IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018), PP 36-43
www.iosrjournals.org

Speech to text and text to speech recognition systems-Areview

Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal
DepartmentofComputerScience, NMIMS University,Mumbai,India.
Corresponding Author:Navya Pant

Abstract*: In present industry, communication is the key element to progress. Passing on information, to the right person, and in the right manner is very important, not just on a corporate level, but also on a personal level. The world is moving towards digitization, so are the means of communication. Phone calls, emails, text messages etc. have become an integral part of message conveyance in this tech-savvy world. In order to serve the purpose of effective communication between two parties without hindrances, many applications have come to picture, which acts as a mediator and help in effectively carrying messages in form of text, or speech signals over miles of networks. Most of these applications find the use of functions such as articulatory and acoustic-based speech recognition, conversion from speech signals to text, and from text to synthetic speech signals, language translation amongst various others. In this review paper, we'll be observing different techniques and algorithms that are applied to achieve the mentioned functionalities.*

Keywords:*Speech to Text, Text to speech, Speech recognition, communication, Machine translation*

Date of Submission: 02-03-2018 Date of acceptance: 17-03-2018

I. Introduction

Over the past few years, Cell Phones have become an indispensable source of communication for the modern society. We can make calls and text messages from a source to a destination easily. It is known that verbal communication is the most appropriate modem of passing on and conceiving the correct information, avoiding misquotations. To fulfil the gap over a long distance, verbal communication can take place easily on phone calls. A path-breaking innovation has recently come to play in the SMS technology using the speech recognition technology, where voice messages are being converted to text messages. Quite a few applications used to assist the disabled make use of TTS, STT, and translation. They can also be used for other applications, taking an example: Siri an intelligent automated assistant implemented on an electronic device, to facilitate user interaction with a device, and to help the user more effectively engage with local and/or remote services [1] makes use of Nuance Communications voice recognition and text-to-speech (TTS) technology. In this paper, we will take a look at the different types of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation. Under speech the recognition we will follow the method i.e. pre-emphasis of signals, feature extraction and recognition of the signals which help us in training and testing mechanism. There are various models used for this purpose but Dynamic time warp, which is used for feature extraction and distance measurement between features of signals and Hidden Markov Model which is a stochastic model and isused to connect various states of transition with each other is majorly used. Similarly for conversion of speech to text we use DTW and HMM models, along with various Neural Network models since they work well with phoneme classification, isolated word recognition, and speaker adaptation. End to end ASR is also being tested as of late 2014 to achieve similar results. Speech synthesis works well in helping convert tokenized words to artificial human speech. Different machine translation methods, as well as engines will also be reviewed and compared in this paper. Following are the components of speech production, which are looked up to while applications use different speech related functionalities[15].

- Phonation (producing sound)
- Fluency
- Intonation
- Pitch variance
- Voice (including aeromechanical components of respiration)

II. Literature Review

In this review paper we have analysed the existing system for speech recognition, speech to text conversion, speech to text conversion and machine leaning methods.

A. SPEECH RECOGNITION

Speech Recognition is the ability of machine/program to identify words and phrases in spoken language and convert them into machine-readable format. Speech Recognition Systems can be classified on basis of the following parameters [1]:

• **Speaker:** All speakers have a different kind of voice. The models hence are either designed for a specific speaker or an independent speaker.

• **Vocal Sound:** The way the speaker speaks also plays a role in speech recognition. Some models can recognize either single utterances or separate utterance with a pause in between.

• **Vocabulary:** The size of the vocabulary plays an important role in determining the complexity, performance, and precision of the system.

1) Basic Speech Recognition Model:

Each speech recognition system follow some standard steps as shown in figure 1 [1].

i) Pre-processing: The analog speech signal is transformed into digital signals for later processing. This digital signal is moved to the first order filters tospectrally flatten the signals. This helps in increasing the signal's energy at a higher frequency.

