TEXT TO SPEECH SYNTHESIS

By M.Vinitha

Abstract

Speech technology has been identified as one of the fastest growing engineering technologies. Nearly 20% of the worldwide population is suffering from various kinds of disabilities such as visual impairment, language-based learning disability and inability to use hands properly [1]. Text-to-speech (TTS) is a technology that converts the normal language input text into synthesised speech (output). Synthesised speech is created by concatenating together small segments of recorded human speech that are stored in database. The main aim is to develop an ideal speech synthesiser that synthesises natural and intelligible human speech [2].

TTS system is built for 13 major Indian languages, namely, Tamil, Telugu, Malayalam, Kannada, Hindi, Marathi, Bengali, Gujarati, Rajasthani, Assamese, Manipuri, Odia and Bodo. A unified framework is required for building TTS for Indian languages [3]. The TTS synthesisers for Indian languages are developed in a uniform manner by exploiting the similarities that exist among the languages. Cross-lingual analysis and borrowing of models will enhance the development of TTSes for new languages. Two successful techniques for speech synthesis are unit selection- based concatenative synthesis (USS) and hidden Markov model (HMM) based text to speech synthesis (HTS) also known as statistical parametric synthesis. The USS is based on waveform concatenation of pre-recorded speech units, while the latter is based on generation of optical parameter sequence from sub-word HMMs [4].

Degradation mean opinion scores (DMOS) and word error rate (WER) are used to evaluate the quality of the synthesised speech.

1. Introduction

Speech is the oldest and the most efficient means of communication between human beings. Text to speech synthesis also called speech synthesis is the artificial production of human synthesised speech. Speech synthesiser is a computer software used for this purpose [3]. A text-to-speech system takes text as input which is analysed, processed and then generated as synthesised speech. The challenging part of building the TTS system is finding ways to extract appropriate information from the input text to make the synthesised speech more intelligible and natural. The quality of a TTS system is analysed by its similarity to human speech (naturalness) and intelligibility. A TTS system is very helpful for people with visual impairments and language based learning disabilities to listen to written works.

In India, there are about 1652 languages with dialectal variations according to the 1961 Census report [5]. As a result, developing speech recognition, translation systems and synthesis is quite a challenging task. Indian languages can be broadly classified as Indo-Aryan, Dravidian or Sino-Tibetan language groups. Dravidian language group is predominant in the southern four states, Indo-Aryan group in the northern states and Sino-Tibetan group in the north-eastern states. Indo-Aryan languages has 700 million Indian speakers, Dravidian languages has 200 million Indian speakers and Sino-Tibetan language has 5.5 million Indian speakers which adds up to one-sixth the population of the world. Indian languages share a common sound base and this is studied to enable cross-lingual borrowing of models [3]. This approach reduces the turnaround time for building TTS system for a new language and without having any compromise on the synthesis quality.

-

Indian languages fall under the category syllable-timed languages. Moreover, syllables are considered to be the fundamental units of speech production. Syllables are defined as C*VC* units in which C denotes consonants and V denotes vowels. C* denotes that it can be zero or more consonants [6].

State- of- the- art high quality Unit Selection Speech Synthesiser (USS) for Indian languages are based on concatenation of actual speech units from the database for synthesising speech and built using syllable as the basic unit. Although the syllable-based USS synthesiser performs really well, the performance is found to be inconsistent in many cases and many flaws are observed. The output human synthesised speech lacks the continuity, flow and the rhythm of natural speech and due to the introduction of different artifacts suffers in terms of intelligibility as well. Artifacts are usually introduced in the speech waveform due to segmentation errors and errors at the sub-word unit level. As a result, statistical parametric speech synthesis system using hidden Markov models (HMM) has gained great popularity in the past few years. The difference is that they model sub-word speech units rather than concatenating actual speech waveforms. Models are built using context information and speech is synthesised by generating speech waveforms from these models. Context- dependent monophone is used as the basic sub-word unit by these systems [7]. An automatic speech segmentation algorithm is used to segment speech waveforms at the monophone level. The accuracy of segmentation plays a crucial role in the performance of these systems. Hence, various techniques to improve the quality of these systems and synthesised speech have been dealt.

2. Speech synthesis

2.1 What is speech synthesis?

A text- to- speech synthesiser (TTS) is a computer- based system that automatically converts the given input text to synthesised speech. TTS is illustrated in the Figure 2.1.

