sound wave from the speaker to the distinct depiction in a computer. This occurrence is called channel variability.

Speech is articulated in an surroundings of sounds, a clock ticking, a computer humming, a radio playing somewhere down the corridor, another human speaker in the background etc. This is frequently called noise, i.e., unwanted information in the speech signal. A further kind of noise is the echo effect, which is the speech signal bounced on some surrounding object, and that arrives in the microphone a few milliseconds later. If the place in which the speech signal has been produced is strongly echoing, then this may give raise to a phenomenon called reverberation, which may last even as long as seconds.

Apart from the above characteristics, the environment variability, sex, age, speed of speech also makes the Speech Recognition system more complex. But the efficient Speech Recognition systems must cope with the variability in the signal. [2][14][21].

3. TECHNIQUES

The speaker recognition system may be viewed as working in three stages. Initially the speech input is given to the system then the pre-processing tasks are done to analyse the input. Furthermore feature extraction and classification of the speech input is done in order to have the resultant string output in Fig 1.

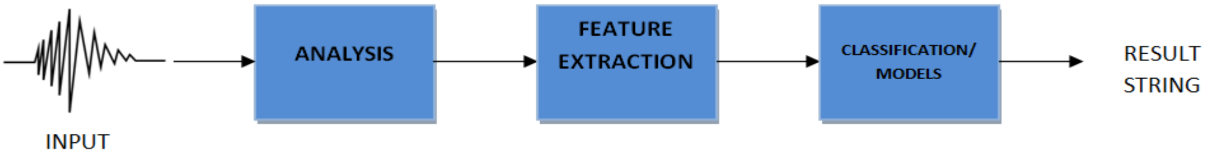

Fig 1: Process of Speech Recognition

3.1 Analysis

Interference during the speech recognition mainly occurs due to noise which further degrades the performance of the system. Before giving the speech signal to the feature extraction block the noise in the speech signal must be removed in order to have a better result. Pre-processing does this task. It eliminates the noise based on zero-crossing rate and energy. The parting of voiced and unvoiced speech based on both energy and zero-crossing rate gives the most excellent result [7]. The start off point and ending points are dogged based on energy and zero-crossing rates. The output speech contains the information and the unwanted noise is eliminated.

3.2 Feature Extraction

Speech feature extraction is very important stage which is for transformation of the speech signals into stream of feature vectors coefficients which contains only that information which is required for the identification of a given utterance. As every speech has dissimilar unique characteristics enclosed in spoken words these characteristics can be pulled out from a wide range of feature extraction techniques and can be employed for speech recognition task. But pulled out feature should meet up certain measures while dealing with the speech signals, such as: extracted speech features should be calculated easily, extracted features should be steady with time, and features should be robust to noise and surroundings. [8]. Different feature extraction techniques are :

1. Mel-Frequency Cepstrum Coefficients (MFCC)
2. Linear Predictive Coding (LPC)
3. Linear Prediction Cepstral Coefficients (LPCC)
4. Perceptual Linear Prediction (PLP)
5. Linear Discriminant Analysis (LDA)
6. Discrete Wavelet Transform (DWT)

7. Relative Spectral (RASTA-PLP)
8. Principal Component analysis (PCA).

3.3 Classification

Speech recognition process deals with speech variability and account for learning the relationship between specific utterance and the corresponding word or word. There are three approaches to speech recognition.[7][9][10]