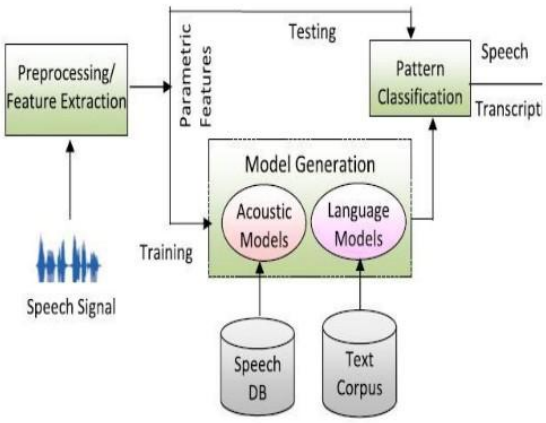

Fig.1. ArchitectureforSpeechRecognitionSystem

ii) Feature Extraction: This step finds the set of parameters of utterances that have a correlation with speech signals. These parameters, known as features, are computed through processing of the acoustic waveform. The main focus is to compute a sequence of feature vectors (relevant information) providing a compact representation of the given input signal. Commonly used feature extraction techniques are discussed below:

• **Linear Predictive Coding (LPC):** The basic idea is that the speech sample can be approximated as a linear combination of past speech samples. Figure 2 shows the LPC process [2]. The digitized signal is blocked into frames of N samples. Then each sample frame is windowed to minimize signal discontinuities. Each framed window is then auto- correlated. The last step is the LPC analysis, which converts each frame of autocorrelations into LPC parameter set.

• **Mel-Frequency Cestrum Co-efficient (MFCC):** It is a very powerful technique and uses human auditory perception system. MFCC applies certain steps to the input signal: Framing: Speech wave- form is cropped to remove interference if present; Windowing: minimizes the discontinuities in the signal; Discrete Fourier Transform: converts each frame from time domain to frequency domain; Mel Filter Bank Algorithm: the signal is plotted against the Mel spectrum to mimic human hearing [2].

• **Dynamic Time Warping:** This algorithm is used for measuring the similarity between two-time series which may vary in speed, based on dynamic programming. It aims at aligning two sequences of feature vectors (1 of each series) iteratively until an optimal match (according to a suitable metrics) between them is found.

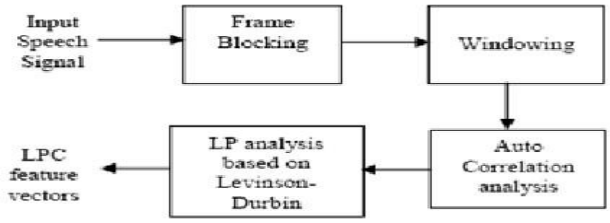

Fig.2.LPCFeatureExtractionProcess

iii) Acoustic Models: It is the fundamental part of Automated Speech Recognition (ASR) system where a connection between the acoustic information and phonetics is established. Training establishes a correlation between the basic speech units and the acoustic observations.

iv) Language Models: This model induces the probability of a word occurrence after a word sequence. It contains the structural constraints available in the language to generate the probabilities of occurrence. The language model distinguishes word and phrase that has a similar sound.

v) Pattern Classification: It is the process of comparing the unknown pattern with existing sound reference pattern and computing similarity between them. After completing the training of the system at the time of testing, patterns are classified to recognize the speech. Different approaches for pattern matching are [1]:

• **Template Based Approach**: This approach has a collection of speech patterns which are stored as a reference representing dictionary words. Speech is recognized by matching the uttered word with the reference template [14].

• **Knowledge Based Approach**: This approach takes set of features from the speech and then train the system to generate set of production rules automatically from the samples.
• **Neural Network Based Approach**: This approach is capable of solving more complicated recognition task. The basic idea is to compile and in- corporate knowledge from a variety of knowledge sources with the problem at hand [8].

• **Statistical Based Approach**: In this approach, variations in speech are modelled statistically (e.g. HMM) using training methods.

2) Speech to Text Conversion Methods:

Speech to text conversion is the process of converting spoken words into written texts. It is synonymous to speech recognition but the latter is used describe the wider process of speech understanding. STT follows the same principles and steps of speech recognition, with different combinations of various techniques for each step. Some widely used conversion methods are discussed below.

i) Hidden Markov Model (HMM): HMM is a statistical model used in speech recognition because a speech signal can be viewed as a piece wise stationary signal or a short-time stationary signal. HMM, models are useful for real-time speech to text conversion for mobile users[10]. It depends on the following parameters:

• **Recognition accuracy**- Recognition is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between the test pattern and each reference pattern. It is the most important factor of any recognition system, ideally it should be 100 % and independent of the speaker.