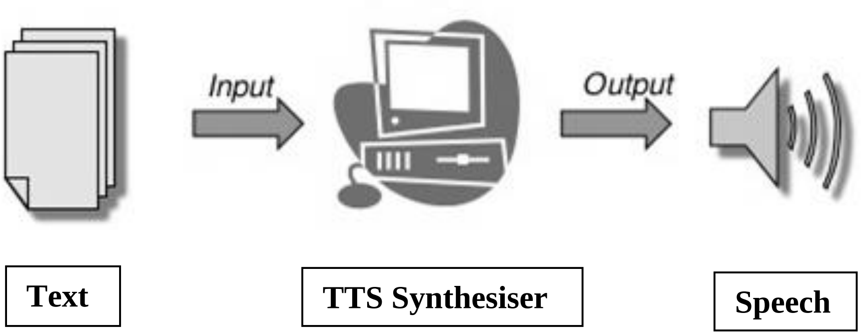


Figure 2.1: Text- to- speech synthesis

TTS systems are composed of two parts: 1) Front- end (text analysis part) and 2) Back-end (speech production part). TTS systems can be classified based on the method used to translate phonemes into audible sound. For example, pre-recorded TTS system where a database of pre-recorded words are maintained, formant TTS system where the voice is generated by the simulation of the behaviour of the human vocal cord and concatenated TTS system. The TTS system that has to be developed for this particular goal is not restricted to any specific domain and hence should be able to synthesise any arbitrary text input [8]. This type of generic domain TTS system can be used for a variety of applications such as telephony, automotive, multimedia, industrial and medical (to aid people with visual impairments and dyslexia).

The main challenge here is to derive maximum acoustic information and produce human-like speech which is natural and intelligible. TTS systems for Indian languages are really essential for people who are illiterate and have difficulties in reading their native language scripts. The similarities that exist among Indian languages make system building easier. This approach to system building is cross- lingual. TTS systems are trained by specific synthesis techniques.

2.2 Source-filter model for speech production

Speech production is the method by which a person can translate thoughts into speech. The difference between speech production and language production is that language can also be produced manually by signs. Conceptualisation, formulation and articulation are the three major processes involved in speech production. In order to synthesise human-like speech, the features and components that are used to represent speech should approximate the speech production system [9]. Lungs, larynx and vocal tract are the main parts involved in producing speech. The air coming from the lungs passes through the larynx and is modulated by the vocal tract in order to produce speech. The larynx consists of the vocal folds, also known as vocal cords. For voiced sounds, the vocal folds vibrate at a periodic interval. The vocal folds remain open for unvoiced sounds. The oral and naval cavities are parts of the vocal tract. The lower and upper lip, teeth, tongue, alveolar ridge, palate (both soft and hard) and glottis are articulators that help in the articulation process [7].

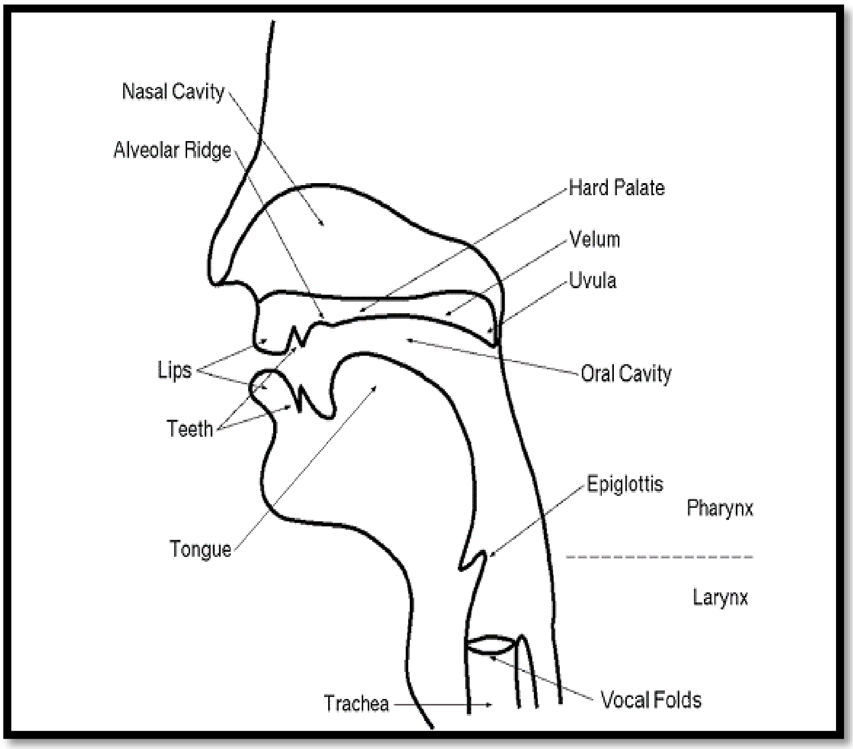
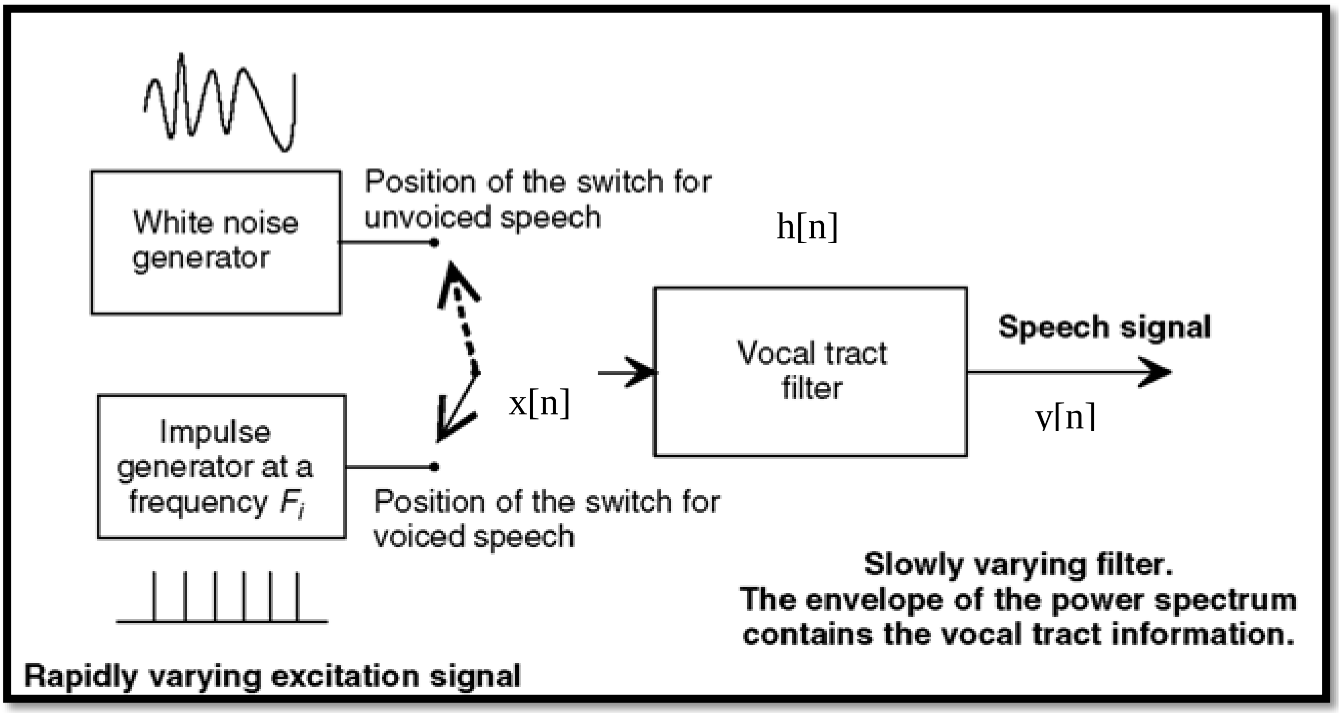