3.3.1 Acoustic Phonetic Approach
3.3.2 Pattern Recognition Approach
3.3.3 Artificial Intelligence Approach

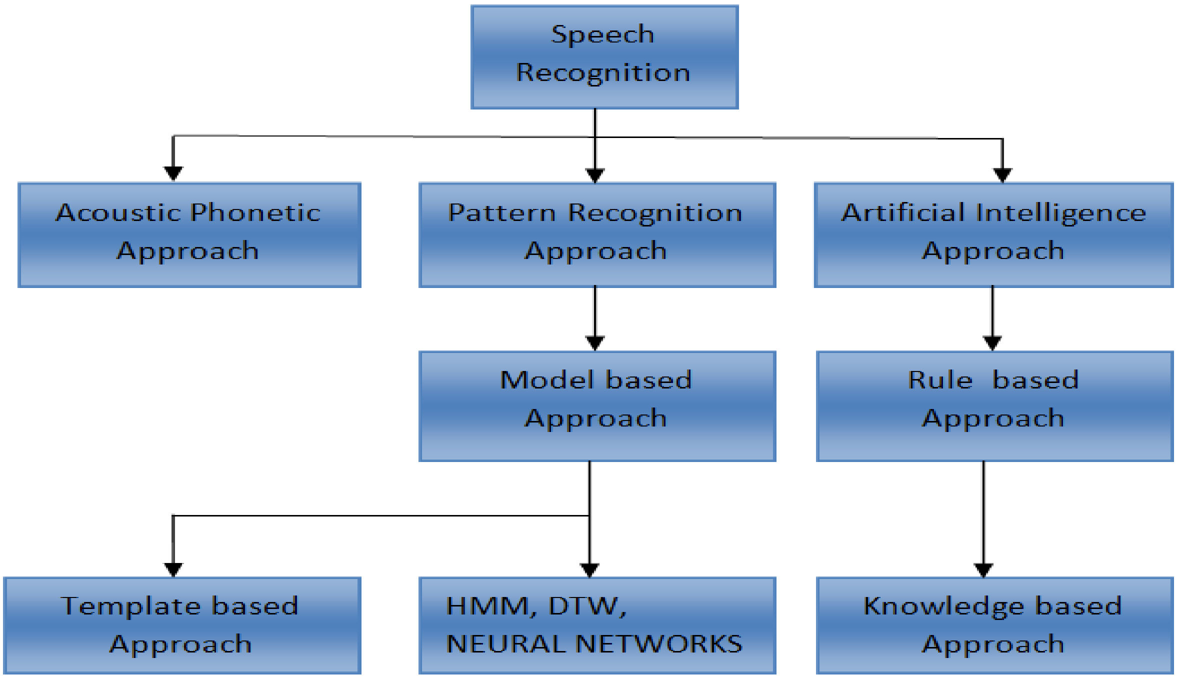
Speech Recognition
Acoustic Phonetic Approach
Pattern Recognition Approach
Artificial Intelligence Approach
Model based Approach
Rule based Approach
Template based Approach
HMM, DTW, NEURAL NETWORKS
Knowledge based Approach

Fig 2: Taxonomy of Speech Recognition

3.3.1 Acoustic Phonetic Approach

In Acoustic Phonetic approach the speech recognition were based on discovering speech sounds and endowing them with appropriate labels. This is the foundation of the acoustic phonetic approach which hypothesize that there exist fixed, distinguishing phonetic units called phonemes and these units are largely regarded as by a set of acoustics properties present in speech.

Even though, the acoustic belongings of phonetic units are highly inconsistent, both with narrator and with neighbouring sounds, it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily well-read by a machine.

The primary stage in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that alter the spectral measurements to a set of features that portray the broad acoustic properties of the different phonetic units. The subsequent step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling.[9][10][11]

3.3.2. Pattern Recognition Approach

Pattern classification mainly embody of the training or development of a system (given a feature vector) which will divide a large number of individual examples into groups called classes. As the source of the speech is often due to a large amount of many causes, the available speech signal results of the combination of the audio channel, noise, additive noise, etc.[12]. Pattern Classification (or recognition) is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. After completing training of the system at the time of testing patterns are classified to recognize the speech.

3.3.2.1 Template based approach:

Template based approach has a collection of exemplary speech patterns. These patterns are stored as reference patterns representing the dictionary of words. Speech is documented by matching an unknown spoken expression with each of these reference templates and selecting the category of the finest matching pattern. Normally Templates for whole words are built. Errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be evaded.

