A REVIEW ON SPEECH TO TEXT CONVERSION METHODS

Miss.Prachi Khilari[1]
[1]Department of E&TC Engineering.
G.H.R.C.O.E.M, Ahmednagar.
Savitribai Phule University of Pune.

Prof. Bhope V. P.[2]
[2]Department of E&TC Engineering.
G.H.R.C.O.E.M, Ahmednagar.
Savitribai Phule University of Pune.

ABSTRACT:

Speech is the first important primary need, and the most convenient means of communication between people. The communication among human computer interaction is called human computer interface. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech to text conversion and also gives overview technique developed in each stage of classification of speech to text conversion. A comparative study of different technique is done as per stages. This paper concludes with the decision on future direction for developing technique in human computer interface system in different mother tongue and it also discusses the various techniques used in each step of a speech recognition process and attempts to analyze an approach for designing an efficient system for speech recognition. However, with modern processes, algorithms, and methods we can process speech signals easily and recognize the text. In this system, we are going to develop an on-line speech-to-text engine. However, the transfer of speech into written language in real time requires special techniques as it must be very fast and almost 100% correct to be understandable. The objective of this review paper is to recapitulate and match up to different speech recognition systems as well as approaches for the speech to text conversion and identify research topics and applications which are at the forefront of this exciting and challenging field.

Keyword : Speech To Text conversion, Automatic Speech Recognition, Speech Synthesis.

I. INTRODUCTION :

In modern civilized societies for communication between human speeches is one of the common methods. Different ideas formed in the mind of the speaker are communicated by speech in the form of words, phrases, and sentences by applying some proper grammatical rules.The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. By classifying the speech with voiced, unvoiced and silence (VAS/S) an elementary acoustic segmentation of speech which is essential for speech can be considered. In succession to individual sounds called phonemes this technique can almost be identical to the sounds of each letter of the alphabet which makes the composition of human speech. Most of the Information in digital world is available to a few who can read or understand a scrupulous language. Language technologies can provide solutions in the form of ordinary interfaces so the digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages[4]. These technologies play a vital role in multi-lingual societies such as India which has about 1652 dialects/native languages. Speech to Text conversion take input from microphone in the form of speech & then it is converted into text form which is display on desktop. Speech processing is the study of speech signals, and the various methods which are used to process them. In this process various applications such as speech coding, speech synthesis, speech recognition and speaker recognition technologies; speech processing is employed. Among the above, speech recognition is the most important one. The main purpose of speech recognition is to convert the acoustic signal obtained from a microphone or a telephone to generate a set of words [13, 23]. In order to extract and determine the linguistic information conveyed by a speech wave we have to employ computers or electronic circuits. This process is performed for several applications such as security device, household appliances, cellular phones ATM machines and computers. Survey of these paper deals with different methods of speech to text conversion which is useful for different languages such as Phonem to Graphem method,conversion for Bengali language, HMM based speech synthesis methods etc[17].

1. Type of Speech:

Speech recognition system can be separated in different classes by describing what type of ullerances they can recognize [1].

1.1 Isolated Word:

Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time .This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class.

1.2 Connected Word:

Connected word system are similar to isolated words but allow to divide or separate sound to be “run together minimum pause between them.

1.3 Continuous speech:

Continuous speech recognizers allows user to talk almost naturally, while the computer determine the content. Recognizer with continues speech capabilities are some of the most difficult to create because they utilize unique sound and special method to determine utterance boundaries.

1.4 Spontaneous speech:

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR System with spontaneous speech ability should be able to handle a different words and variety of natural speech feature such as words being run together.

2. Types of Speaker Model:

All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into main categories based on speaker models, namely, speaker dependent and speaker independent [1].

2.1 Speaker independent models:

Speaker independent systems are designed for variety of speakers. It recognizes the speech patterns of a large group of people. This system is most difficult to develop, most expensive and offers less accuracy than speaker dependent systems. However, they are more flexible.