Fig 2.2: Various parts involved in speech production

The signal termed as source signal x[n] is produced from the vocal fold activity. The vocal tract configuration whose impulse response is h[n] acts like a linear filter (system). The result of the discrete convolution of the source and the filter is speech signal y[n].

$$\mathbf{y[n] = x[n] * h[n]}$$

The source-filter model shown in Figure 2.3 is the most important theory of speech production. The source and the filter are assumed to be independent of each other. Pitch values are used to represent the source information of speech. System parameters are represented by mel generalised cepstral (MGC) coefficients. By taking into account the behaviour of the human auditory system, which has higher resolution in the low frequency regions, the MGC coefficients model the spectrum of speech [7].



White noise generator
Position of the switch for unvoiced speech
h[n]
Impulse generator at a frequency F_i
x[n]
Vocal tract filter
Speech signal
y[n]
Position of the switch for voiced speech
Slowly varying filter.
The envelope of the power spectrum contains the vocal tract information.
Rapidly varying excitation signal

Fig 2.3: Source-filter model of speech production [10]

2.3 State-of-the-art techniques to build a TTS system

As shown in Figure 2.4 the training data consists of some recorded speech and the corresponding transcriptions or text. A large amount of text was collected from the internet which included news, sports, blogs and short stories in native languages. The training text is chosen very carefully to avoid long words and the sentences are mostly declarative [11]. The text for speech recording is selected such that all the aksharas and monophones in the language are covered, which are back-up units in the absence of syllables. Using language-specific letter-to-sound (LTS) or unified parser, the text is converted to sub-word sequence also known as labels. The labels may be phones, diphones, syllables, etc. This is followed by the process of segmentation where wavefiles are time-aligned or labelled according to the sub-word sequence. Later, the training data is used to build a TTS system. In synthesis phase, the test sentence is parsed into the sub-word sequence and the speech is synthesised. Unit selection synthesis (USS) and HMM based speech synthesis system (HTS) are the existing state-of-the-art techniques used to build TTS systems for Indian languages.