• **Recognition speed** - If the system takes a long time to recognize the speech, users would become restless and the system loses its significance. The signals *undergoes* the following steps: [3]

• Pre-processing: The input speech signals are *converted* into speech frames and give a unique sample, reducing noise.
• HMM Training: Training involves creating a pat- tern representative of the features of a class using one or more test patterns that correspond to speech sounds of the same class.
• HMM Recognition: It is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity (distance). Maximum likelihood is used for recognition.

ii) Artificial Neural Network Classier(ANN) based Cuckoo Search Optimization: ASR with Cuckoo Search Optimization technique is used for better communication, better recognition and to remove unwanted noise. ASR is built for a better interface of human and machine interaction. For the same, a three-step process is followed: [4]

• Pre-processing of the speech signals is the most important part of speech recognition which is executed to remove avoidable waveform of the signal. The signals are fed to the high-pass filters to remove the background noises.
• Two kinds of acoustic features are extracted, from the speech signal. They are Mel Frequency Cep- strum Coefficients (MFCC) and Linear Predictive Coding coefficients (LPCC).
• Classification: In this, artificial neural network is used as the classifier. The neural network is a three-layered classifier with n input nodes, l hidden nodes and k output nodes. In CSO (Cuckoo Search Optimization), ANN is implemented by two-layered Feed Forward Backpropagation Neural Network (FFBNN) with 3 units; two input unit, three Hidden units and one output unit. Here, the input layer consists of two inputs having two feature extracted which are MFCC and LPCC features. These features are given as input in which networks get trained and it produces a corresponding output.

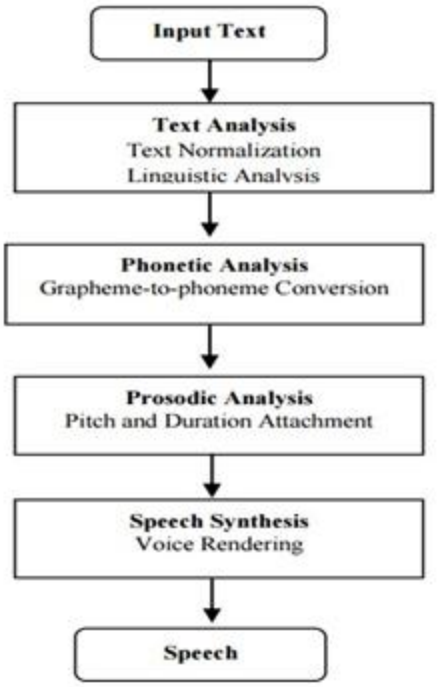

Fig.3. Texttospeechsystemflow

B. TEXT TO SPEECH CONVERSION

Text-To-Speech is a process in which input text is first analysed, then processed and understood, and then the text is converted to digital audio and then spoken. Figure 3 shows the block diagram of TTS. The figure shows all the steps involved in the text to speech conversion but the main phases of TTS systems are [5]:

• **Text Processing**: The input text is analysed, normalized (handles acronyms and abbreviation and match the text) and transcribed into phonetic or linguistic representation.
• **Speech Synthesis:** Some of the speech synthesis techniques are [5]:

i) Articulator Synthesis: Uses mechanical and acoustic model for speech generation. It produces intelligible synthetic speech but it is far from natural sound and hence not widely used.

ii) Formant Synthesis: In this system, representation of individual speech segments are stored on a parametric basis. There are two basic structures in formant synthesis, parallel and cascade, but for better performance, some kind of combination of these 2 structures is used. A cascade formant synthesizer consists of band-pass resonators connected in series. The output of each formant resonator is applied to the input of the successive one. The cascade structure needs only formant frequencies as control information. A parallel formant synthesizer consists

of resonators connected in parallel. The excitation signal is applied to all formants simultaneously and their outputs are summed. [5]

iii) Concatenative Synthesis: This technique synthesizes sound by concatenating short samples of sound called units. It is used in speech synthesis to generate user specific sequence of sound from a database built from the recording of other sequences. Units for Concatenative synthesis are [5]: Phone- a single unit of sound; Diphone- is defined as the signal from either midpoint of a phone or point of least change within the phone to the similar point in the next phone; Triphone- is a section of the signal taking in a sequence going from middle of a phone completely through the next one to the middle of a third.