2.2 Speaker dependent models:

Speaker dependent systems are designed for a specific speaker. This systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems. They are generally more accurate for the particular speaker, but much less accurate for others speakers.

3. Types of Vocabulary:

The size of vocabulary of a speech recognition system affects the complexity, processing necessities, performance and the precision of the system. Some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. direction machines). In ASR systems the types of vocabularies can be classified as follows.

a. Small vocabulary - ten of words

b. Medium vocabulary - hundreds of words

c. Large vocabulary – thousands of words

d. Very-large vocabulary – tens of thousands of words

e. Out-of-Vocabulary – Mapping a word from the vocabulary into the unknown word.

Apart from the above characteristics, the environment variability, channel variability, speaker style, sex, age, speed of speech also make the ASR system more complex. But the efficient ASR systems must cope with the variability in the signal.

II. LITERATURE REVIEW:

1.Yee-Ling Lu, Man-Wai and Wan-Chi Siu explains about text-to-phoneme conversion by using recurrent neural networks trained with the real time recurrent learning (RTRL) algorithm [3].

2.Penagarikano, M.; Bordel, G explains a technique to perform the speech to text conversion as well as an investigational test carried out over a task oriented Spanish corpus are reported & analytical results also.

3.Sultana, S.; Akhand, M. A H; Das, P.K.; Hafizur Rahman, M.M. explore Speech-to-Text (STT) conversion using SAPI for Bangla language. Although achieved performance is promising for STT related studies, they identified several elements to recover the performance and might give better accuracy and assure that the theme of this study will also be helpful for other languages for Speech-to-Text conversion and similar tasks [3].

4.Moulines, E., in his paper "Text-to-speech algorithms based on FFT synthesis," present FFT synthesis algorithms for a French text-to-speech system based on diaphone concatenation. FFT synthesis techniques are capable of producing high quality prosodic adjustments of natural speech. Several different approaches are formulated to reduce the distortions due to diaphone concatenation.

5.Decadt, Jacques, Daelemans, Walter and Wambacq describes a method to develop the readability of the textual output in a large vocabulary continuous speech recognition system when out-of-vocabulary words occur. The basic idea is to replace uncertain words in the transcriptions with a phoneme recognition result that is post-processed using a phoneme-to-grapheme converter. This technique uses machine learning concepts.

III. SPEECH TO TEXT SYSTEM:

Speech is an exceptionally attractive modality for human computer interaction: it is “hands free”; it requires only modest hardware for acquisition (a high-quality microphone or microphones); and it arrives at a very modest bit rate. Recognizing human speech, especially continuous (connected) speech, without burdensome training (speaker-independent), for a vocabulary of sufficient complexity (60,000 words) is very hard. However, with modern processes, flow diagram, algorithms, and methods we can process speech signals easily and recognize the text which is talking by the talker. In this system, we are going to develop an on-line speech-to-text engine [4]. The system acquires speech at run time through a microphone and processes the sampled speech to identify the uttered text. The recognized text can be stored in a file. It can supplement other larger systems, giving users a different choice for data entry.

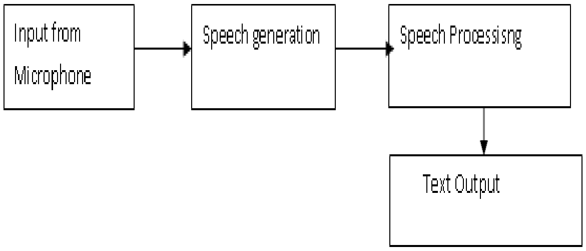

Figure 1 . Basic Block diagram of speech to text System.