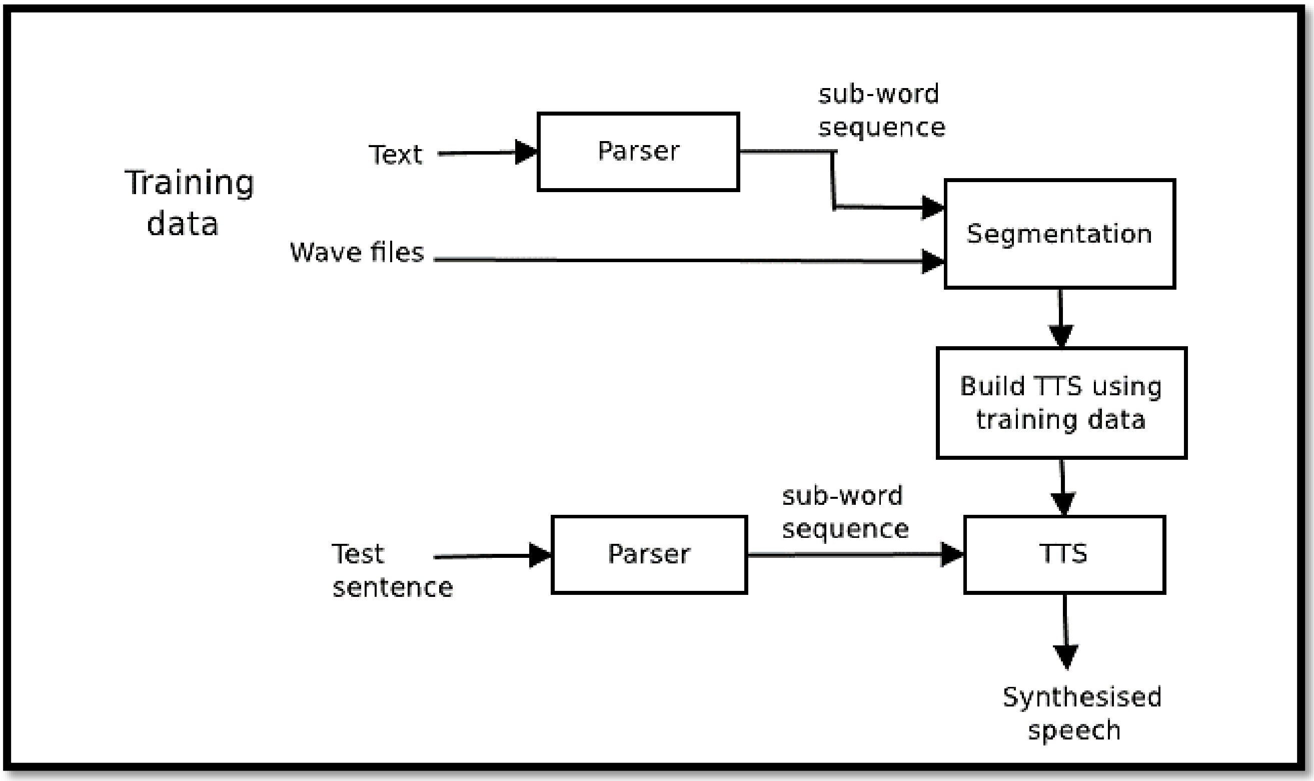


Fig 2.4: Procedure to build a TTS system [7]

2.3.1 Segmentation of speech waveforms

The process by which a waveform of continuous speech is segmented into the sub-word units it is composed of is called speech segmentation. In order to obtain high quality synthesis, the sub-word units have to be segmented very precisely [8]. The methods generally used for segmentation are flat start method, bootstrap method and hybrid segmentation technique. While the bootstrap method is semi-automatic, flat start method and hybrid segmentation techniques are fully automatic.

2.3.1.1 Flat start method

The entire duration of a wave file is divided equally among the phones that make up that utterance, in the flat start method. Then, monophone HMMs are built from the time-aligned data. The state means and variances are basically initialised with global mean and variance. Flat start initialised embedded training of monophone HMMs is iteratively re-estimated using embedded re-estimation and then followed by forced Viterbi algorithm [7].

2.3.1.2 Bootstrap method

In bootstrap segmentation, data is chosen with enough phone coverage and are manually segmented at the phone level. From the manually segmented data, HMMs are built and are used to segment the data using forced Viterbi algorithm at the phone level. Segmentation is the obtained from the re-estimated models that are iteratively re-estimated [7].

The bootstrap segmentation is used to time-align the speech data in the source language. Two sets of TTSes are built where the first one is built for target languages using flat start segmentation. The second TTS is built for different target languages by cross-lingual borrowing from the source languages (Tamil or Hindi depending on the language group) for annotation. By conducting experiments it is determined that the models in the source language built by bootstrap segmentation are definitely better models than the ones built in the target language using flat start segmentation.

2.3.1.3 Hybrid segmentation method

HMM based segmentation and group delay (GD) based segmentation are performed iteratively in the hybrid segmentation approach in order to obtain accurate segmentation automatically [12]. In hybrid segmentation algorithm, machine learning is used in tandem with signal processing [6]. The first set of TTSes is developed for the target languages using the hybrid segmentation algorithm. By using the CI-HMMs from the source language as initial models for segmentation, another set of TTSes is built for various target languages.

