INTERNATIONAL ACADEMIC RESEARCH JOURNAL
ISSN :2289-8433
International Academic Research Journal of Business and Technology
Journal homepage : www.iarjournal.com

Automatic Speech Recognition: A Review

Naima Zerari[1] , Bilal Yousfi[2] and Samir Abdelhamid[3]

[1]University Batna2 Algeria, [2] International Islamic University Malaysia , [3]University Batna2, Algeria

Corresponding email: n.zerari@yahoo.fr, yousfi.bilal@hotmail.com, Samir_abdel@hotmail.com

Article Information

Keywords
Speech, Automatic Speech Recognition, Utterance, Accuracy.

Abstract

Speech is the most natural way of human communication and also the most efficient form of exchanging information. Speech can be identified and converted to a machine readable format via a technology called speech recognition, or speech to text (STT). This inter-disciplinary sub-field of computational linguistics is one of the premier applications for machine learning and pattern technology. Speech recognition is defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. This conversion is accomplished according to one of the three approaches: the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach. The objective of this review paper is to present the basic idea of speech recognition, to give a description of existing definitions on Speech Recognition and to discuss the major approaches.

INTRODUCTION

Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program (Sanjivani S. Bhabad and Gajanan K. Kharate, 2013). Due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. Therefore several research efforts have been oriented to this area where computer scientists have been researching ways and means to make computers able to record, interpret and understand human speech. It has been an intensive research area for decades. ASR system includes two phases. Training phase and Recognition phase. In training phase, known speech is recorded, and then the features (parametric representation of the speech) are extracted and stored in the speech database. In the recognition phase, the features of the given input speech signal are extracted and compared with the reference templates (stored in the speech database) to recognize the utterance.

Speech recognition process contains different phases to process the speech signal which are the acoustic pre-processing, feature extraction, feature classification and the recognition/identification as shown in Fig.1. Acoustic pre-processing is a phase of filtering speech signal to eliminate unwanted noise sounds. The pre-processing identifies the word boundaries to determine the start and the end point of word utterances before it can be analyzed in feature extraction process. It represents the computation of a sequence of feature vectors (used to differentiate between different words) and provides a compact representation of the given speech signal. Feature classification is the process for determining similarities spoken words between an input feature vector sequence and a set of acoustic models to determine the word which was supposed to have been delivered.

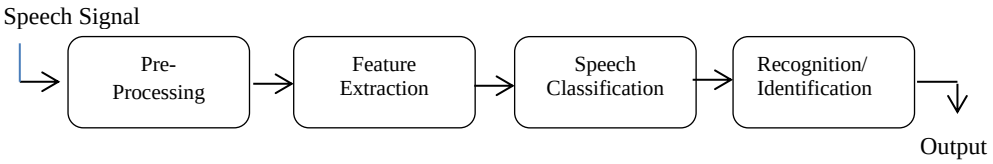

Fig. 1. Speech Recognition Process

TYPES OF SPEECH UTTERED

Different classes of speech recognition system can be made according to the type of utterance they have ability to recognize.

a. Isolated speech

Isolated word recognizer usually sets necessary condition that each utterance having little or no noise on both sides of sample window. It requires single utterance at a time. Often, these types of speech have "Listen/Not-Listen states", where they require the speaker to have pause between utterances. Isolated word might be better name for this type (Nidhi Desai et al, 2013).

b. Connected words

Connected words require minimum pause between utterances to make speech flow smoothly. They are almost similar to isolated words (Kishori R. Ghule & R. R. Deshmukh, 2015). Connected word speech recognition is a class of fluent speech strings where the set of strings is derived from small-to-moderate size vocabulary such as digit strings, spelled letter sequences, combination of alphanumeric (Wiqas Ghai & Navdeep Singh, 2012)**.**

c. Continuous speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content (Basically, it's computer dictation). Continuous speech recognition deals with the speech where words are connected together instead of being separated by pauses. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries (Wiqas Ghai & Navdeep Singh, 2012) (Santosh K.Gaikwad et al, 2010) (Shanthi Therese S. & Chelpa Lingam, 2013).

d. Spontaneous speech

A spontaneous speech is a speech which is natural sounding and not rehearsed (Wiqas Ghai & Navdeep Singh, 2012). An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run at the same time (Nidhi Desai et al, 2013) (Kishori R. Ghule & R. R. Deshmukh, 2015) (Santosh K.Gaikwad et al, 2010).