A discourse to-content framework can likewise enhance framework availability by giving information passage alternatives to visually impaired, hard of hearing, or physically debilitated clients. Voice SMS is an application grew in this work that permits a client to record and believer talked messages into SMS instant message. Client can send messages to the entered telephone number. Speech recognition is done via the Internet, connecting to Google's server. The application is adapted to input messages in English. Speech recognition for Voice uses a technique based on hidden Markov models (HMM - Hidden Markov Model). It is currently the most successful and most flexible approach to speech recognition.

1. Speech production & Speech perception :

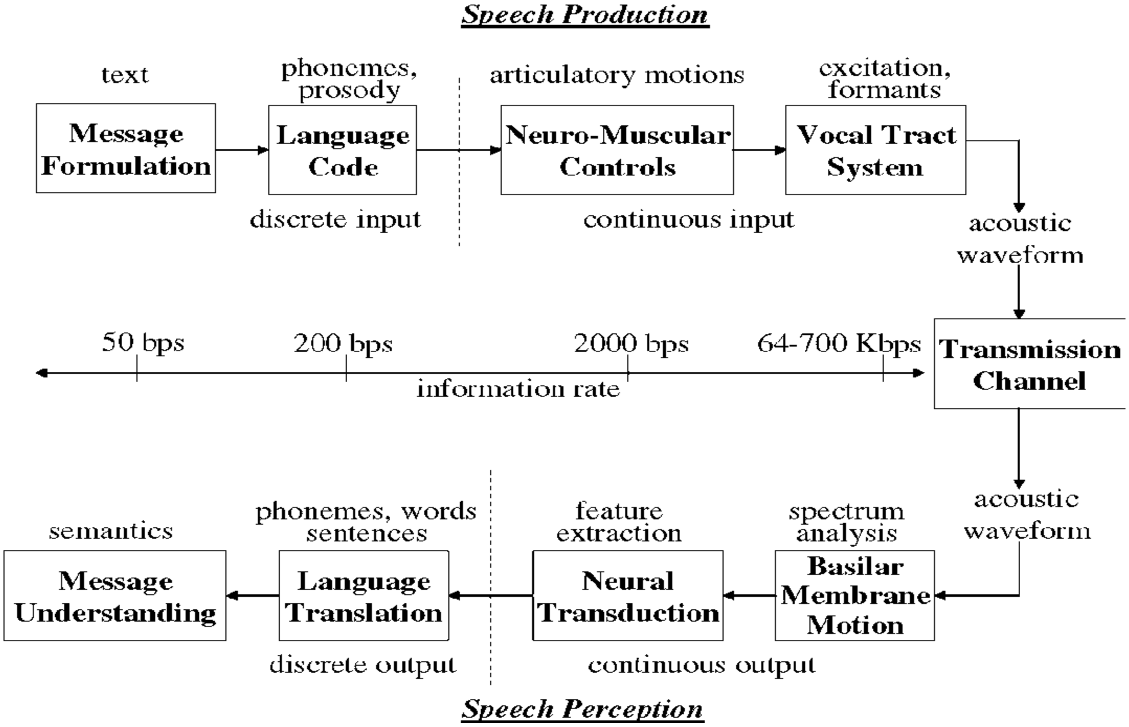

Figure 2 . Speech Chain

The process starts with the message information which can be thought of as having a number of different representations during the process of speech production. For example the message could be represented initially as English text. In order to “speak” the message, the talker implicitly converts the text into a symbolic representation of the sequence of sounds corresponding to the spoken version of the text. This step, called the language code generator which converts text symbols to phonetic symbol (along with stress and durational information) that describe the basic sounds of a spoken version of the message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced. The third step in the speech production process is the conversion to “neuro-muscular controls,” i.e., the set of control signals that direct the neuro-muscular system to move the speech articulators, namely the tongue, lips, teeth, jaw and velum, in a manner that is consistent with the sounds of the desired spoken message and with the desired degree of emphasis. The end result of the neuro-muscular controls step is a set of articulator motions (continuous control) that cause the vocal tract articulators to move in a prescribed manner in order to create the desired sounds. Finally the last step in the Speech Production process is the “vocal tract system” that physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform, such as the one shown in fig. that encodes the information in the desired message into the speech signal.