Template based approach to speech recognition has provided a family of technique that has advanced the field considerably during the last two decades. This approach is simple. It is the procedure of matching unknown speech with a set of pre-recorded words in order to find the best match. This approach has the benefit of using perfectly accurate word models; but this has the downside that the pre-recorded templates are predetermined. So difference in speech signals can only be modelled by using many templates per word, which certainly becomes impractical. Template training and matching turn into prohibitively costly or impractical as vocabulary size amplify beyond a few hundred words. This method is rather unproductive in terms of both requisite storage and processing power needed to complete the matching. Template matching is also mind-numbingly speaker reliant. Continuous speech recognition is not possible using this approach.[13]

3.3.2.2 Dynamic Time Warping:

The time alliance of various utterances is the central part difficulty for distance measurement in speech recognition. A little swing or change results into false detection. Dynamic time warping is an algorithm for measuring resemblance among two sequences which may fluctuate in time or speed. DTW is a technique that uncovers the most favourable match between two given sequences with certain constraints. The sequences are perverted nonlinearly in the time dimension. DTW was accepted as the most appropriate method for speech recognition for the reason of its potential to deal with dissimilar speaking swiftness. [7]

$D (i, j) = d (i, j) + \min (D (i, j) + T_{10}, D (i, j) + T_{11}, D (i, j) + T_{01})$

A variety of segments of the utterances are broadened and compacted so as to uncover arrangement that ends with the preeminent promising match among test and location vectors feature by facet origin. The stages incorporated in Dynamic time warping are in the following:

1) Record, consider and stock up dictionary of reference words.
2) Record analysed word to be familiar with and parameterize.
3) Evaluate distance amid test word and apiece reference word.
4) Pick reference word next to test word. [14]

The foremost trouble in dynamic time warping is to set up reference pattern. Beforehand it was geared up by selecting an illustration of each word that is to be recognized. It is referred as reference template although a distinct template is not adequate since it is not feasible to speak again and again the alike word in analogous approach as previously by the speaker, thus to steer clear of this crossword reference template is owned for the designing of crossword reference template is with multiple examples. Then the average length of the extracted template is estimated. Next template with length closed to average length is elected to be superlative template. Preliminary reference is afterwards template. DTW's time support the other templates. Final reference template is achieved by performing the average time aligned template across frame.

3.3.2.3 Hidden Markov Model:

Hidden Markov Models (HMMs) present an effortless and valuable structure for moulding time-varying spectral vector sequences. In consequence, nearly the entire present day large vocabulary continuous speech recognition (LVCSR) systems are based lying on HMMs.

While the fundamental theories underlying HMM-based LVCSR are quite easy, the approximations and abridge suppositions involved in a direct carrying out of these principles would result in a system which has pitiable accurateness and improper sensitivity to changes in operating environment. Thus, the practical application of HMMs in modern systems involves significant erudition. [15]

In HMM, state series are concealed and the observations are probabilistic tasks of the state. A Hidden Markov Model is an assortment of states associated by alterations. The selection of alteration and output symbol are both haphazard, managed by probability distributions. The order of output symbols created over time is noticeable, however the order of states stayed over time is out of sight from vision. HMM model can be depicted by these three set of parameters a, b and π and the model of N states and M observations referred to by:

$\lambda = (A, B, \pi);$

where $A = \{ a_{ij} \}$, $B = \{ b_j (w_k)\}$ and $1< I, j\leq N$ and $1\leq K \leq M$. [13]

The potency of HMM is its framework which contains mathematical things and the configuration of implementation. It is computationally viable. The three cases for HMM are Evaluation, Decoding and Training.HMM method is prompt in its preliminary training and when a new vocal is entered in the training course of action to build a new HMM model.[14]

The following figure shows the general architecture of HMM model in which its shape signifies arbitrary variable which is capable to accept a number of values.