Using HMMs from a source language leads to a better synthesis quality in terms of intelligibility and naturalness. The main reason for this is the better use of initial models due to accurate phone-level in the segmentation. It is noticeable that hybrid segmentation gives better phone boundaries. On experimental basis, it can be said that the language specific and cross-lingual systems developed using hybrid segmentation outperforms both the language specific and cross-lingual systems developed using bootstrap segmentation [7].

2.3.2 Unit selection synthesis (USS)

Unit selection synthesis (USS) is the most simplest and successful technique for speech synthesis which selects and concatenates the pre-recorded speech units in the database to ensure that the cost criteria, which is a combination of target and concatenation costs are minimised. The corresponding waveforms are concatenated to produce the synthesised speech. To build a USS, the commonly chosen sub-word unit is called a phoneme. In the case of Indian languages, the sub-word unit chosen is a syllable. Syllables are of the form C*CV* and are usually composed of at least one vowel and may/may not have consonants preceding and/or succeeding the vowel.

The training phase is the phase in which the speech database is organised into structures that makes the process of retrieving the most suitable unit easy. The structures in this case which organise the speech database, is called classification and regression trees (CART). The text has to be broken down into its sub-word unit representation in order to build these CART structures. Acoustic, linguistic and phonetic features are used to build the CART structures [7].

Letter to sound rules, are a set of rules that breaks the given text into its corresponding sub-word unit representation. In this case, the letter to sound rules are a set of handwritten rules written for the two different types of language sub-categories, Aryan and Dravidian languages. The rules for both the sub-categories are almost the same, except for the deletion of schwa in Aryan languages [8].

Despite, the synthesis quality being natural, discontinuities are observed at the concatenation points. Moreover, the footprint size of the synthesiser is very large (about hundreds of MB depending on the training data) as the original speech waveforms are stored in the database. This makes it unsuitable for the process. Statistical parametric speech synthesis (SPSS) is another paradigm in speech synthesis which stores models of the sub-word units instead of the original waveforms in the database. Although, the synthesised speech is muffled due to averaging, the footprint size reduces to 6-10 MB. One famous SPSS technique is HTS also known as hidden Markov model (HMM) based speech synthesis system [7,8].

2.3.3 HMM based speech synthesis system (HTS)

In Figure 2.5, the training and synthesis phases of HTS are shown. They have much better intelligibility and smaller footprint compared to USS. A well labelled database is required for building a good quality HTS system for any language. In the training phase of the HTS, native script is converted to sequence of labels (using language specific parser), spectral parameters (Mel generalised cepstral coefficients) and their dynamic features, excitation parameters (log of the fundamental frequency f_0) and its dynamic features are extracted from the speech data [4]. The HMMs are four-stream with five states and a single mixture component per state. The first stream corresponds to the static and dynamic values of MGC. The next three streams is for the log f_0 features, velocity and acceleration values respectively [8]. Both voiced and unvoiced regions are modelled by the excitation features.

Three types of context are used by HTS: prosodic, linguistic and phonetic. For the HMM-based system, the phonetic context (the basic sub-word unit) used is the context-dependent pentaphone. Based on that, the UTF-8 text is converted to a sequence of pentaphones. With these contexts, the context dependent models are initialised with a set of context independent monophone HMM, as in the conventional HTS and stored in the database. Based on the question set, tree based context clustering is performed to tie states. This tree based context clustering is mainly done to address two problems associated with the use of context dependent phones and modelling. (1)Inadequate data to build all context dependent HMMs separately and (2) unseen models.

In the synthesis phase, to obtain the sentence HMM, the required context-dependent HMMs are concatenated and labels are generated from the sentence. By traversing the decision tree built during tree based context clustering, appropriate models are chosen. State durations are determined based on the duration HMMs. Using a source-filter model, the speech waveform is synthesised from the excitation parameters and generated spectral [13]. The Mel log spectral approximation (MLSA) filter is used, if cepstral coefficients are used as features [12].

