# A Review on Text-to-Speech System

Vaishnavi R. Ambaskar1, Vinod M. Lokhande, Dr. Avinash S. Kapse3

*Student, IT. Dept Anuradha Engineering College Maharashtra, India1,2*

*Head of Dept. Anuradha Engineering College Maharashtra, India3*

## ABSTRACT

*Speech is one amongst the oldest and most natural means that of data exchange between human. Over the years, tries are created to develop vocally interactive computers to understand voice/speech synthesis. Clearly such AN interface would yield nice advantages. During this case a pc will synthesize text and provides out a speech. Text-To-Speech Synthesis may be a Technology that gives a way of changing written communication from a descriptive type to a speech communication that's simply comprehensible by the top user (Basically in English Language). It runs on JAVA platform, and also the methodology used was Object orientating Analysis and Development Methodology; whereas knowledgeable System was incorporated for the interior operations of the program. This style is going to be double-geared towards providing a unidirectional communication interface whereby the pc communicates with the user by reading out matter document for the aim of fast assimilation and reading development. These days, communication is that the key part to progress. Passing on information, to the right person, and inside the correct manner is implausibly very important, not merely on a corporation level, but to boot on a personal level. The world is moving towards conversion, so unit the suggests that of communication. Phone calls, emails, text messages etc. became Associate in Nursing integral an area of message conveyance throughout this tech-savvy world. Thus, on serve the aim of effective communication between two parties whereas not hindrances, many applications have come to image, that acts as a negotiator and facilitate in effectively carrying messages fashionable of text, or speech signals over miles of networks. Most of these applications understand the use of functions like pronunciation and acoustic-based speech recognition, conversion from speech signals to text, and from text to artificial speech signals, language translation amongst varied others.*

**Keywords:-** *Speech-to-Text, TTS, Speech Recognition*

## 1. INTRODUCTION

Over the past few years, Cell, Phones became a crucial provide of communication for the trendy society. We are going to build calls and text messages from a provided to a destination merely. It's known that verbal communication is that the foremost applicable equipment of passing on and conceiving the right data, avoiding misquotations. To fulfil the gap over an extended distance, verbal communication can surface merely on phone calls. A path-breaking innovation has recently come to play at intervals the SMS technology mistreatment the speech recognition technology, where voice messages unit being regenerated to text messages. Quite a few applications accustomed assist the disabled build use of TTS, STT, and translation. They're going to even be used for different applications, example: Siri academic degree intelligent automatic assistant implemented on academic degree device, to facilitate user interaction with a tool, and to help the user plenty of effectively engage with native and/or remote services makes use of which means Communications voice recognition and text-to-speech (TTS) technology. The text-to-speech (TTS) synthesis procedure consists of 2 main phases. The primary is text analysis, wherever the input text is transcribed into a phonetic or another linguistic illustration, and also the second is that the generation of speech waveforms, wherever the output is created from this phonetic and speech data. These 2 phases are typically known as high and low-level synthesis. A text to speech convertor convert's traditional language text into speech. Synthesized speech may be created by concatenating items of recorded speech that is keep in an exceeding info. Systems dissent within the size of the keep speech units; a system that stores phones or iPhones provides the biggest output vary, however might lack clarity. For specific usage domains, the storage of entire words or sentences permits for high-quality output. Instead, a synthesizer will incorporate a model of the vocal tract and alternative human voice characteristics to form a very "synthetic" voice output. Throughout this paper, we have a tendency to are progressing to take a look at the assorted types of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation. Beneath speech the recognition we have a tendency to AR progressing to follow the strategy i.e., pre-emphasis of signals, feature extraction and recognition of the signals that facilitate U.S.A. in coaching job and testing mechanism. There unit varied models used for this purpose but Dynamic time warp, that's utilized for feature extraction and distance measurement between choices of signals and Hidden Andre Mark off Model

that will be a random model and issued to connect varied states of transition with each other is majorly used. Equally, for conversion of speech to text we have a tendency to tend to use DTW and HMM models, at the facet of assorted Neural Network models since they work well with sound classification, isolated word recognition, and speaker adaptation. End to end ASR is to boot being tested as of late 2014 to understand similar results. Speech synthesis works well in serving to convert tokenized words to artificial human speech. Fully totally different AI ways that, still as engines are going to be reviewed and compared throughout this paper. Following unit, the weather of auditory communication, that unit searched to whereas applications use fully totally different speech connected functionalities [15

- Phonation (producing sound)
- Fluency
- Intonation
- Pitch Variance
- Voice (including aeromechanical parts of respiration)

## 2. LITERATURE REVIEW

In this review paper we've analysed the present system for speech recognition, text to speech conversion, speech to text conversion and machine leaning strategies.

