Chapter 2

A Survey on Speech Synthesis Techniques*

Producing synthetic speech segments from natural language text utterances comes with a unique set of challenges and is currently under serviced due to the unavailability of a generic model for all available languages. This chapter presents a study on the existing speech synthesis techniques along with their major advantages and deficiencies. The classification of different standard speech synthesis techniques are presented in Figure 2.1. This chapter also discusses about the current status of the text to speech technology in Indian languages focusing on the issues to be resolved in proposing a generic model for different Indian languages.

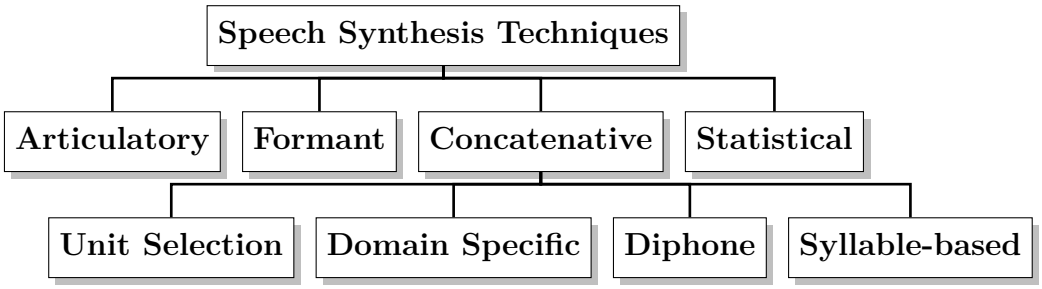

Figure 2.1: Classification of speech synthesis techniques

*S. P. Panda, A. K. Nayak, and S. Patnaik, "Text to Speech Synthesis with an Indian Language Perspective", International Journal of Grid and Utility Computing, Inderscience, Vol. 6, No. 3/4, pp. 170-178, 2015

2.1 Articulatory Synthesis

Articulatory synthesis models the natural speech production process of human. As a speech synthesis method this is not among the best, when the quality of the produced speech is the main criterion. However, for studying speech production process it is the most suitable method adopted by the researchers [48]. To understand the articulatory speech synthesis process the human speech production process is needed to be understood first. The human speech production organs (i.e. main articulators - the tongue, the jaw and the lips, etc as well as other important parts of the vocal tract) is shown in Figure 2.2 [49] along with the idealized model, which is the basis of almost every acoustical model.

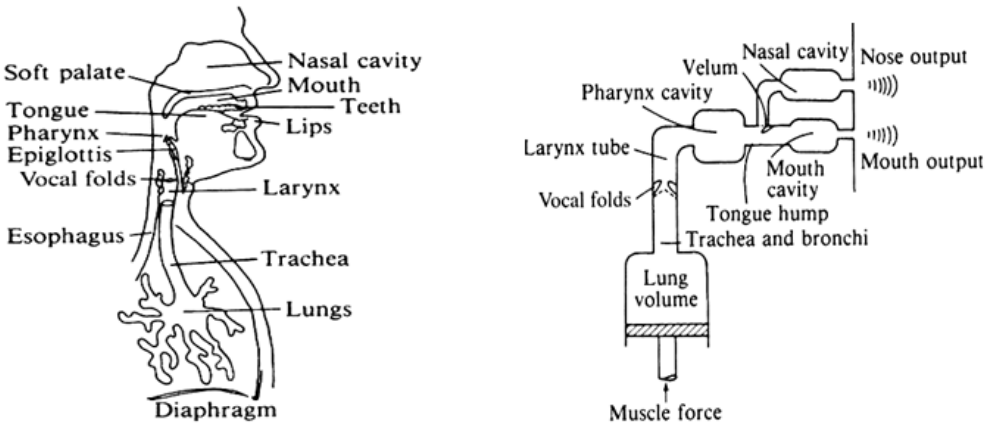

Figure 2.2: The human speech production organs (left) and an idealized model of speech production (right)[‡]