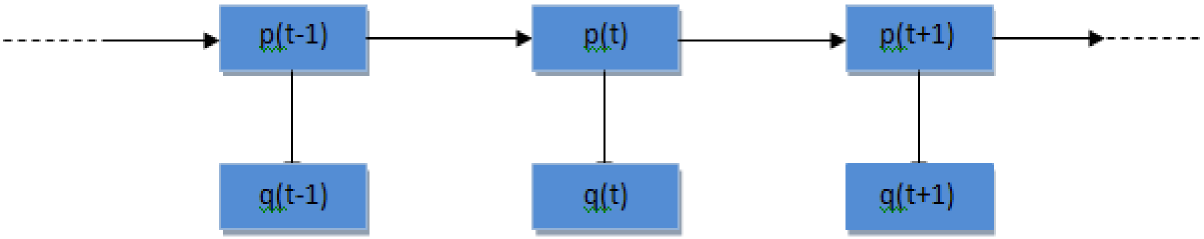

Fig 3: HMM Architecture

3.3.2.3 Neural Network:

Another approach in acoustic modelling is the use of neural networks. An artificial neural network (ANN) is a tensile mathematical arrangement which is proficient of identifying complex nonlinear relationships among input and output data sets. ANN models have been found valuable and well-organized, mainly in troubles for which the characteristics of the processes are complicated to describe using physical equations. Information that surges through the network influences the structure of the ANN because a neural network modifies - or learns, in a sense - based on that input and output[18][20].

They are able to solve a lot more problematical recognition tasks, but do not scale as excellent as Hidden Markov Model (HMM) when it comes to huge dictionaries. More willingly than being used in general-purpose speech recognition applications they can deal with low quality, noisy data and speaker independence [19] . Such systems can accomplish greater correctness than HMM based systems, as long as there is training data and the vocabulary is finite. A more common approach using neural networks is phoneme recognition. This is an active field of research, but generally the results are improved than HMMs [19].

3.3.3 Artificial Intelligence Approach

Computers behave like human being so it is called as Artificial intelligence. AI makes machine very intelligent and useful[16]. The approach of artificial intelligence is the most enlarge and effective techniques, which supports error-free and accurate speech recognition and being used for decoding. It is because; artificial intelligence includes certain algorithmic techniques, which fosters logical conversion and transformation of speech into understandable patterns, and vice versa.

The artificial intelligence approach is the mixture of the pattern recognition approach and acoustic phonetic approach so it is called hybrid approach of pattern recognition and acoustic phonetic approach[7]. It is due to the fact that it incorporates the concepts and ideas of pattern recognition methods and acoustic phonetic approach. It has been established that artificial intelligence is also mentioned as knowledge based approach and it uses the information, which is related to spectrogram, phonetic, and linguistic[16][17].

Artificial intelligence approach plays an essential role in different activities of speech recognition, including designing of recognition algorithm, demonstration of speech units, and representation of proper and appropriate inputs. It is considerable to bring into the notice that, some of all methods of speech recognition, artificial intelligence is the most reliable and proficient methods[17].

In general, artificial intelligence can be understood as the emerging and constantly developing fields of computer science. It has been examined that artificial intelligence mainly focus attention on the development of such machines, which are effective enough to get engross in the behaviours of people.

But this approach is not that much victorious in quantifying accomplished knowledge as compare with other two approaches. More well founded method for this type of approach is Artificial Neural Network method. Artificial Neural Network includes huge number of simple processing element that is called neurons. These neurons smack each other's performance via a network of excitatory weights. It is a feed - forth artificial neural network which possesses more than one layer of concealed units between its inputs and its outputs. Neural Network comes up with three types of learning methods which are supervised, unsupervised and reinforced.

3.3.3.1 Knowledge/Rule Based Approach

Knowledge /rule based approach is used for continuous speech recognition which has been proposed by several researchers and applied to speech recognition. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram [7]. The "expert" knowledge about deviation in speech is hand-coded into a system It uses set of features from the speech, and then the training system generates set of production rules automatically from the samples. These rules are derived from the parameters that provide valuable information about taxonomy. At the frame level, the effort of recognition is performed using an inference engine to implement the decision tree and classify the firing of the rules. This approach has the advantage of explicitly modelling variation in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach is well thought-out as impractical and automatic learning procedures were sought instead[9][13][10].

