International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 03, Issue: 04 (JULY–AUGUST, 2017)

STUDY ON SPEECH RECOGNITION SYSTEMS

Geetha K
Department of Computer Science
D. J. Academy for Managerial Excellence
Coimbatore India
Geethakab [AT] gmail [Dot] com

Vadivel R
Department of Information Technology
Bharathiar University
Coimbatore India
vlr_vadivel [AT] yahoo [Dot] co [Dot] in

***Abstract* —** Speech processing is one of the most exciting research areas of signal processing. Speech processing is the study of speech signals and the processing methods of speech signals. Since processing methods handle digital form of the speech signals, speech processing can be treated as an intersection of digital signal processing and natural language processing. Some of the speech processing technologies are Automatic Speech Recognition (ASR), digital speech coding, spoken language dialog systems, text-to-speech synthesis. In addition, information such as speaker, gender or language identification can also be extracted from speech signals. In this paper, architecture of the general speech recognition system and the issues related to it are presented.

Keyword — *Automatic Speech Recognition System, Acoustic Model, Lexical Model, Language model.*

INTRODUCTION

In the early days, the field of computing dominated by vector-oriented processors and linear algebra-based mathematics, but in the current scenario, Digital Signal Processing based systems rely on sophisticated statistical models implemented using a complex software paradigm. These systems are now capable of understanding not only the isolated or connected speech input, but also continuous and also spontaneous speech input for vocabularies of several thousand words in operational environments [1]. Purpose of speech processing: *(a)* to understand speech as a means of communication. *(b)* To represent speech for transmission and reproduction. *(c)* To analyze speech for automatic recognition and extraction of information. *(d)* To discover some physiological characteristics of the talker.

Automatic Speech Recognition including Automatic Speech Segmentation (ASS) faces many problems. In natural speech, there are no pauses between phonemes or even parts of words and are recognized by their context phones or words. Co-articulation often causes canonically expected phones to be modified or completely go missing since the speech organs like lips, tongue, etc. are in continuous movement during the pronunciation of words. Loudness pitch, duration and other elements of prosody such as voice quality which all affect the physical attributes of the signal and carry additional linguistic information.

APPROACHES OF ASR

There are three types of approaches in speech recognition systems [2],

a. Acoustic-Phonetic Approach
b. Pattern Recognition Approach
c. Artificial Intelligence Approach

a. Acoustic-phonetic Approach

Acoustic phonetic systems use knowledge of the human body such as speech production, hearing etc. to compare speech features. For every spoken language, there exist a fixed number of distinctive phonetic units. These phonetic units are broadly characterized by a set of acoustics properties varying with respect to time in a speech signal. In this approach, acoustic properties like nasality, frication, voiced-unvoiced classification and continuous features such as formant locations, ratio of high and low frequencies can be analyzed to find the phonetic units. Figure 1 shows the diagram of Acoustic-Phonetic Speech recognition System.

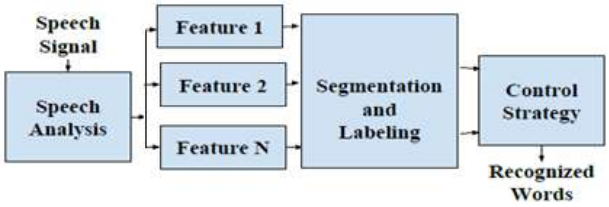

Fig. 1 Acoustic-Phonetic Speech Recognition System

b. Pattern Recognition Approach

Pattern training and pattern matching are the two essential steps in this approach. It uses a well formulated mathematical framework to develop a consistent speech pattern representation for a set of labeled speech training samples via a formal training algorithm. Steps involved in pattern matching are

1. Parameter measurement
2. Compare the patterns
3. Decision making.

Figure 2 shows the diagram of Pattern Recognition Approaches for Speech Recognition System. Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) are the two popular non linear pattern matching technique used in speech processing [chang 2009].