### 2.1 Speech Recognition

Speech Recognition is that the ability of machine/program to spot words and phrases in oral communication and convert them into machine-readable format. Speech Recognition Systems may be classified on basis of the subsequent parameters [:

1. **Speaker:** All speakers have a unique quite voice. The models therefore are either designed for a selected speaker or AN freelance speaker.
2. **Vocal Sound:** The method the speaker speaks conjointly plays a job in speech recognition. Some models will acknowledge either single vocalizations or separate utterance with a disruption in between.
3. **Vocabulary:** the scale of the vocabulary plays a crucial role in determinant the complexes, performance, and preciseness of the system.

### 2.2 Basic Speech Recognition Model:

Every speech recognition system follows some commonplace steps: [1].

1. **Pre-processing:** The analog speech signal is remodelled into digital signals for later process. This digital signal is stirred to the primary order filters spectrally flatten the signals. This helps in increasing the signal's energy at the next frequency.
2. **Feature Extraction:** This step finds the set of parameters of utterances that have a correlation with speech signals. These parameters, called options, are computed through process of the acoustic undulation. The most focus is to figure a sequence of feature vectors (relevant information) providing a compact illustration of the given input. Usually used feature extraction techniques are mentioned below:

    A. **Linear prophetic writing (LPC):** the essential plan is that the speech sample may be approximated as a linear combination of past speech samples. Figure a pair of shows the LPC method [2]. The digitized signal is blocked into frames of N samples. Then every sample frame is windowed to reduce signal discontinuities. Every framed window is then auto-correlative. The last step is that the LPC analysis, that converts every frame of autocorrelations into LPC parameter set.

    B. **Mel-Frequency asterid dicot genus Co-efficient (MFCC):** it's an awfully powerful technique and uses human sound perception system. MFCC applies bound steps to the input signal: Framing: Speech wave-kind is cropped to get rid of interference if present; Windowing: minimizes the discontinuities within the signal; separate Fourier Transform: converts every frame from time domain to frequency domain; Mel Filter Bank Algorithm: the signal is premeditated against the Mel spectrum to mimic human hearing [2].

C. **Dynamic Time Warping:** This rule is employed for activity the similarity between cuckold series which can vary in speed, supported dynamic programming. It aims at positioning 2 sequences of feature vectors (1 of every series) iteratively till associate degree best match (according to an acceptable metrics) between them is found.

3. **Acoustic Models:** it's the elemental a part of machine-driven Speech Recognition (ASR) system wherever an affiliation between the acoustic data and acoustics is established. Coaching establishes a correlation between the essential speech units and also the acoustic observations.
4. **Language Models:** This model induces the chance of a word incidence when a word sequence. It contains the structural constraints offered within the language to come up with the chances of incidence. The language model distinguishes word and phrase that incorporates a similar sound.
5. **Pattern Classification:** it's the method of comparison the unknown pattern with existing sound reference pattern and computing similarity between them. When finishing the coaching of the system at the time of testing, patterns are classified to acknowledge the speech.

Totally different approaches for pattern matching [1]:

A. **Example based mostly Approach:** This approach incorporates an assortment of speech patterns that are keep as a reference representing wordbook words. Speech is recognized by matching the verbalized word with the reference example [14].
B. **Data based mostly Approach:** This approach takes set of options from the speech so train the system to come up with set of production rules mechanically from the samples.
C. **Neural Network based mostly Approach:** This approach is capable of determination additional difficult recognition task. The essential plan is to compile and in- company data from a spread of information sources with the matter at hand [8].
D. **Applied mathematics based mostly Approach:** during this approach, variations in speech are modelled statistically (e.g., HMM) mistreatment coaching strategies.