In producing pulmonic sounds, the breathing muscles act as an energy source and the lungs provide storage of pressurized air. The vocal fold or vocal chords separates the lungs from the vocal tract. The signals generated by the vocal folds are filtered by the vocal tract and are then radiated to the surroundings via the mouth and/or nostrils as speech signals [49]. Articulatory synthesis models this natural speech production process as accurately as possible by creating a synthetic model of human physiology and making it speak. For rule-based synthesis the considered articulatory control parameters may be the lip aperture, lip protrusion, tongue tip height, tongue

[‡]Source: D. Qinsheng, Z. Jian, W. Lirong, and S. Lijuan, "Articulatory speech synthesis: A survey", In Proc: 14th IEEE International Conference on Computational Science and Engineering, pp. 539–542, 2011

tip position, tongue height, tongue position, velic aperture, etc and the excitation parameters may be the glottal aperture, cord tension, and lung pressure. While speaking, the vocal tract muscles cause articulators to move and change shape of the vocal tract which causes different sounds.

Articulatory speech synthesis may be used as a tool in basic speech research to understand the speech production process at the deepest levels for construction of speech synthesizer indistinguishable from a natural speaker. Articulatory models are also used for speech recognition for understanding the process of speech production and recognition by modeling it as a whole [50]. Apart from this, the possible application of the synthesis technology may also includes use as a virtual language tutor, in speech therapy, in general purpose audio-visual speech synthesis [51], in speech encoding, in imitation of real speakers [52], as the speech engine of virtual actors and even as a toy [53].

Articulatory synthesis produces highly intelligible speech. However, its output is far from natural sounding speech. The data for articulatory model are normally derived from X-ray analysis of natural speech production process which was traditionally only 2-D making it difficult to optimize in real implementation (the real vocal tract is naturally 3-D). Also, the 2-D X-ray data do not characterize the degrees of freedom of the articulators [54] and hence making it relatively difficult to model the complicated tongue movements. Therefore, traditionally articulatory synthesis technique was considered to be one of the most difficult methods to be implement and thus has received less attention than other synthesis methods and has also not achieved the same level of success. However, as the analysis methods are developing and the computational resources are increasing, researchers are using these techniques these days for analyzing different articulatory parameters of speech production and are able to produce different language speech from a single voice called polyglot synthesis [55]. It has many possible applications today in different fields of voice based feature analysis including the forensic sciences. However, further research may be undertaken to provide completely "tunable" high quality synthesized speech.

2.2 Formant Synthesis

Formant synthesis technique is a rule-based technique closely related to articulatory synthesis. It produces speech segments by generating artificial signals based on a set

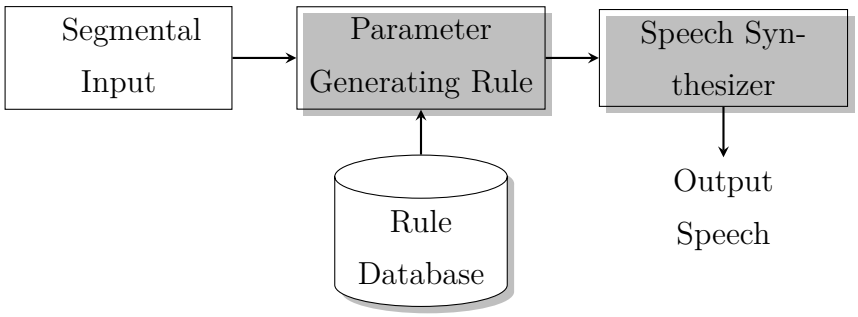

Figure 2.3: Rule-based synthesis approach

of specified rules, that mimics the formant structure and other spectral properties of natural speech as closely as possible [56]. The synthesized speech is produced using an additive synthesis and an acoustic model. The acoustic model uses parameters like, voicing, fundamental frequency, noise levels, etc that varied over time [57]. The overview of the rule based formant synthesis technique is presented in Figure 2.3.