4. CONCLUSION

This paper offers the fundamentals of speech recognition system along with various approaches available for feature extraction and pattern matching has been discussed. Numerous advanced concept of classification techniques have been exercised freshly in speech recognition systems. With the help of these techniques, rate of speech recognition can be enhanced and better quality speech recognition can be developed. The difficulty always arises due to variation of the speech in time and environmental noise makes the recognition precision difficult. In future the focal point will be on development of large vocabulary speech recognition system and speaker independent continuous speech recognition system.

5. REFERENCES

[1] Ms.Ayushi Y. Vadwala, Ms.Krina A. Suthar, Ms.Yesha A. Karmakar, P. N. P. (2017). Intelligent Android Voice Assistant - A Future Requisite. *International Journal of Engineering Development and Research*, 5(Issue 3), 337-339.

[2] Radha, V., & Vimala, C. (2012). A review on speech recognition challenges and approaches. *doaj. org*, *2*(1), 1-7.

[3] Bhabad, S. S., & Kharate, G. K. (2013). An Overview of Technical Progress in Speech Recognition. *International Journal of advanced research in computer science and software Engineering*, *3*(3).

[4] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, *10*(3), 16-24.

[5] Kalamani, M., Valamrthy, S., Mohan, R., & Anitha, S. A review on clustering techniques in continous speech recognition.

[6] Nidhi Desai, Prof. Kinnal Dhameliya, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," *International Journal of Emerging Technology and Advanced Engineering (IJETAE)*, Vol. 3, Issue 12, December 2013.

[7] Gamit, M. R., Dhameliya, P. K., & Bhatt, N. S. (2015). Classification Techniques for Speech Recognition: A Review. *vol*, *5*, 58-63.

[8] Madan, A., & Gupta, D. (2014). Speech Feature Extraction and Classification: A Comparative Review. *International Journal of computer applications*, 90(9).

[9] Kaur, R. D. N. Speech Recognition Using Stochastic Approach: A Review.

[10] Bhabad, S. S., & Kharate, G. K. (2013). An Overview of Technical Progress in Speech Recognition. *International Journal of advanced research in computer science and software Engineering*, 3(3).

[11] Tran, D. T. (2000). Fuzzy Approaches to Speech and Speaker Recognition(Doctoral *dissertation, university of Canberra*).

[12] Gaudard, C., Aradilla, G., & Bourlard, H. (2007). Speech Recognition based on Template Matching and Phone Posterior Probabilities (No. LIDIAP-REPORT-2007-006). *IDIAP*.

[13] Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications*, 115(22).

[14] Kaur, P., Singh, P., & Garg, V. (2012). Speech recognition system; challenges and techniques. *International Journal of Computer Science and Information Technologies*, 3(3), 3989-3992.

[15] Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing*, 1(3), 195-304.

[16] Gohil, M. G. Artificial Intelligence for Speech Recognition.

[17] Alhawiti, K. M. (2015). Advances in artificial intelligence using speech recognition. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *9*(6), 1351-1354.

[18] Hsu, K. L., Gupta, H. V., & Sorooshian, S. (1995). *Artificial neural network modeling of the rainfall‐ runoff process. Water resources research, 31(10)*, 2517-2530.

[19] Pour, M. M., & Farokhi, F. (2009). *An advanced method for speech recognition*.

[20] Kamble, B. C. (2016). *Speech Recognition Using Artificial Neural Network–A Review*. *IEEE trans*.

[21] Forsberg, M. (2003). Why is speech recognition difficult? *Chalmers University of Technology*.

[22] Google voice search: faster and more accurate, *Google Research blog*, Thursday, September 24, 2015, https://research.googleblog.com/2015/09/google-voice-search-faster-and-more.html