C. LANGUAGE TRANSLATION

In India, we have a variety of languages spoken. The 2001 Census recorded 30 languages which were spoken by more than a million native speakers and 122 which were spoken by more than 10,000 people, which is why it is very necessary to have applications and processes that can convert text from one language to another, keeping the sanctity of the message. Machine Translation (MT) is a field of Artificial Intelligence and Natural Language Processing which deals with translation from one language to another using machine translation system. [6]. The human translation process may be described as: Decoding the meaning of the source text, and Re-encoding this meaning in the target language. Some of the machine translation models are discussed below:

i) Rule Based Machine Translation (RBMT): Translation is generated on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages. Such a system consist of collection of rules: Grammar rules- basically consist of analysis of languages in terms of grammar structures (syntax, semantic, morphology, part of speech tagging and orthographic features); bilingual or multilingual lexicon dictionary for looking up words during translation while the software program allows effective and efficient interaction of components; and software programs to understand a process those rules. There are three types of rule-based model:

• Direct: It is dictionary based.
• Transfer: It uses lexicons and structural analysis into every SL input text after which it's converted to intermediate representation.
• Interlingual: source language is transformed into an intermediate language which is independent of any of the languages involved in the translation.

ii) Statistical machine translation (SMT): It is characterized by the use of machine learning methods. SMT is a data-driven approach which uses parallel aligned corpora and treats translation as a mathematical reasoning problem. In that, every sentence in the target language is a translation with probability from the source language. The higher the probability, the higher is the accuracy of translation and vice-versa. Basic SMT architecture includes:

• Language model for calculating the probability of the target language
• Translation model for calculating conditional probability of target language output given source language input
• Decoder model- gives the best translation possible t by maximize the two probability mentioned above.

iii) Example based machine translation (EBMT): It is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components. The similar sentences are then used to translate the sub-sentential components of the original sentence into the target language, and these phrases are put together to form a complete translation. The Analogy translation uses three stages; matching, adaption and recombination

• **Matching**-The SL input text is fragmented, followed by search for examples from database which closely matches the input SL fragment string and the relevant fragments are picked. The TL fragments corresponding to the relevant fragments are extracted.

• **Adaption**-If the match is exact, the fragments are recombined to form TL output, else find the TL portion of the relevant match correspond to specific portion in SL and align them.
• **Recombination-** Combination of relevant TL fragments in order to form legal grammatical target text.

iv)Hybrid machine translation: The expansion of methodologies in the past decade and the introduction of new applications for automated translation have highlighted the limitations of adopting one single approach to the problems of translation. [7] Hybrid MT is a method of machine translation that is characterized by the use of

multiple machine translation approaches within a single machine translation system. It is a combination of RMBT and SMT method, and it makes the use of the advantages of both these methods. Statistical data is hence, put to use in generation of lexicon and syntax.
Engines like IBM Watson Developer Cloud, Google Translate, and Microsoft translators are widely used by various applications or work independently in helping effectively translated languages for better understanding.

III. Observations

MODELS	TECHNIQUES	FINDINGS	ISSUES
SPEECH RECOGNITION: FEATURE EXTRACTION	Linear Predictive Coding (LPC)	Static feature extraction method. Spectral analysis is done with a fixed resolution along a subjective frequency scale. [1]	Frequencies are weighted equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic
	Mel-Frequency Cestrum Co-efficient (MFCC)	It is the nearest feature extraction method to the actual human auditory speech perception.	MFCC values are not very robust in the presence of additive noises. Normalization is required [1]
	Dynamic Time Warping (DTW)	It is used to cope with different speaking speed. Simple hardware implementation.	Difficulty in selecting the reference template.
SPEECH RECOGNITION: PATTERN MATCHING	Template Based	Simple Approach Errors due to segmentation or classification of smaller acoustically more variable units is avoided. It is speaker dependent.	The pre-recorded templates are fixed. Template training and matching become impractical as vocabulary size increases. Continuous speech recognition is not possible.
	Knowledge Based	Uses the information regarding linguistic, phonetic and spectrogram. [1]	Explicit modelling variation in speech is difficult to obtain and use successfully, so, this approach is impractical.
	Neural Based	Solve complicated recognition task. Reduces modelling unit. Can be used to develop hybrid models	
	Statistical based	Present models use this approach	Low accuracy of priori modelling presumption reducing its trend
	Hidden Markov Model (HMM)	HMMs are simple, automatically trained and computationally feasible to use.	Lack in discrimination property for classification
SPEECH TO TEXT CONVERSION	Artificial Neural Network based Cuckoo Search Optimization	Simple Fast convergence rate Increase the recognition accuracy of the speech recognition system. [4]	Not effective in modelling time-variability of speech