The technique produces highly intelligible synthesized speech, even at high speeds, avoiding the acoustic glitches commonly plagued in concatenative synthesis systems (discussed next). These are usually smaller programs compared to the concatenative techniques as they do not depend on a speech corpus to produce the output speech. Therefore, formant synthesis may be a suitable speech synthesis technique for embedded systems, where memory and microprocessor power are limited. However, the major drawback of the technique is, the system generates artificial, robotic-sounding speech that is far from the natural speech spoken by a human. Also, it is relatively difficult to design rules that specify the timing of the source and the dynamic values of all filter parameters for even simple words [58].

As, formant-based systems have complete control on all aspects of the output speech, today a wide variety of emotions and different tone voices may be produced by incorporating some prosodic and intonations modeling techniques [59]. the formant synthesis technique are widely being used for utterance copy. i.e. for mimicking the voice features that takes speech as input and find the respective input parameters that produces speech, mimicking the target speech [60]. Mimicking the voice characteristic is relatively a difficult and ongoing area of research these days [61]. However, a lot of further research may be undertaken to obtain more natural sounding speech segments by optimizing different speech parameters.

2.3 Concatenative Synthesis

The Concatenative speech synthesis technique is a corpus-based technique that uses some pre-recorded speech samples (words, syllables, half-syllables, phonemes, diphones or triphones) in a database and produces the output speech by concatenting appropriate units based on the entered text utterances [62]. The simplicity of the model and highly natural speech production quality makes it suitable for its use in designing human computer interactive systems for different domains [63]. The overview of a concatenative speech synthesis system based on unit selection technique is presented in Figure 2.4 and is discussed next.

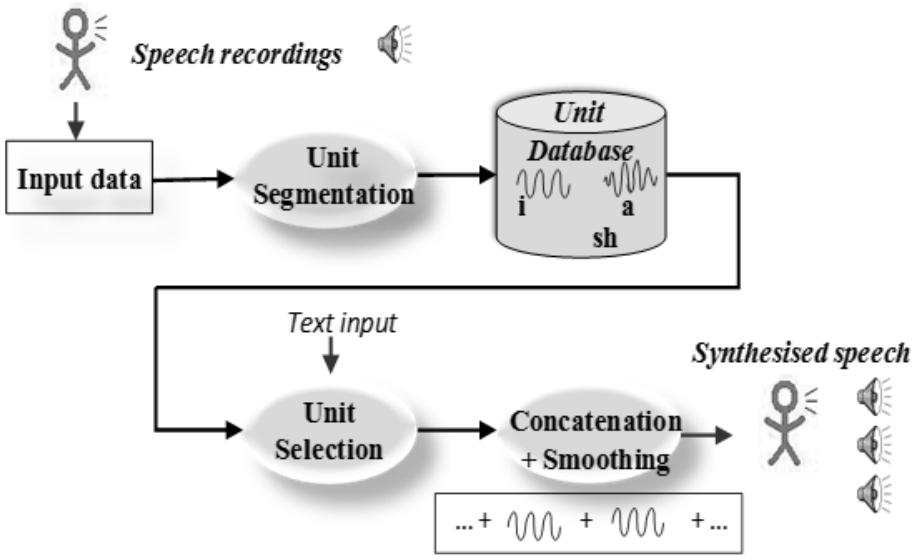

Figure 2.4: Unit selection approach in concatenative synthesis

The quality of the synthesized speech is affected by the unit length in the database [64]. The naturalness of the synthesized speech increases with longer units while by using longer units less concatenation points are there reducing the formation of unnatural segments at concatenation points. However, more memory is needed and the number of units stored in the database becomes very numerous. On the other hand with shorter units, the memory requirement is less but the complexity of sample collection and labeling techniques increases. The concatenative technique may broadly be categorized into the following three types based on the unit type stored in its database.

2.3.1 Unit Selection Synthesis

Unit selection synthesis uses a large databases of recorded speech. During database creation, each recorded utterance is segmented into individual phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones [65].

At runtime, the desired target utterance is created by determining the best chain of candidate units from the database. This process is typically achieved using a specially weighted decision tree [66]. In the unit selection scheme, by using the target cost and the concatenation cost, speech units are selected from the whole speech database, and concatenated in run-time. In this scheme, a heuristic distance is defined between contexts to measure the target cost. To avoid this, a clustering-based scheme may be used which clusters the contexts in advance, and selects each unit from a cluster [67]. A typical decision tree based on concatenation cost for unit selection and is presented in Figure 2.5 [67] along with an overview of the clustering based unit selection scheme.