PARAMETERS AFFECTING THE ACCURACY OF RECOGNITION

In a speech recognition system, many parameters affect the accuracy of recognition such as vocabulary size, speaker dependency, speaker independency, time for recognition, type of speech (isolated, connected, continuous, and spontaneous), recognition environment condition (quit, noise) and different pronunciations of one word by one person in several times etc.

AUTOMATIC SPEECH RECOGNITION SYSTEM CLASSIFICATIONS

Speech Recognition systems can be classified as shown in Fig.2. Where Speech recognition is one of the most important areas in digital signal processing and is highly demanded technology, which consists of many useful applications.

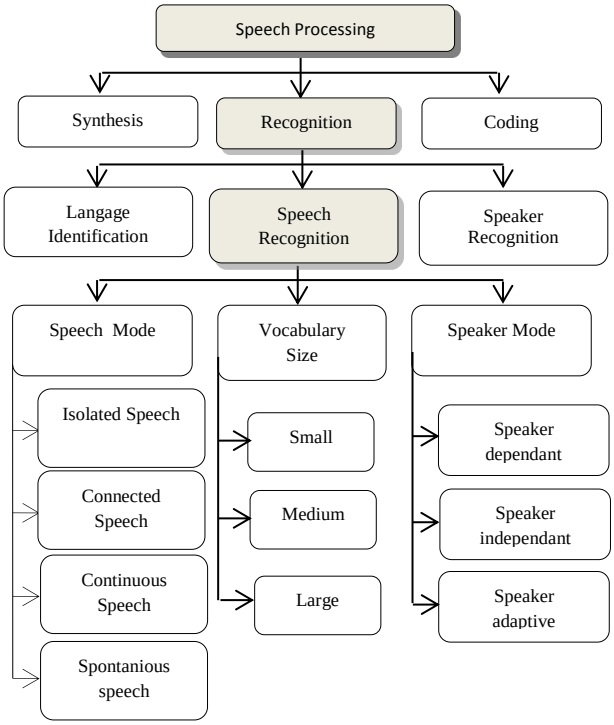

Fig.2**.** Speech Processing Classification

APPROACHES TO AUTOMATIC SPEECH RECOGNITION BY MACHINE

Broadly speaking, there are three approaches to speech recognition, as illustrate by Fig3, namely (Rabiner L. R. & B. H. Juang, 1993) :

1. The acoustic-phonetic approach
2. The pattern recognition approach
3. The artificial intelligence approach

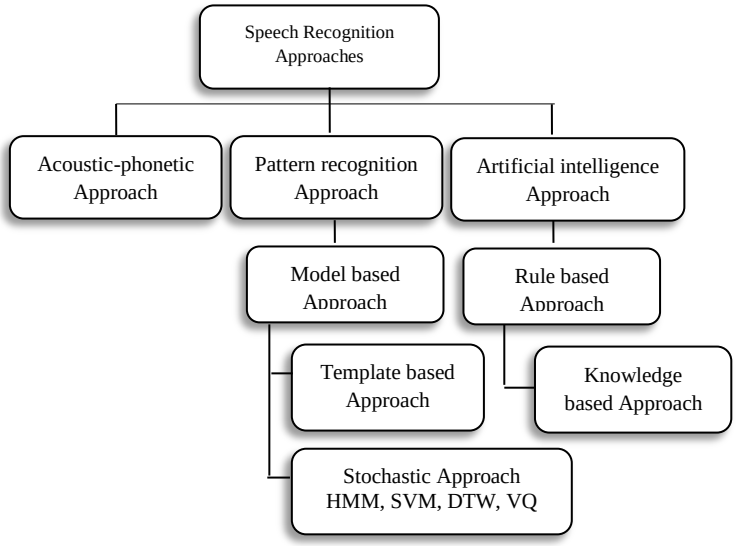

Fig.3. Speech Recognition approaches

1. The acoustic-phonetic approach

The term acoustic is deals with the different sounds in speech and phonetic is the study of phonemes in the language. Acoustic phonetic approach is based on the fact that, there exist finite and distinctive phonetic units in spoken language and these units are broadly characterized by a set of acoustic properties that are manifest in the speech signal over time. The acoustic properties of phonetic units depend on the speaker and co-articulation