6. **Speech to Text Conversion Methods:** Speech to text conversion is that the method of changing spoken words into written texts. It's substitutable to speech recognition however the latter is employed describe the broader method of speech understanding. STT follows identical principles and steps of speech recognition, with totally different mixtures of varied techniques for every step. Some wide used conversion strategies are mentioned below.
7. **Hidden Andre Mark off Model (HMM):** HMM may be an applied mathematics model utilized in speech recognition as a result of a speech signal is viewed as a bit wise stationary signal or a short-time stationary signal. HMM, models are helpful for period of time speech to text conversion for mobile users [10]. It depends on the subsequent parameters:

A. **Recognition accuracy:** Recognition is that the method of comparison the unknown takes a look at pattern with every sound category reference pattern and computing a life of similarity between the take a look at pattern and every reference pattern. It's the foremost necessary issue of any recognition system, ideally it ought to be 100% and freelance of the speaker.
B. **Recognition speed:** If the system takes a protracted time to acknowledge the speech, users would become restless and also the system loses its significance. The signals undergo the subsequent steps [3]:
    i. **Pre-processing:** The input speech signals are born-again into speech frames and provides a singular sample, reducing noise.
        1. **HMM coaching:** Training involves making a pat-l arid representative of the options of {a category a category} mistreatment one or additional take a look at patterns that correspond to speech sounds of identical class.
C. **HMM Recognition:** it's the method of comparison the unknown takes a look at pattern with every sound category reference pattern and computing a life of similarity (distance). Most chances are used for recognition.

8. **Artificial Neural Network Classier (ANN) based mostly Cuckoo Search improvement**: ASR with Cuckoo Search Optimization technique is employed for higher communication, higher recognition and to get rid of unwanted noise. ASR is made for a more robust interface of human and machine interaction. For identical, a three-step method is followed: [4]

- Pre-processing of the speech signals is that the most vital a part of speech recognition that is dead to get rid of evitable undulation of the signal. The signals are fed to the high-pass filters to get rid of the background noises.
- 2 varieties of acoustic options are extracted, from the speech signal. They're Mel Frequency CEP-strum Coefficients (MFCC) and Linear prognostic committal to writing coefficients (LPCC).

9. **Classification:** during this, artificial neural network is employed because the classifier. The neural network may be a three-layered classifier with n input nodes, l hidden nodes and k output nodes. In CSO (Cuckoo Search Optimization), ANN is enforced by 2-layered Feed Forward Back Propagation Neural Network (FFBNN) with three units; two input unit, 3 Hidden units and one output unit. Here, the input layer consists of 2 inputs having 2 features extracted that are MFCC and LPCC options. These options are given as input during which networks get trained, and it produces a corresponding output.

### 2.3 Text to Speech Conversion

Text-To-Speech may be a method during which input text is initial analysed, then processed and understood, so the text is born-again to digital audio so spoken. Figure three shows the diagram of TTS. The figure shows all the steps concerned within the text to speech conversion however the most phases of TTS systems [5]:

- **Text Processing:** The input text is analysed, normalized (handles acronyms and abbreviation and match the text) and transcribed into phonetic or linguistic illustration.
- Speech Synthesis: a number of the speech synthesis techniques area unit [5]:
- **Articulatory Synthesis:** Uses mechanical and acoustic model for speech generation. It produces intelligible artificial speech however it's removed from natural sound and thus not wide used.
- Format Synthesis: during this system, illustration of individual speech segments area unit holds on a constant quantity basis. There are a unit 2 basic structures in format synthesis, parallel and cascade, except for higher performance, some quite combination of those two structures is employed. A cascade format synthesizer consists of band-pass resonators connected serial. The output of every format resonator is applied to the input of the sequent one. The cascade structure desires solely formant frequencies as management info. A parallel formant synthesizer consists of resonators connected in parallel. The excitation signal is applied to any or all formants at the same time and their outputs' area unit summed [5].
- Concatenative Synthesis: this system synthesizes sound by concatenating short samples of sound referred to as units. it's utilized in speech synthesis to come up with user specific sequence of sound from a info engineered from the recording of alternative sequences. Units for Concatenative synthesis are [5]: Phone- one unit of sound; Diphones outlined because the signal from either canter of a phone or purpose of least modification within the phone to the similar purpose within the next phone; Triphone- may be a section of the signal taking in an exceedingly sequence going from middle of a phone fully through successive one to the canter of a 3rd.