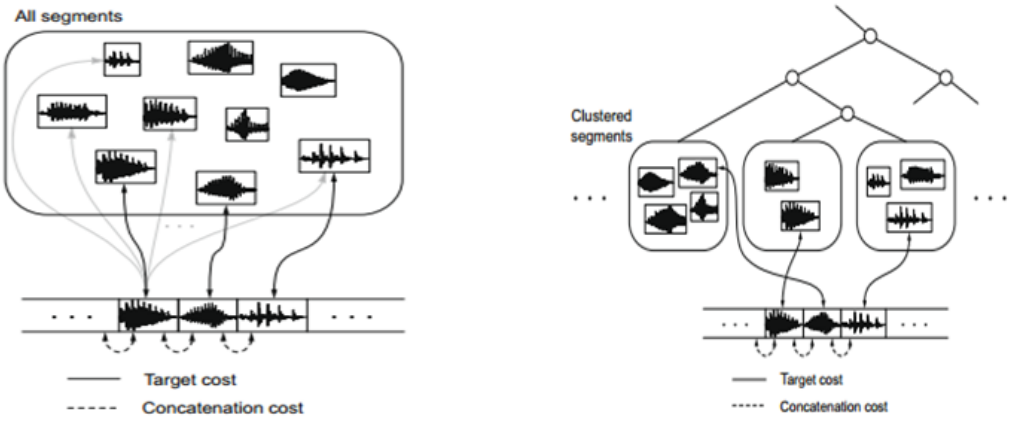

Figure 2.5: Overview of general unit-selection scheme (left) and clustering-based unit-selection scheme (right)[¶]

[¶]Source: N. P. Narendra and K. S. Rao, "Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis", Applied Soft Computing, Vol. 13, pp. 773–781, 2013

As only a small amount of digital signal processing is applied to the recorded speech, unit selection technique produces highly natural speech segments [67]. However, maximum naturalness typically requires unit-selection speech databases to be very large. Also, unit selection algorithms have been known to select segments from a place that results in less than ideal synthesis. For example minor words become unclear, even when a better choice exists in the database [68]. Recently, researchers have proposed various automated methods to detect unnatural segments in unit-selection speech synthesis systems [66, 69–71] however, a lot of work may further be done to achieve better performance.

2.3.2 Diphone Synthesis

In this type of synthesizers, all the diphones (sound-to-sound transitions) occurring in a language are contained in the speech database [72]. A diphone is made of two connected half phones and captures the transition between two phones by starting in the middle of the first phone and ending in the middle of the second one [73]. The number of diphones depends on the phonotactics of the language: for example, 800 diphones for Spanish, and about 2500 diphones for German. Identifying the number of diphone units in a language is a challenging task. Only one example of each diphone is contained in the speech database and at runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as LPC [74], PSOLA [75] or MBROLA [76].

In diphone synthesis, with only one instance of the speech unit being available, extensive prosodic modifications have to be applied to obtain good quality speech. Diphone synthesis suffers from the robotic-sounding synthesized speech. Although due to a number of freely available software implementations, it continues to be used in research but its use in commercial applications is declining. TTS systems based on diphone synthesis need prosodic models to produce good speech output. The prosodic analysis for these models require a database of speech annotated with linguistic and prosodic labels. Tools are also required to generate appropriate linguistic information essential to predict prosody from text [77].

2.3.3 Domain Specific Synthesis

The domain specific synthesis stores recorded speech samples of some commonly used words or phrases for particular domains and concatenation of those segments is

performed to create complete utterances [78]. It is mostly used in applications like announcement of transit schedules, weather reports, railway inquiries etc where the variety of texts the system will output is limited to a particular domain [79]. This technology is in commercial use for a long time for its simple and easy to implement characteristics. It can be implemented in devices like calculators, talking clocks, etc [80].