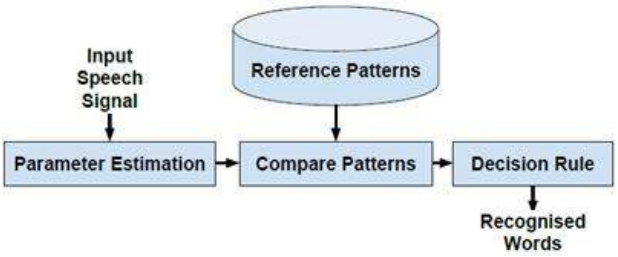

Fig. 2 Pattern Recognition Approach for Speech Recognition System

d. Artificial Intelligence Approach

Artificial intelligence approach to speech recognition is a hybrid of the acoustic-phonetic approach and the pattern recognition approach in that it exploits ideas and concepts of both methods. Both acoustic phonetic and template based approach failed at their own to explore considerable insight into human speech processing. As a result, error analysis and knowledge based system enhancement couldn't get strength. In traditional Knowledge based approach, the production rules are created heuristically from empirical linguistic knowledge or from the observations from the speech spectrogram. Knowledge helps the algorithm to perform better and also in the selection of a suitable input representation, the definition of units of speech and the design of the recognition algorithms. Most the new speech recognition systems are based on hybrid approach Hidden Markov Model/Artificial Neural Network (HMM/ANN) [3]. HMM has a great capacity to treat events in time, while ANN is an expert in the classification of patterns. In recent years, Deep Neural Networks (DNN) has driven tremendous improvements in large vocabulary continuous speech recognition (LVCSR) Systems [4].

ARCHITECTURE OF SRS

A high level overview of speech recognition systems is presented in figure 3. In the general speech recognition system there are two phases and they are speech training and speech testing. The major components of a typical speech recognition system are the feature extraction, construction of pronunciation dictionary, language models, acoustic model, the design of the decoder to find the text form of the spoken speech. The digitized speech signal is first transformed into a set of useful measurements or features at a fixed rate typically one every 10-20 mili seconds. These measurements are then used to search for the most likely word, making use of acoustic, lexical and language models. Throughout the process, training data are used. Speech acquisition and feature extraction modules are common to both training and testing phases of ASR. Mel Frequency Cepstral Coefficients, Linear Predictive Cepstral Coefficients(LPCC), Perceptual linear prediction (PLP), Relative Spectral Transform Perceptual Linear Prediction(RASTA-PLP), Digital filter banks are some of the features extraction techniques using in speech processing[5][6][7].

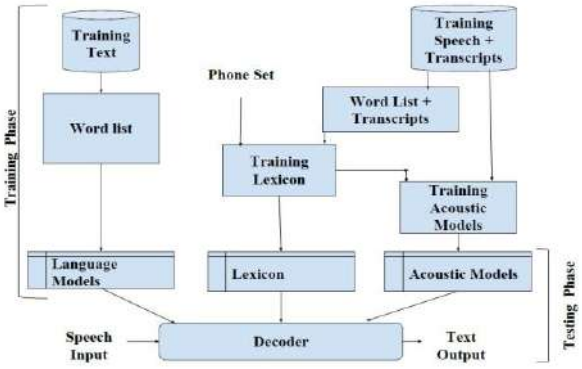

Figure 3 Overview of the General Speech Recognition System

One of the key issues in acoustic modeling has been the decision of a selection of the unit. In words based acoustic model, as the vocabulary increase, the models to be trained are also increased and it is tedious to training data to build all words. But in small vocabulary systems of a few words, it is possible to build separate models. To build large vocabulary systems, it is necessary to represent words in terms of sub-word units.

Acoustic model is one of the most important knowledge sources for automatic speech recognition system, which represents acoustic features for phonetic units to be recognized. In building an acoustic model, one fundamental and important issue is choosing of basic modeling units. Generally, if the target language of the speech is specified, there are different types of sub word phonetic units can be suggested for acoustic modeling. Using different acoustic model units can result in difference in the performance.