### 2.4 LANGUAGE TRANSLATION

In India, we've got a spread of languages spoken. The 2001 Census recorded thirty languages that were spoken by quite 1,000,000 native speakers and 122 that were spoken by quite ten,000 people, that is why it's terribly necessary to own applications and processes which will convert text from one language to a different, keeping the holiness of the message. computational linguistics (MT) may be a field of AI and linguistic communication process that deals with translation from one language to a different victimization computational linguistics system. [6]. The human translation method is also represented as: cryptography the which means of the supply text, and Reencoding this which means within the target language. a number of the computational linguistics models area unit mentioned below:

1. **Rule based mostly computational linguistics (RBMT):** Translation is generated on the premise of morphological, syntactic, and linguistics analysis of each the supply and therefore the target languages. Such a system encompasses assortment of rules: descriptive linguistics rules-primarily encompass

analysis of languages in terms of descriptive linguistics structures (syntax, semantic, morphology, a part of speech tagging and writing features); bilingual or multilingual lexicon wordbook for wanting up words throughout translation whereas the computer code program permits effective and economical interaction of components; and computer code programs to grasp a method those rules. There are a unit 3 forms of rule-based model:

- **Direct:** it's wordbook based mostly.
- **Transfer:** It uses lexicons and structural analysis into each FTO input text when that it's reborn to intermediate illustration.
- **Interlingual:** language is reworked into Associate in Nursing intermediate language that is freelance of any of the languages concerned within the translation.

2. **Applied mathematics computational linguistics (SMT):** it's characterised by the utilization of machine learning strategies. SMT may be a data-driven approach that uses parallel aligned corpora and treats translation as a mathematical reasoning downside. In that, each sentence within the target language may be a translation with likelihood from the language. The upper the likelihood, the upper is that the accuracy of translation and vice-versa. Basic SMT design includes:
   - Language model for shrewd the likelihood of the target language.
   - Translation model for shrewd {conditional like contingent probability | probability | chance} of target language output given language input.
   - Decoder model-offers the most effective translation potential t by maximize the 2 probabilities mentioned higher than.
3. Example based mostly computational linguistics (EBMT): it's supported the concept of analogy. During this approach, the corpus that's used is one that contains texts that have already been translated. Given a sentence that's to be translated, sentences from this corpus area unit selected that contain similar sub-sentential parts. The similar sentences area unit then accustomed translate the sub-sentential parts of the first sentence into the target language, and these phrases area unit place along to create an entire translation. The Analogy translation uses 3 stages; matching, adaption and recombination
   - **Matching-** The FTO input text is fragmented, followed by explore for examples from info that closely matches the input FTO fragment string and therefore the relevant fragments area unit picked. The metal fragments similar to the relevant fragments area unit extracted.
   - **Adaption-** If the match is precise, the fragments area unit recombined to create metal output, else notice the metal portion of the relevant match correspond to specific portion in FTO and align them.
   - **Recombination-** Combination of relevant metal fragments so as to create legal grammatical target text.

## 3. FINDINGS

**Model - A:** Speech Recognition Pattern Matching

Technique – I): Knowledge Based

Findings: Uses the information regarding linguistic, phonetic and spectrogram. [1]

Issues: Explicit modelling variation in speech is difficult to obtain and use successfully, so, this approach is impractical.

Technique – II): Template Based

Findings: Simple Approach Errors due to segmentation or classification of smaller acoustically more variable units are avoided. It is speaker dependent.

Issues: The pre-recorded templates are fixed. Template training and matching become impractical as vocabulary size increases. Continuous speech recognition is not possible.

Technique – III): Hidden Markov Model (HMM)

Findings: HMMs are simple, automatically trained and computationally feasible to use.

Issues: Lack in discrimination property for classification.