The major advantage of this approach is the level of naturalness. These systems can have a very high naturalness as the variety of sentence types is limited, which closely matches the prosody and intonation of the recorded original speech [81]. The main disadvantage of this type of techniques is the limitation of words and phrases in the databases [82]. These types of systems can only produce speech that combines the words and phrases in the database with which they are preprogrammed. Therefore, these types of TTS systems are not general-purpose. The blending of words within naturally spoken language however can still cause problems unless the many variations are taken into account.

2.3.4 Syllable-based Synthesis

Apart from the three broad categories of the concatenative technique discussed above, the syllable-based technique is another concatenative technique used these days that relies on the syllable units in any language to achieve high quality speech segments. The syllable-based speech synthesis technique stores recorded speech samples for a common set of syllable units in any language. The syllables are often considered to be the phonological building blocks of a word [83]. A syllable is typically made up of a syllable nucleus (a vowel) with an optional initial and final margins (consonants). Creating a syllable based speech synthesis system requires identifying the respective syllable units in the language [84]. The total number of syllable units in any language may vary between 400-800 units depending on the language [84].

There are a number of syllable based techniques available for different languages, however, this technique achieves high quality results for the syllable-timed languages like the Indian languages, where the pronunciation relies on the syllable units in the word. As a syllable may comprise different combination of consonant or vowels highly natural quality speech may be generated. The syllable units used for deriving the pronunciation vary from language to language and unavailability of a unit may cause audio discontinuity. Thus adding a new language requires identification of

syllable units in the respective language and creation of a new speech corpus [85,86]. Designing a speech database of syllable units that may work for different syllable-based languages is a challenging task.

2.4 Statistical Parametric Synthesis

Statistical approaches have recently been shown as very effective methods in various fields of speech processing. While, the spectral features in Statistical Parametric Synthesis (SPS) are represented by mel-log spectral approximation based cepstral coefficient, line spectral pairs and harmonic noise models features etc, the excitation features are represented by fundamental frequency and voicing strengths [87, 88]. Speech signals are generated by excitation and spectral features using source filter models. The overview of a statistical speech synthesis model using HMM based approach is presented in Figure 2.6 [89]. The statistical models extracts the suitable parameters needed for synthesizing speech from a set of training utterances and models the parameters using some statistical methods. A set of parameters are then generated from the trained model and synthesized speech is produced from the generated parameters.

To build a database with multiple instances of each phone in different context for unit selection is a time consuming task and the database size increases in an enormous way. Also, the concatenative synthesis approach has a limitation on recreation of the recorded samples. The major advantage of the statistical techniques is that instead of storing data, the parameters of the model are stored reducing the memory requirement of the system. Also, different speech parameters can be modified causing more variations of the recorded speech. For example, the original voice can be converted into another voice [90–93].

While designing the model, the extracted parameter's behavior considered should be simple enough to model easily in the subsequent steps. Also, they should contain much information to generate less distorted speech. In case of the Statistical parametric methods, a large amounts of speech information is discarded for applying spectral modifications. Therefore, the quality of the output speech in statistical parametric technique is lower than the unit selection techniques [94]. However, as statistical methods are parametric, they are flexible and the generated speech may be modified by changing the system parameter values [94]. Speaker adaptation, interpolation

and eigen voices are three examples of this flexibility which enable the synthesized speech to be modified without the necessity of large data sets.

The emergence of statistical speech synthesis techniques such as ANN based speech synthesis [95], Gaussian Process Regression based synthesis [96], etc has enhances the quality of synthesized speech up to certain level. However, a huge set of work is still needed to achieve high quality synthesized speech as like natural speakers [97]. The HMM-based [89] synthesis achieves the highest level of success in this regard and is used by most of the researchers in designing TTS system in many languages including some Indian languages. The HMM based synthesis is discussed next.