Hidden Markov Model is one most common type of acoustic models. Other acoustic models include segmental models, super-segmental models, neural networks, maximum entropy models and conditional random fields, etc. An acoustic model maintained if a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label. An acoustic model is created by taking a large database of speech and using special training algorithms to create statistical representations for each basic linguistic unit considered.

These statistical representations are called Hidden Markov Models. HMMs are constructed for each basic unit considered to build ASR [3]. The speech decoder finds the probability of the distinct sounds spoken by

a user and compares with all HMMs and selects the unit which has the highest probability. Later, Gaussian Mixer model were used in computing the posterior probability [8]. To speed up statistical pattern classification, by reducing the time consumed in likelihood evaluations of feature vectors, by using optimal number of Gaussian mixture components selected on the basis of empirical observations, was tried[9]

An acoustic model, all phonetic states share a common Gaussian Mixture Model structure, in which the means and mixture weights vary in a subspace of the total parameter space is called as a Subspace Gaussian Mixture Model (SGMM) in which globally shared parameters define the subspace. This type of acoustic model suits, when the training data is less, since it allows for a compact representation of the signal and gives better results.

Even though there are similar sounding phone in a speech, people generally do not find it difficult to recognize the word, since they know the context. It is because of the prediction of the words or phrases which can occur in the context. The same phenomenon can be used in ASR to predict the most relevant phone or word to be occurred next. Providing this context to a speech recognition system is the purpose of language model. The language model specifies what the valid words are in the language and in what sequence they can occur.

Language models are used to constrain search in a decoder by limiting the number of possible words that need to be considered at any one point in the search which leads to faster execution and higher accuracy. They constrain search can be computed either by enumerating some small subset of possible expansions or by computing a likelihood for each possible successor word. The former will usually have an associated grammar. This is compiled down into a graph; the latter will be trained from a corpus.
Language model is the single largest component trained on billions of words, consisting of billions of parameters and developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary. ASR systems utilize *n*-gram language models to guide the search for correct word sequence by predicting the likelihood of the n^{th} word on the basis of the $n-1$ preceding words. The probability of occurrence of a word sequence W is calculated as:

$$\boldsymbol{P(w) = P(w_1, w_2, \ldots w_{n-1}, w_n)}$$

$$\boldsymbol{= P(w_1).P(w_2|w_1).P(w_3|w_2w_1) \ldots P(w_n|w_{n-1} \ldots w_1)}$$

Language models are cyclic and non-deterministic. Both these features make it complicated to compress its representations. Language models can be classified into several categories that include

- **Uniform Models**: Each word has equal probability of occurrence.
- **Stochastic Models**: Probability of occurrence of a word depends on the words preceding it.
  - P(w) = probability of word w
  - $P(w_j/w_i)$ = probability of w_j given a one word history w_i
  - $P(w_k/w_i,w_j)$ = probability of wk given a two word history w_i; $_{wj}$
- **Finite State Languages**: Language uses a finite state network to define allowed word sequences.
- **Context Free Grammar**: Context free grammar can be used to encode sentences.

Lexicon is developed to provide the pronunciation of each word in a given language. Through lexical model, various combinations of phones are defined to give valid words for the recognition. Neural networks have helped to develop lexical model for non-native speech recognition.

ISSUES IN SPEECH PROCESSING

Duration variability: At the physical level of natural speech, the rate of speech is governed by the inertia of the articulators. The body of the tongue moves relatively slowly and the rate of sonorant phones is limited by the rate at which the tongue moves. The lips and tongue can move faster and so plosive sounds occur over a much shorter time interval. In addition to durational variation due to phone differences, vowel duration may change by a factor of eight, depending on speaking rate, syntax and stress. The factors that influence the phone duration while speech production and speech perception which result in fairly complex models. Van Santen et al [10] proposed a model to account for 86% of the variance of vowel durations only in a large corpus of manually segmented speech.