IV. AUTOMATIC SPEECH RECOGNITION (ASR):

1. Basic Principle:

ASR systems operate in two phases. First, a training phase, during which the system learns the reference patterns representing the different speech sounds (e.g. phrases, words, phones) that constitute the vocabulary of the application. Each reference is learned from spoken examples and stored either in the form of templates obtained by some averaging method or models that characterize the statistical properties of pattern [6]. Second, a recognizing phase, during which an unknown input pattern, is identified by considering the set of references.

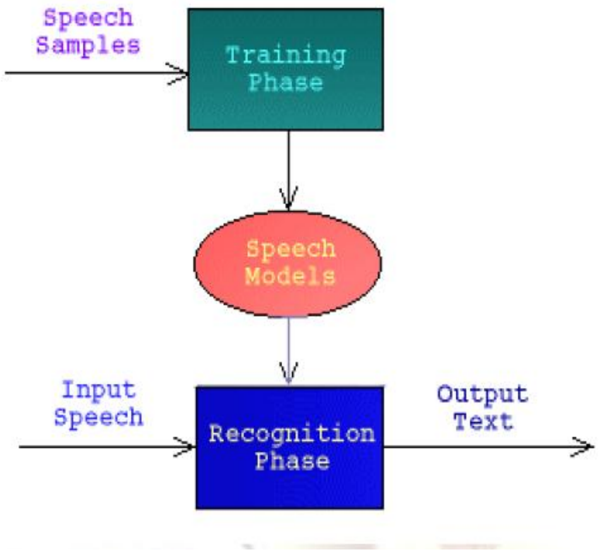

Figure 3 . Basic Principle of ASR

2. Speech Recognition Techniques:

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories. The goal of automatic speaker recognition is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognition system may be viewed as working in a four stages-

a. Analysis

b. Feature extraction

c. Modeling

d. Testing

a. Speech analysis:

In Speech analysis technique Speech data contains different types of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The physical structure and dimension of vocal tract as well as excitation source are unique for each speaker. This uniqueness is embedded in the speech signal during speech production and can be used for speaker used for speaker recognition.

b. Feature Extraction Technique:

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances[6]. These characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

a. Easy to measure extracted speech features.

b. It should not be susceptible to mimicry.

c. It should show little fluctuation from one speaking environment to another.

d. It should be stable over time.

e. It should occur frequently and naturally in speech.