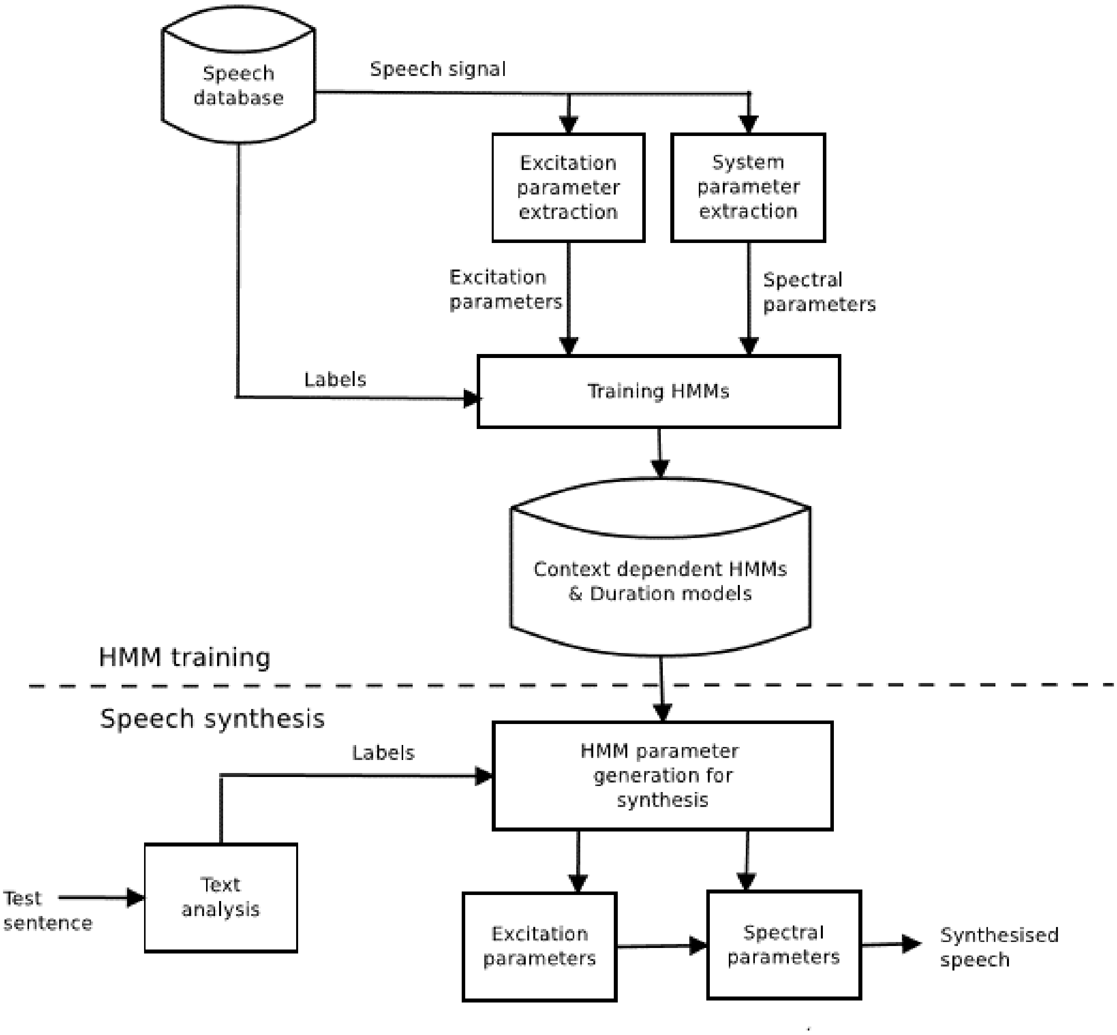


Fig 2.5: Overview of HMM based speech synthesis system (HTS) [8]

2.4 Unified framework

It is quite time consuming and expensive to build a TTS system for each Indian language from scratch. A common framework is developed for building Indian language TTSes called unified framework, given the increasing demand for faster development of TTS for multiple Indian languages and the lack of annotated data for each language. The main aim is to exploit the similarities that exist between languages to help in the system building. A natural design choice is to gradually build TTSes from only speech data and transcriptions by making the system modules as language independent as possible.

In the HTS framework, the quality of the synthesised speech depends on an accurate representation of the sounds in the language and also the question set for the clustering of the phones. The following modules are developed for the generic framework as part of the preliminary work:

1) Common label set (CLS)
2) Common question set (CQS)

2.4.1 Common label set (CLS)

The very first step to TTS system building is to list down the phones in that language. Scripts of the Indian languages are non-Latin and hence they require a separate representation in the Latin alphabets. The scripts of most of the Indian languages trace their origin to the Brahmi script. In the Brahmi script it is observed that the characters are arranges according to their place and manner of articulation [7]. Each language is observed to have 11-13 vowels. It is also observed that most of the consonants among different languages are phonetically similar. Except Tamil language which has only 26 consonants, all other languages have 33 consonants in common [5]. Dravidian languages also have long vowels for e and o in addition to the long vowels for a, i and u present in most Indo-Aryan languages.

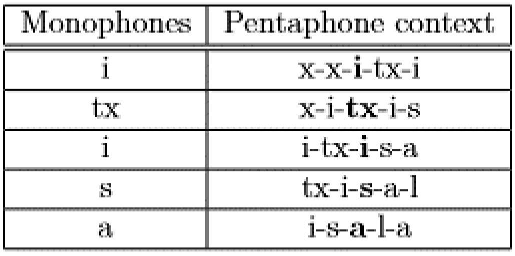
Language family	Indo-Aryan	Dravidian	Sino-Tibetan
Languages	Assamese, Bengali, Marathi, Gujarati, Odia, Hindi, Rajasthani	Tamil, Telugu, Kannada, Malayalam	Bodo, Manipuri

Table: Major Indian Languages grouped according to their language families

A common label set (CLS) is designed by exploiting the common attributes present across different Indian languages. The labels in the CLS are standard notations for the phones across thirteen Indian languages. In the CLS, similar sounds across the different languages are mapped together and denoted by a single label. These labels are represented using Roman alphabet set and certain suffixes are also used as the number of sounds in a language exceeds the number of Roman alphabets. International Phonetic Alphabet (IPA) symbols are used as references for the mapping process. The native text represented by the CLS labels is converted to the spoken form by the LTS rules or the parser [5].