effect. Even though the acoustic properties of phonetic units are highly variable, the rules governing their variability can be learned and applied in practical situations (Nidhi Desai et al, 2013) ((Kishori R. Ghule & R. R. Deshmukh, 2015) (Rabiner L. R. & B. H. Juang, 1993). The acoustic phonetic approach has not been widely used in most commercial applications (Santosh K.Gaikwad et al, 2010). Different steps are included in the acoustic phonetic approach to speech recognition as follow (Nidhi Desai et al, 2013) (Kishori R. Ghule & R. R. Deshmukh, 2015) (Rabiner L. R. & B. H. Juang, 1993):

a. Segmentation and labelling the speech into discrete regions where the acoustic properties of the signal are representative of phonetic units. Then attaching phonetic labels to each segmented region according its properties.
b. Determination of string of words: determines the string of words from the sequence of phonetic labels produced in the first step.

The real problem with this method is the difficulty in getting a reliable phoneme lattice for the lexical access stage.

2. The pattern recognition approach

The essential feature of this approach is that the speech patterns are used directly without feature determination and segmentation. This approach involves two important steps: pattern training and pattern comparison. In the first step, if enough versions of patterns are included in a training set provided to the algorithm, the system will characterize the acoustic properties of the pattern. This type of characterization of speech is called classification because the machine knows which property of the speech class is reliable and repeatable through all training taken of the pattern. Whereas pattern comparison step includes a direct comparison of the unknown speech with each possible pattern learned in the training step and classifies the unknown speech according to the goodness of match of the patterns. The pattern recognition approach contain two methods namely template approach and stochastic approach.

a. Template based approach

This approach has provided a family of techniques that have advanced the field considerably during the last six decades. The underlying idea is simple. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Usually templates for entire words are constructed. This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. In turn, each word must have its own full reference template; template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. One key idea in template method is to derive typical sequences of speech frames for a pattern (a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker. But it also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical (Santosh K.Gaikwad et al, 2010).

b. Stochastic approach

Stochastic modelling entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition (Sanjivani S. Bhabad & Gajanan K. Kharate, 2013). This approach cover different methods like HMM, SVM, DTW, and VQ. Among all these methods hidden markov model is the most used in speech recognition (Kishori R. Ghule & R. R. Deshmukh,, 2015). This method is characterized by a finite state markov model and a set of output distributions.

The transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities. These two types of variabilities are the essence of speech recognition.

3. The artificial intelligence approach

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. This approach attempts to mechanize the recognition procedure that depend to the way a person applies intelligence in visualizing, analyzing and finally making a decision on the measured acoustic features (Santosh K.Gaikwad et al, 2010) (Jamaliah Ibrahim Noor et al, 2013). A More reliable method for this type of approach is Artificial Neural Network (ANN) (Kishori R. Ghule & R. R. Deshmukh, 2015) (Rabiner L. R. & B. H. Juang, 1993). Artificial Neural networks are represented as a set of nodes called neurons and a set of connections between them. The connections have weights associated with them, representing the strength of those connections (Irfan Y. Khan et al, 2013). Most neural network architecture has three layers in its structure. First layer is the input layer which provides an interface with the environment, the second layer is the hidden layer where computation is done and the last layer is the output layer where output is stored as shown by Fig.4. Data is propagated through successive layers, with the final result available at the output layer. Many different

types of neural networks are available whereby multi-layer neural networks are the most popular. Their popularity is due to more than one hidden layer in their structure which helps sometimes to solve complex problems that a single hidden layer neural network cannot solve (Koushal Kumar & Abhishek , 2012).

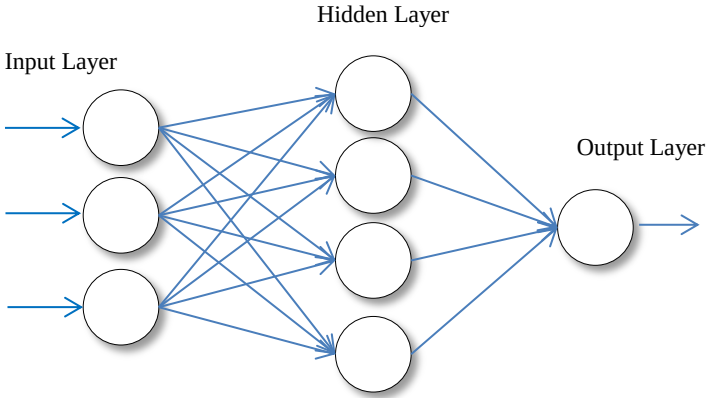

Fig..4. ANN Architecture

PERFORMANCE EVALUATION OF SPEECH RECOGNITION SYSTEMS

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor (M.Kalamani et al, 2014).