V. SYSTEM DESIGN & IMPLEMENTATION :

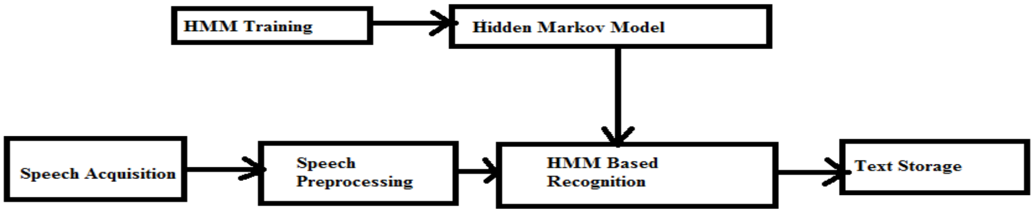

Figure 4 . System Architecture

The system equipped a speech-to-text system using isolated word recognition with a vocabulary of ten words (digits 0 to 9) and statistical modeling (HMM) for machine speech recognition. In the training phase, the uttered digits are recorded using 16-bit pulse code modulation (PCM) with a sampling rate of 8 KHz and saved as a wave file using sound recorder software. We use the MATLAB software's wavered command to convert the .wav files to speech samples. Generally, a speech signal consists of noise-speech-noise in a noisy environment. The recognition of actual speech in the given samples is important. We divided the speech signal into frames of 450 samples each with an overlap of 300 samples, i.e., two-thirds of a frame length. The speech is alienated from the pauses using voice activity detection (VAD) techniques, which are discussed in detail later in the paper. The system performs speech analysis and synthesis using the linear predictive coding (LPC) method. From the LPC coefficients we get the weighted cepstral coefficients and cepstral time derivatives, which form the characteristic vector for a frame. Then, the system performs vector quantization using a vector codebook. The resulting vectors form the observation sequence. For each word in the vocabulary, the system builds an HMM model and trains the model during the training phase. The training steps, from VAD to HMM model building, are performed using PC-based C programs. We load the resulting HMM models onto an FPGA for the recognition phase. In the recognition phase, the speech is acquired vigorously from the microphone through a codec and is stored in the FPGA's memory. These speech samples are preprocessed, and the probability of getting the observation sequence for each model is calculated. The uttered word is recognized based on maximum likelihood estimation.

VI. APPLICATIONS OF SPEECH TO TEXT SYSTEM:

The application field of STT is expanding fast whilst the quality of STT systems is also increasing steadily. Speech synthesis systems are also becoming more affordable for common customers, which makes these systems more suitable for everyday use & becomes a cot effective. Some uses of STT are described below[5].

1.Aid to Vocally Handicapped

A hand-held, battery-powered synthetic speech aid can be used by vocally handicapped person to express their words. The device will have especially designed keyboard, which accepts the input, and converts into the required speech within blink of eyes.

2.Source of Learning for Visually Impaired

Most important fact for listening is an important skill for people who are blind. Blind individuals rely on their ability to hear or listen to gain information quickly and efficiently. Students use their sense of hearing to gain information from books on tape or CD, but also to assess what is happening around them.

3.Games and Education

Synthesized speech can also be used in many educational institutions in field of study as well as sports. If the teacher can be tired at a point of time but a computer with speech synthesizer can teach whole day with same performance, efficiency and accuracy.

4.Telecommunication and Multimedia

STT systems make it possible to access vocal information over the telephone. Queries to such information retrieval systems could be put through the user's voice (with the help of a speech recognizer), or through the telephone keyboard. Synthesized speech may also be used to speak out short text messages in mobile phones.

5.Man-Machine Communication

Speech synthesis can be used in several kinds of human machine interactions and interfaces. For example, in warning, alarm systems, clocks and washing machines synthesized speech may be used to give more exact information of the current situation [5]. Speech signals are far better than that of warning lights or buzzers as it enables to react to the signal more fast if the person is unable to get light due some obstacles.

6.Voice Enabled E-mail

Voice-enabled e-mail uses voice recognition and speech synthesis technologies to enable users to access their email from any telephone. The subscriber dials a phone number to access a voice portal, then, to collect their email messages, they press a couple of keys and, perhaps, say a phrase like "Get my e-mail." Speech synthesis software converts e-mail text to a voice message, which is played back over the phone. Voice-enabled e-mail is especially useful for mobile workers, because it makes it possible for them to access their messages easily from virtually anywhere (as long as they can get to a phone),without having to invest in expensive equipment such as laptop computers or personal digital assistants.

VII. CONCLUSION:

In this paper, we discussed the topics relevant to the development of STT systems .The speech to text conversion may seem effective and efficient to its users if it produces natural speech and by making several modifications to it. This system is useful for deaf and dumb people to Interact with the other peoples from society. Speech to Text synthesis is a critical research and application area in the field of multimedia interfaces. In this paper gathers important references to literature related to the endogenous variations of the speech signal and their importance in automatic speech recognition. A database has been created from the various domain words and syllables. The desired speech is produced by the Concatenative speech synthesis approach. Speech synthesis is advantageous for people who are visually handicapped. This paper made a clear and simple overview of working of speech to text system (STT) in step by step process. The system gives the input data from mice in the form of voice, then preprocessed that data & converted into text format displayed on PC. The user types the input string and the system reads it from the database or data store where the words, phones, diaphones, triphone are stored. In this paper, we presented the development of existing STT system by adding spellchecker module to it for different language. There are many speech to text systems (STT) available in the market and also much improvisation is going on in the research area to make the speech more effective, and the natural with stress and the emotions.