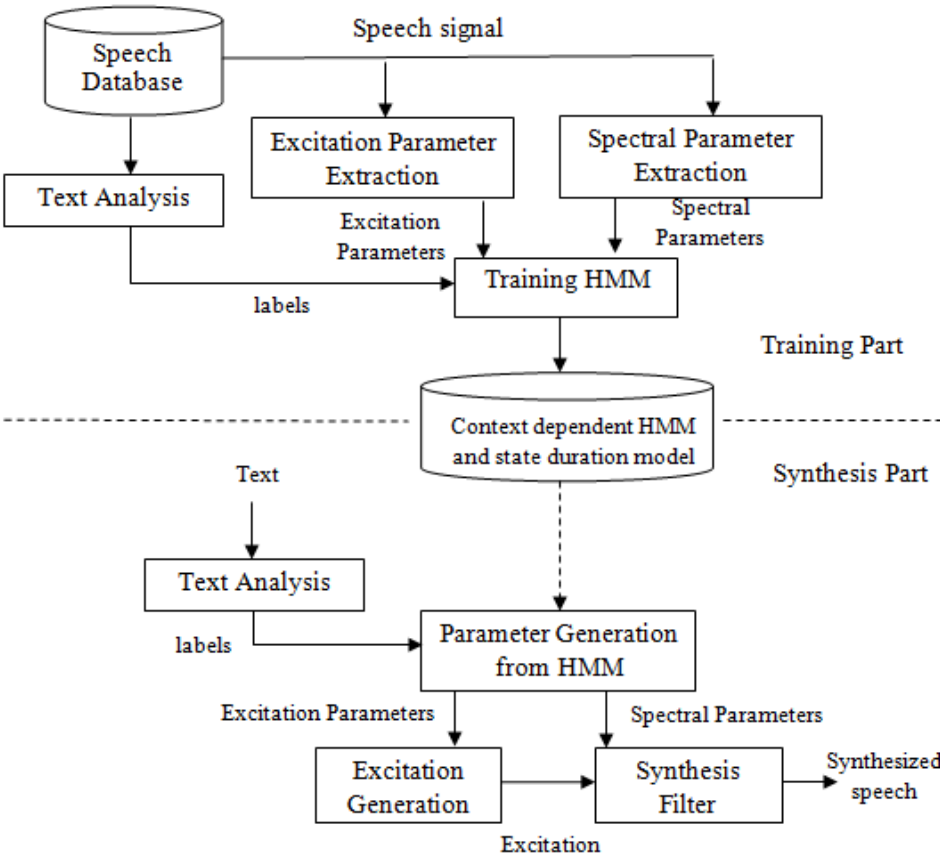

Figure 2.6: The HMM based statistical speech synthesis system‖

‖Source: K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models", Proceedings of the IEEE, Vol. 101, pp. 1234–1252, 2013

2.4.1 Hidden Markov Model based Synthesis

HMM-based synthesis is a statistical parametric speech synthesis method based on Hidden Markov Models (HMM). In this technique, the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion [98]. Figure 2.6 [89] shows the overall HMM based speech synthesis process. It consists of two phases, the training phase and the synthesis phase. The spectrum and excitation parameters are extracted from speech database and modeled by context dependent HMMs in the training phase. A model usually consists of three states that represent the beginning, the middle and the end of the phone. The synthesis phase deals with generation of speech signals by concatenating the context dependent HMMs according to the text to be synthesized [99].

The major advantages of the HMM-based speech synthesis approach is its voice characteristics can be easily modified and can be applied to various languages with little modification [100]. A variety of speaking styles or emotional speech can be synthesized using the small amount of speech data. Also the techniques developed in Automatic Speech Recognition (ASR) can be easily applied to it and its footprint is relatively small.

The major disadvantage of the HMM-based synthesis approach against the unit selection approach is the quality of synthesized speech. There seems to be three factors which degrade the quality: vocoder(artificial sounds from an analysis of speech input), modeling accuracy, and over-smoothing. The synthesized speech by the HMM-based generation synthesis approach sounds buzzy since it is based on the vocoding technique [101]. To alleviate this problem, a high quality vocoder such as multi-band excitation scheme or STRAIGHT [102] have been integrated. Several groups have recently applied LSP-type parameters [103] instead of mel-cepstral coefficients to the HMM-based generation synthesis approach. The Hidden Semi-Markov models (HSMMs) [104], trajectory HMMs [105], and stochastic markov graphs [106] are some other variations to obtained enhanced modeling accuracy. Further, the accuracy of acoustic parameter prediction and the naturalness of synthesized speech can be improved when shared clustering and asynchronous-state model structures are adopted for combined acoustic and articulatory features [107].