a. WER

Word Error Rate is a common metric of the performance of a speech recognition or machine translation system. WER is derived from the Levenshtein distance (is a measure of the similarity between two strings, the source string (s) and the target string (t)). The distance is the number of deletions, insertions, or substitutions required to transform (s) into (t) (Rishin Haldar & Debajyoti Mukhopadhyay, 2011). The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). Word error rate can then be computed as (Bhabad & Kharate, 2013) (Shanthi Therese S. & Chelpa Lingam, 2013) (M.Kalamani et al, 2014):

$$WER = (S+D+I)/N \qquad (1)$$

Where

N: the number of words in the reference.
S : the number of substitutions (incorrect words substituted)
D: the number of the deletions (words deleted)
I: the number of the insertions (extra words inserted).

The WER threshold for acceptable performance is different for different applications. It has been shown that good document retrieval performance is possible even with a 66% WER but that precision begins to fall off rapidly when WER gets above 30% (Randall Fish et al, 2006).

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead (Santosh K.Gaikwad et al, 2010) :

$$WRR = 1- WER = (N-S-D-I)/N \qquad (2)$$

b. Speed

Real Time Factor is parameter to evaluate speed of automatic speech recognition. It is defined as (Wiqas Ghai & Navdeep Singh, 2012) :

$$RTF = P/ I \qquad (3)$$

Where

P: the time taken to process an input
I: duration of the input

CONCLUSION

Speech recognition is one of the most integrating areas of machine intelligence since humans carry out daily activities of speech recognition. It has attracted scientists as an important discipline and has created a technological impact on society. The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, databases and performance evaluation. An attempt has been made through this paper to present mainly the core architecture of automatic speech recognition that is essential for its comprehension. The classification of automatic speech recognition and the description of its various approaches have also been tackled by mean of this paper.

REFERENCES

Irfan Y. Khan, P.H. Zope and S.R. Suralkar, (2013). Importance of Artificial Neural Network in Medical Diagnosis disease like acute nephritis disease and heart disease, *2*(2), 210–217.

Jamaliah Ibrahim Noor, Yamani Idna Idris Mohd Razak Zaidi and Naemah and Abdul Rahman Noor, (2013),"Automated tajweed checking rules engine for Quranic learning", Multicultural Education & Technology Journal, Vol. 7 Iss 4 pp. 275 – 287.

Kishori R. Ghule and R. R. Deshmukh, (2015). Feature-Extraction-Techniques-for-Speech-Recognition-A-Review. *International Journal of Scientific & Engineering Research*, *6*(5), 143–147.

Koushal Kumar and Abhishek, (2012). Artificial Neural Networks for Diagnosis of Kidney Stones Disease, (July), 20–25.

M.Kalamani, S.Valarmathy, C.Poonkuzhali and Catherine J N, (2014). Feature Selection Algorithms for Automatic Speech Recognition. *International Conference on Computer Communication and Informatics (ICCCI).*

Nidhi Desai, Kinnal Dhameliya and Vijayendra Desai, (2013). Feature Extraction and Classification Techniques for Speech Recognition : A Review. International Journal of Emerging Technology and Advanced Engineering, 3(12), 367–371.

Rabiner L. R. and B. H. Juang, (1993). Fundamentals of Speech Recognition Englewood Cliffs, NJ: Prentice-Hall.

Randall Fish, Qian Hu and Stanley Boykin. Navdeep Singh, (2006). Using Audio Quality To Predict Word Error Rate In An Automatic Speech Recognition System. The MITRE Corporation.

Rishin Haldar and Debajyoti Mukhopadhyay, (2011). Levenshtein Distance Technique in Dictionary Lookup Methods : An Improved Approach, ([Online]. Available: http://arxiv.org/abs/1101.1232).

Sanjivani S. Bhabad and Gajanan K. Kharate, (2013). An Overview of Technical Progress in Speech Recognition. International Journal of Advanced Research in Computer Science and Software Engineering, 3(3), 2277–128.

Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications* (0975 – 8887).

Shanthi Therese S. and Chelpa Lingam, (2013). Review of Feature Extraction Techniques in Automatic Speech Recognition. *Engineering and Technology*, *484*(2), 479–484.

Wiqas Ghai and Navdeep Singh, (2012). Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications, 41*(8), 42–50.