TEXT TO SPEECH CONVERSION	Articulator Synthesis	Use mechanical and acoustic model	Output is far from natural voice.
	Formant Synthesis	Based on the source filter-model of speech	The cascade structures has been found better for non-nasal voiced sounds and because it needs less control information than parallel structure, it is then simpler to implement. Combination of 2 can be used.
	Concatenative Synthesis	Duration of units is not fixed, can be varied as per implementation.	Complex Method
MACHINE TRANSLATION	Rule Based Machine Translation (RBMT)	Collection of Grammar rules, and grammar structure. Is of 3 types as discussed	It is hard to deal with rule interactions in big systems, ambiguity, and idiomatic expressions. Insufficient amount of really good dictionaries
	Statistical Machine Translation (SMT)	Generated on the basis of statistical model, Probabilistic modelling. Makes use of Bayes theorem, pdf etc.	Can be costly, doesn't work well between languages with different word orders.
	Example Based Machine Translation (EBMT)	Bilingual corpus with parallel texts as its main knowledge.	Computational efficiency for large database is less.
	Hybrid Machine Translation (HMT)	Integration of advantages of rule based and SMT.	

IV. Conclusion

We have learned about various techniques that fall under STT and TTS, and have also read about their applications and usage. After having looked upon closely, at the different types of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation systems, we can draw a conclusion as such: In STT, under the 2 studied we can say that HMM works as a better speech signal to text converter despite its drawbacks because of their computational feasibility. Similarly under TTS systems studied formant synthesis that makes use of parallel and cascade synthesis works as the best converter. Hybrid machine translation is widely used due to its inculcation of advantages of both rule- based as well as statistical machine translation techniques. It makes sure that there is a creation of syntactically connected and grammatically correct text while also taking care of smoothness in a text, fast learning ability, data acquisition which are a part of SMT.

References

[1]. Suman K. Saksamudre, P.P. Shrishrimal, R.R. Deshmukh, A Review on Different Approaches for Speech Recognition System, International Journal of Computer Applications (0975 8887) Volume 115 No. 22, April 2015.

[2]. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, A Comparative Study of Feature Extraction Techniques for Speech Recognition System, International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 12, December 2014.

[3]. Ms. Anuja Jadhav, Prof. Arvind Patil, Real Time Speech to Text Con- verter for Mobile Users, National Conference on Innovative Paradigms in Engineering Technology (NCIPET-2012) Proceedings published by International Journal of Computer Applications (IJCA)

[4]. Sunanda Mendiratta, Dr. Neelam Turk, Dr. Dipali Bansal, Speech Recognition by Cuckoo Search Optimization based Artificial Neural Network Classifier, 2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015.

[5]. Suhas R. Mache, Manasi R. Baheti, C. Namrata Mahender, Review on Text-To-Speech Synthesizer, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.

[6]. Aditi Kalyani, Priti S. Sajja, A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies, Interna- tional Journal of Computer Applications (0975 8887) Volume 121 No.23, July 2015

[7]. Mouiad Fadiel Alawneh, Tengku Mohd Sembok Rule-Based and Example-Based Machine Translation from English to Arabic, 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications

[8]. F. Seide, G. Li, D. Yu,Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, In Interspeech, pp. 437440, 2011.

[9]. Kamini Malhotra,Anu Khosla, Automatic Identification of Gender Ac- cent in Spoken Hindi Utterances with Regional Indian Accents, 978-1- 4244-3472-5/08/25.00 2008 IEEE

[10]. y Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Speech Synthesis Based on Hidden Markov Models, Proceedings of the IEEE — Vol. 101, No. 5, May 2013. Junichi Yamagishi, Member IEEE, and Keiichiro Oura

[11]. G. E. Dahl, D. Yu, L. Deng, A. Acero, Large vocabulary continuous speech recognition with context-dependent DBN-HMMs, In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4688-4691, 2011.
[12]. Pere Pujol Marsal, Susagna Pol Font, Astrid Hagen, H. Bourlard, and C. Nadeu, Comparison And Combination Of Rasta-Plp And Ff Features In A Hybrid Hmm/Mlp Speech Recognition System, Speech and Audio Processing, IEEE Transactions on Vol.13, Issue: 1, 20 December 2004.
[13]. Tatsuhiko KINJO, Keiichi FUNAKI,"ON HMM SPEECH RECOG- NITION BASED ON COMPLEX SPEECH ANALYSIS",1-4244-0136- 4/06/20.00 '2006 IEEE
[14]. Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Template Based Continuous Speech Recognition,IEEE Transs. On Audio, Speech Language Processing, vol.15, issue 4,pp 1377-1390, May 2007.
[15]. Lawrence Rabiner, Biing-Hwang Juang, B.Yegnanarayana, Fundamentals of Speech Recognition.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

*Navya Pant "Speech to text and text to speech recognition systems-Areview." IOSR Journal of Computer Engineering (IOSR-JCE) 20.2 (2018): 36-43.