2.4.2 Common question set (CQS)

Full context-dependent pentaphone model is the basic unit in HTS. In the pentaphone context, the pentaphone model is in fact the monophone model. The table given below is an example of the pentaphone model. “It is a lovely day.” The monophone and pentaphone context models for the first three words of the text are given [7].



Monophones	Pentaphone context
i	x-x-**i**-tx-i
tx	x-i-**tx**-i-s
i	i-tx-**i**-s-a
s	tx-i-**s**-a-l
a	i-s-**a**-l-a

Table: Pentaphone model (example)

Suprasegmental contexts are included to obtain natural sounding TTS. Context such as position of the phoneme in the syllable, whether the previous syllable is stressed/accented or not, position of the syllable in the word, position of the word in the phrase, the number of phonemes in the current syllable, etc. are considered.

If each language has approximately 50 phones, then the number of monophone models are 50 and the number of pentaphone models are 50^5 [5]. If suprasegmental contexts are included, it will result in a huge number of combinations of models. However, some of the models may not be valid in particular languages and it may also not be possible to cover all these models in the training data. When unseen models are present in the test sentence, problem arises. A decision-tree based context clustering technique is used to overcome this problem. The question set contains a set of questions based on linguistic and phonetic context. A binary tree is used for clustering where the nodes split into two sub-categories based on yes/no questions. Linguistic questions are based on the structure of the sentence while phonetic questions are based on the characteristics of sound such as vowels, consonants, front vowels, stop consonants, nasals, back consonants, sonorants, fricatives, affricatives, etc. It is needed to design the phonetic questions carefully for each language. To design these questions knowledge in acoustic-phonetic characteristics of all sounds in a language are required [5].

Using a common question set (CQS) derived from the common label set (CLS) makes the task of system building quite easier. The CQS is used for training TTSes for multiple Indian languages instead of designing a set of questions for each language. The CQS is represented as the superset of questions across the 13 Indian languages. About 60 questions from the English question set relevant to the Indian languages have been included. The CQS has a fixed number of questions so that irrelevant questions in the question set can be ignored while clustering. One Indo-Aryan language (Hindi) and one Dravidian language (Tamil) have been chosen and questions were first carefully designed for them. Hindi and Tamil were chosen as they covered most of the labels in CLS. After that the CQS was extended to include the remaining labels from CLS [5].

3. Evaluation of TTS system

A variety of subjective evaluations are conducted to evaluate the quality of text to speech synthesis systems. The test sentences that are used in the evaluation process should not be from the training data. Hence, the test sentences are obtained from the internet belonging to different domains of interest like sports, news, travel, business and nature [7]. Mean opinion score (MOS) plays an important role in the evaluation of speech synthesis system. Since the quality of synthesised speech is dependent on quality of original voice, the score obtained by synthesised speech is normalized to natural speech. Hence the name degraded Mean opinion score (DMOS) and it is used for evaluating the naturalness of the TTS system. Word error rate (WER) is used to analyse the intelligibility of the TTS system. TTS system that has high DMOS and low WER has very good quality. The TTS system evaluation is conducted in a noise- free environment and using headphones.

3.1 Degraded mean opinion score (DMOS)

The DMOS test is conducted by playing natural sentences and synthesised sentences using both approaches in some random order. It is better to have a large number of listeners for the proper evaluation of TTS system. The listeners are asked to score the quality of synthesised speech on a scale of 1-5, 5 being excellent (human- like speech) and 1 being poor [3].

3.2 Word error rate (WER)

In WER test, a set of semantically unpredictable sentences synthesised using both approaches are played randomly and the listeners are asked to transcribe the synthesised sentences based on whatever they understood. In WER test, insertions, deletions and substitution of words are considered as errors [3]. WER can be calculated as,

WER= ((insertions+ deletions+ substitutions)/total number of words)*100

4. Applications of synthetic speech

Synthetic speech may be used in widespread applications like electronic mail readers, reading machines for the blind, etc. which require TTS system and good vocabulary. The quality of the TTS system is increasing in a steady rate and so is the use of synthetic speech application. Due to the affordability of TTS systems, it has become more suitable for everyday use. With the availability of TTS system people with communication disabilities can find employment opportunities. Speech synthesis in reading and communication aids is helpful for the visually impaired. The deafened and vocally handicapped have an opportunity to communicate with people who do not know the sign language by using speech synthesis. A computer with speech synthesiser can be used for educational purposes to teach 365 days a year and 24 hours a day. The latest application of speech synthesiser is in multimedia and telecommunication. It can also be used in warning and alarm systems [14].