Effect of co-articulation Co-articulation is the impact of one phone to its neighboring phone which is showed as a smooth change in formant frequencies of the current phone to the next phone in the context. This smooth transition between phones makes difficult to determine the exact location of phonetic boundary. Ohman [11] proposed a model to handle co articulation in Vowel-Consonant-Vowel (VCV) utterances using a vocal tract shape based information. Even though this model was successful on VCV utterances, it had its limitations in handle co-articulation effects in Consonant-Vowel-Consonant (CVC) utterances.

Speaker Variability: In speech processing, the speech signal not only conveys the linguistic information but also information about the speaker like gender, age, social, and regional origin, health and emotional state, etc.

Pronunciation Variability: Following co-articulation and pronunciation effects, speaker related spectral characteristics have been identified as another major dimension of speech variability. Speaking faster or slower also has influence on the speech signal. This impacts both temporal and spectral characteristics of the signal, both affecting the acoustic models. Obviously, faster speaking rates may also result in more frequent and stronger pronunciation changes.

Vocabulary: Vocabularies or Dictionaries are list of words or utterance that can be recognized by the Speech Recognition (SR) system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as sentence. Smaller vocabularies can have as few as one or two recognized utterances while very large vocabularies can have a hundred thousand or more. Recognition is generally more difficult when vocabularies are large or have many similar-sounding words. However, the vocabulary size is not a reliable measure of task complexity. The grammar constraints can also influence the complexity of the system.

Speaking Mode – Isolated word vs. connected word: Early Systems used discrete speech, in which the user had to speak one word at a time, with a pause between words (e.g., digit recognition). In 1952, Davis, K. H. et. al. has tried an isolated digit recognition system for a single speaker [12]. Connected word systems recognize separate utterances to be 'run-together' with a minimal pause between them.

Rabiner et al [2] analyzed three algorithms designed for connected word recognition: Two level Dynamic Programming (DP) approach, Level Building approach and One Pass approach. These three algorithms are found to be provided identical best matching string with the identical matching score for connected word recognition.

Like isolated word speech recognition, it also requires the basic input speech utterance as a word or phrase. HMM/Continuous Density (HMM/CD) and Linear Predictive cepstral coefficients/Dynamic Time Warping (LPCC/DTW) are the best models to construct isolated ASR model [2] [13]. This type of ASR suits for command and control applications.

Speaking Style - Continuous speech vs. Spontaneous speech: Continuous ASR allows the user to speak in a more natural way and the word boundaries are not so evident. The level of difficulty also varies due to the type of interaction. That is, recognizing speech from human-human interactions like recognition of conversational telephone speech, broadcast news etc. is more difficult than human-machine interactions like dictation software. Spontaneous speech is much more difficult to recognize than speech read from script. In this type, the speech signal to be recognized is natural sounding and not rehearsed. An ASR system for such speech handles a variety of natural speech features. Large structured collection of speech is essential. Labeling and annotation of spontaneous speech is difficult. Some points to be noted are how to handle extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, repetitions, style shifting etc. Due to these factors the performance of spontaneous ASR gets degraded.

Speaker mode: Speech recognition systems can be either speaker dependent or speaker independent. Speaker dependent system requires to be trained on the user, and generally they achieve better recognition results. The speaker dependent ASR is usually easier to develop, less expensive. In speaker independent systems, training is not required for the user and some speaker-independent ASR systems uses adaptation techniques to improve the recognition performance.

Channel type: The characteristics of the channel can affect the speech signal and it may range from telephone channels to wireless channels with fading and with a sophisticated voice.

Transducer type: The quality of the device used to record the speech also influence the performance of the speech recognition systems. Some of them are high-quality microphones, telephones, cell phones, array microphones, etc.

Environment variability: Speech recognitions systems are also suppressed with the background noise and suppressed noise which leads the research in noise reduction techniques.

These issues conclude that on these different conditions of speech, it is highly complex and difficult to understand and build models.