Technique – IV): Statistical based

Findings: Present models use this approach.

Issues: Low accuracy of priori modelling presumption reducing its trend.

Technique – V): Neural Based

Findings: Solve complicated recognition task. Reduces modelling unit. Can be used to develop hybrid models.

**Model B:** Machine Translation

Technique – I): Statistical Machine Translation (SMT)

Findings: Generated on the basis of statistical model, Probabilistic modelling. Makes use of Bayes theorem, pdf etc.

Issues: Can be costly, doesn't work well between languages with different word orders.

Technique – II): Rule Based Machine Translation (RBMT)

Findings: Collection of Grammar rules, and grammar structure is of 3 types as discussed.

Issues: It is hard to deal with rule interact in big systems, ambiguity, and idiomatic expressions. Insufficient amount of really good dictionaries.

Technique – III): Hybrid Machine Translation (HMT)

Findings: Integration of advantages of rule based and SMT.

Technique – IV): Example Based Machine Translation (EBMT)

Findings: Bilingual corpus with parallel texts as its crucial knowledge.

Issues: Computational efficiency for large database is less.

**Model C:** Speech-to-Text Conversion

Technique – I): Artificial Neural Network based Cuckoo Search Optimization

Findings: Simple Fast convergence rate Increase the recognition accuracy of the speech recognition system. [4]

Issues: Not effective in modelling time-variability of speech.

**Model D:** TEXT TO SPEECH CONVERSION

Technique – I): Formant Synthesis

Findings: It is based on the source filter model of speech.

Issues: The cascade structures have been found better for non-nasal voiced sounds and because it needs less control information than the parallel structure, it is then simpler to implement. Combination of 2 can be used.

Technique – II): Concatenative Synthesis

Findings: Duration of units are not fixed, can be varied as per implementation.

Issues: Complex Method.

Technique – III): Articulator Synthesis

Findings: Use mechanical and acoustic model.

Issues: Output is far from natural voice.

## 4CONLUSIN

We have learned concerning numerous techniques that comprise STT and TTS, and have conjointly examined their applications and usage. We've learned concerning numerous techniques that comprise STT and TTS, and have conjointly examined their applications and usage. When having looked upon closely, at the various sorts of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation systems, we are able to draw a conclusion as such: In STT, below the two studied we are able to say that HMM works as a much better speech signal to text convertor despite its drawbacks as a result of their process feasibleness. Equally, below TTS systems studied formant synthesis that produces use of parallel and cascade synthesis works because the best convertor. After having looked upon closely, at the various sorts of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation systems, we are able to draw a conclusion as such: In STT, below the two studied we are able to say that HMM works as a much better speech signal to text convertor despite its drawbacks as a result of their process feasibleness. Equally, below TTS systems studied formant synthesis that produces use of parallel and cascade synthesis works because the best convertor. Hybrid AI is wide used thanks to its instilling of benefits of each rule-primarily based similarly as applied math AI techniques. It makes positive that there's a creation of syntactically connected and grammatically correct text whereas conjointly taking care of smoothness in an exceeding text, quick wit, knowledge acquisition that square measure an area of SMT.

## 5. ACKNOWLEDGMENT

We would like to sincerely acknowledge the uncourageous efforts of Information Technology Department of Anuradha Engineering College, Chikhli. Our heartful thanks to faculty members who helped us for preparing this review paper and give the direction with their suggestions and rich experience.

## 6. REFERENCES

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.

[1]. 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.

[2]. 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)

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.

[3]. 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.

[4]. 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

[5]. 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

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

[6]. 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

[7]. 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

[8]. 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.

[9]. 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.

[10]. Tatsuhiko KINJO, Keiichi FUNAKI,"ON HMM SPEECH RECOG- NITION BASED ON COMPLEX SPEECH ANALYSIS",1- 4244-0136- 4/06/20.00 '2006 IEEE

[11]. 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.

[12]. Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik and Supriya Agrawal, Speech to text and text to speech recognition systems-Areview, 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.

[13]. Lawrence Rabiner, Biing-Hwang Juang, B.Yegnanarayana, Fundamentals of Speech Recognition.