VIII. ACKNOWLEDGMENT:

The authors would like to thanks the Director/Principal Dr. Vinod Chowdhary,Prof.Aade K.U. and Prof.Bhope V.P. Savitribai Phule, Pune University, for their useful discussions and suggestions during the preparation of this technical paper.

IX. REFERENCES:

[1] Sanjib Das, *"Speech Recognition Technique: A Review"*,International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012.

[2] Ms. Sneha K. Upadhyay,Mr. Vijay N. Chavda,*"Intelligent system based on speech recognition with capability of self learning"* ,International Journal For Technological Research In Engineering ISSN (Online): 2347 - 4718 Volume 1, Issue 9, May-2014.

[3] Deepa V.Jose, Alfateh Mustafa, Sharan R.*"A Novel Model for Speech to Text Conversion"* International Refereed Journal of Engineering and Science (IRJES)ISSN (Online) 2319-183X, Volume 3, Issue 1 (January 2014).

[4] B. Raghavendhar Reddy,E. Mahender,*"Speech to Text Conversion using Android Platform"*, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,Vol. 3, Issue 1, January -February 2013.

[5] Kaveri Kamble, Ramesh Kagalkar*," A Review: Translation of Text to Speech Conversion for Hindi Language"*, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.Volume 3 Issue 11, November 2014.

[6] Santosh K.Gaikwad,Bharti W.Gawali,Pravin Yannawar, *"A Review on Speech Recognition Technique"*,International Journal of Computer Applications (0975 – 8887)Volume 10– No.3, November 2010.

[7] Penagarikano, M.; Bordel, G., *"Speech-to-text translation by a non-word lexical unit based system,"*Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on , vol.1, no., pp.111,114 vol.1, 1999.

[8]. Olabe, J. C.; Santos, A.; Martinez, R.; Munoz, E.; Martinez, M.; Quilis, A.; Bernstein, J., *"Real time text-to-speech conversion system for spanish,"* Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84. , vol.9, no., pp.85,87, Mar 1984.

[9]Kavaler, R. et al., *"A Dynamic Time Warp Integrated Circuit for a 1000-Word Recognition System",* IEEE Journal of Solid-State Circuits, vol SC-22, NO 1, February 1987, pp 3-14.

[10] Aggarwal, R. K. and Dave, M., *"Acoustic modelling problem for automatic speech recognition system: advances and refinements (Part II)",* International Journal of Speech Technology (2011) 14:309–320.

[11] Ostendorf, M., Digalakis, V., & Kimball, O. A. (1996). *"From HMM's to segment models: a unified view of stochastic modeling for speech recognition".* IEEE Transactions on Speech and Audio Processing, 4(5), 360–378.

[12] Yasuhisa Fujii, Y., Yamamoto, K., Nakagawa, S., "AUTOMATIC SPEECH RECOGNITION USING HIDDEN CONDITIONAL NEURAL FIELDS", ICASSP 2011: P-5036-5039.

[13] Mohamed, A. R., Dahl, G. E., and Hinton, G., *"Acoustic Modelling using Deep Belief Networks",* submitted to IEEE TRANS. On audio, speech, and language processing, 2010.

[14] Sorensen, J., and Allauzen, C., *"Unary data structures for Language Models"*, INTERSPEECH 2011.

[15] Kain, A., Hosom, J. P., Ferguson, S. H., Bush, B., *"Creating a speech corpus with semi-spontaneous, parallel conversational and clear speech",* Tech Report: CSLU-11-003, August 2011.