APPLICATIONS OF SRS

Hands-free computing is possible in computes user ASR. In the health care monitoring systems, speech recognition can be implemented as front-end or back-end of the medical documentation process. This kind of innovation can assist those with dyslexia to command and control appliances like wheel bed, chair etc. Training for Air Traffic Controllers (ATC) is also an excellent application for speech recognition systems. It is used in education, to learn second language and to improve the pronunciation skill. With the exponential development of information and processing power, ASR innovation has progressed to the new phase where more challenging applications are turning into a reality. There are new direction of speech recognition in mobile devices [14],

multilingual speech recognition systems [15] and structures like Embedded Speech Recognition Systems, Network Speech Recognition (NSR) and Distributed Speech Recognition (DSR).

PERFORMANCE OF THE SRS

The performance of the speech recognizer is measured in terms of Word Error Rate (WER) and Word Recognition Rate (WRR). In Equation (1), the WER [16] is defined as

$$WER = \frac{S+I+D}{N} \times 100 \quad (1)$$

Where N is the total number of words in the test set, and S, I and D are the total number of substitution, insertions and deletions respectively. Word Recognition Rate is defined as [17]

$$WRR = \frac{N-(S+I+D)}{N} \quad (2)$$

CONCLUSION

Approaches, overview of a Speech Recognition System, challenges in developing an ASR system and application are presented in this review paper. Building a model to recognize spoken language as human being is really a complicated task. Robust and multilingual speech recognition systems are also the emerging research areas of speech processing.

REFERENCE

[1] X. Huang and L. Deng, “An Overview of Modern Speech Recognition”, Handbook of Natural Language Processing||, Second Edition, Chapter 15, Chapman & Hall/CRC, 2010.

[2] R. Lawrence and B.-H. Juang , “Fundamentals of Speech Recognition||”, Prentice-Hall, Inc., Engel wood, NJ, 1993

[3] Xi, Xiaoping, et al., "A new hybrid HMM/ANN model for speech recognition.” IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Boston, MA, 2005.

[4] Maas, Andrew L., et al., "Building DNN acoustic models for large vocabulary speech recognition.” Computer Speech & Language (2017): 195-213.

[5] Anusuya, M. A., and S. K. Katti, "Front end analysis of speech recognition: a review.” International Journal of Speech Technology ,14.2 (2011), 99-145.

[6] Shrawankar, Urmila, and Vilas Thakare. "Feature extraction for a speech recognition system in noisy environment: A study.” Computer Engineering and Applications (ICCEA), 2010 Second International Conference on. Vol. 1. IEEE, 2010.

[7] Pujol, Pere, et al, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system.” IEEE Transactions on Speech and Audio processing 13.1 (2005), 14-22.

[8] Aggarwal, R. K., and M. Dave, “Using Gaussian mixtures for Hindi speech recognition system”, International Journal of Signal Processing, Image Processing and Pattern Recognition, 4.4, 157-170, 2011.

[9] Povey, Daniel, “Subspace Gaussian mixture models for speech recognition”, Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference, 2010.

[10] Van Hemert., Jan P., “Automatic segmentation of speech”, IEEE Transactions on Signal Processing, 39.4, 1008-1012, 1991.

[11] S. E. G.¨Ohman. “Co-articulation in VCV utterances: Spectrographic measures”, Journal of Acoustical Society of America, 39:151–168, 1966

[12] Davis, K. H., R. Biddulph, and Stephen Balashek. "Automatic recognition of spoken digits.” The Journal of the Acoustical Society of America, 24.6 (1952): 637-642.

[13] Fang, Chunsheng. "From dynamic time warping (DTW) to hidden markov model (HMM)." University of Cincinnati 3 (2009): 19.

[14] Schmitt, A., Zaykovskiy, D. & Minker, W., Speech Recognition for Mobile Devices, Int J Speech Technology (2008) 11: 63. doi:10.1007/s10772-009-9036-6

[15] Vu, Ngoc Thang, et al., "Multilingual deep neural network based acoustic modeling for rapid language adaptation.” Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.

[16] L. R. Rabiner and B. H. Juang, “Statistical Methods of Speech Recognition”, Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005

[17] Anusuya, M. A., and Shriniwas K. Katti. "Speech recognition by machine, a review.” arXiv preprint arXiv: 1001.2267 (2010).