5. Conclusions

5.1 Summary

TTSes are built for multiple Indian languages in a unified framework, provided that Indian languages are low-resourced languages. TTS synthesisers help people who are visually impaired to listen to the written work. People with learning difficulties like dyslexia, with literacy difficulties and people who can speak the language but cannot read have an opportunity to learn. Children may find it easy to learn the pronunciation of words. People who multi-task, access content on mobile phones and people with different learning styles/methods find TTS system highly efficient [15]. It can also be used in domain specific applications such as railway announcement systems.

On the other hand, it may lack the naturalness of human sound quality. Problems are faced while concatenating the sounds based on words. Difficulty to create a logic for the pronunciation of all the words of a dictionary with accuracy arises [14].

5.2 Future scope

- For the purpose of building robust models for the fundamental units, speech synthesis requires accurate segmentation. Despite machine learning being very successful for segmentation, it is not good enough to build high quality TTS systems. Signal processing techniques are quite accurate in particular while machine learning techniques are quite robust on average. Lack of a mechanism to detect boundaries of nasals can be explored [12].
- The speech looks unnatural when there is a delay in sound wave. The method of integration can be observed where the delay is automatically recognised and removed [14].
- Explore the different factors of prosody.
- In cross-lingual studies the main focus is on textual and durational analyses of syllables. Explore the other acoustic properties of syllables across languages and gender specific studies [7].
- To build TTS systems for new languages with the help of small amounts of adaptation data.
- Explore person-centric approach to build an efficient and practical device for enhancing dysarthric speech.
- To build TTS systems for under-sourced languages. For languages that do not have a written script, the common set phone transliteration can be used to develop systems.

References

1) An implementation of speech recognition for desktop application, IEEE, Chengdu, China, 9-11 July 2010.

2) Arnab Ghoshal and Ayaskant Shrivastava, A Text-To-Speech System For Indian Languages, IIT Kanpur, 2002

3) Hemant A Patil, Tanvina B Patel, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, G R Kasthuri, T Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, S P Kishore, S R M Prasanna, Nagaraj Adiga, Sanasam Ranbir Singh, Konjengbam Anand, Pranaw Kumar, Bira Chandra Singh, S L Binil Kumar, T G Bhadran, T Sajini, Arup Saha, Tulika Basu, K Sreenivasa Rao, N P Narendra, Anil Kumar Sao, Rakesh Kumar, Pranhari Talukdar, Purnendu Acharyaa, Somnath Chandra, Swaran Lata, Hema A Murthy, A Syllable-Based Framework for Unit Selection Synthesis in 13 Indian Languages, IEEE, pages 1-8, Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 25-27 Nov 2013.

4) B. Ramani, S. Lilly Christina, G. Anushiya Rachel, V. Sherlin Solomi, Mahesh Kumar Nandwana, Anusha Prakash, Aswin Shanmugam S., Raghava Krishnan, S. P. Kishore, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan and Hema A. Murthy, “A Common Attribute based Unified HTS framework for Speech Synthesis in IndianLanguages”,in SpeechSynthesisWorkshop(SSW8),pages311-316,Barcelona, Spain, August 2013.

5) Anusha Prakash, M Ramasubba Reddy, T Nagarajan and Hema A Murthy, An Approach to Building Language-Independent Text-to-Speech Synthesis for Indian Languages, pages 1-5 , Twentieth National Conference on Communications (NCC), 2014.

6) Anusha Prakash, Jeena J Prakash, Hema A Murthy, Acoustic Analysis of Syllables across Indian Languages, pages 327-331, Inter speech, 2016.

7) Anusha Prakash, Cross-lingual Speech Synthesis and Enhancement of Dysarthric Speech.

8) Raghava Krishnan K, Prosodic Analysis of Indian Languages And Its Applications To Text To Speech Synthesis.

9) Wikipedia. Source filter model for speech production, https://en.wikipedia.org/wiki/Speech_production

10) Source filter model of speech production (Figure), http://what-when-how.com/voip-protocols/overview-of-speech-signals-voip-protocols/

11) Arun Baby, Anju Leela Thomas, Nishanthi N L and TTS Consortium, Resources for Indian languages, pages 37-43, CBBLR Workshop, International Conference On Text, Speech And Dialogue. Springer, 2016.

12) Aswin Shanmugam S, A hybrid approach to segmentation of speech using signal processing cues and hidden Markov models.

13) S. Aswin Shanmugam, Hema A. Murthy, Group Delay Based Phone Segmentation for HTS, pages 1648-1652, Interspeech, 28 Feb-2 March 2014.

14) N. Swetha, K. Anuradha, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 2, No. 6, pages 269-278 (2013).

15) The advantages of TTS system, http://www.readspeaker.com/benefits-of-text-to-speech/