There are a number of other variations of the HMM based approach like the model

discussed in [108], that integrates the Harmonic plus Noise Model (HNM) [109] into the HMM-based speech synthesis system (HTS) [110] to minimize the development cost and time. A HMM based on multi-space probability distribution is discussed in [111] to characterize the sequence of speech spectra. this is being used successfully in speech recognition systems and is also useful for modeling pitch patterns of speech. Also, the modeling of fundamental frequency, or F_0, in HMM-based synthesis is important for output speech which is both natural and accurate [112]. Continuous F_0 modeling yields better synthesized F_0 trajectories and provides significant improvement to the naturalness of synthesized speech.

[113] Literatures a hybrid concatenative technique, that combines concatenative and statistical synthesis units and where, the positions of the statistical models are defined by a hybrid dynamic path algorithm. The DNN [114] (Deep Neural Network) based approach is another variation of the statistical synthesis approaches that is used to overcome the inefficiency of decision trees used in HMMs to model complex context dependencies. In these techniques, the relationship between input texts and their acoustic realizations is modeled by a DNN. The acoustic features are created using maximum likelihood parameter generation(MLPG) trajectory smoothing. The DNN based synthesis techniques are currently the most emerging area of research. However, a huge set of research is needed for their adaptation in different Indian languages.

2.5 Comparative Study of Speech Synthesis Techniques

Table 2.1 presents a comparative study of the discussed speech synthesis techniques with respect to their advantages and disadvantages.

2.6 Status of Text-to-Speech Synthesis Technology in Indian Languages

Different studies have investigated speech synthesis in different cultures, most of which have focused on achieving natural sounding speech in the respective languages [115]. There are fewer models available in different Indian languages for text to speech

Table 2.1: Comparisons of speech synthesis techniques

Technique	**Method Used**	**Advantages**	**Disadvantages**
Articulatory	Mathematical model of human speech production	Needs no speech database	Robotic sounding speech
Formant	Rule-based	Needs no speech database	Robotic sounding speech
Unit Selection	Concatenative	highly natural speech	Database dependency
Diphone	Concatenative	highly natural speech	Database dependency
Domain Specific	Concatenative	highly natural speech	Speech for limited words may be produced
Syllable-based	Concatenative	highly natural speech	Database dependency
SPS	Statistical parametric	provides voice modifications	needs large training data
HMM-based	Statistical parametric	provides voice modifications	needs large training data

synthesis. These studies demonstrated that, although speech synthesis in different languages may have different issues due to the pronunciation variations, the quality of the synthesized speech is regulated by intelligibility and naturalness of the produced speech [31]. While intelligibility refers to the understandability of the artificially produced speech, naturalness refers to how closely it seems like a human generated speech. Individuals are able to produce intelligible speech in different regional languages [116]. In Xia's study [109], the unit selection scheme achieves better results for languages like, English, French, Japanese, etc where the major objective is to achieve highly natural speech. However, when the judged expressions were short the unit selection scheme provides highly natural speech compared to other approaches. At the same time by increasing the length of the text, the processing overhead increases.

A discussion in [117] is presented for the size of the speech database in Indian languages. The researchers in [45] presented the efforts to build a high quality syllable-based framework for unit selection for 13 Indian languages. However, the appropriate size of the speech database needed for unit selection synthesis in all official Indian languages is still an open question. Kishore's study in [118] discusses about the appropriate unit size for unit selection speech synthesis. A discussion in

[119] and [120] is presented on the associated challenges for designing corpus based speech synthesis systems. Murty in [46] discusses about the initiative in building unit selection speech synthesis system in Indian languages. However, the larger size of the unit selection speech database increases the difficulty for its use in small hand held devices having limited storage resources. The Festival speech synthesis system [121] provides an open architecture for multi lingual speech synthesis research, where some Indian languages like, Hindi, Marathi, Tamil and Telugu are included along with other western languages (English).

There are fewer researches arguing the syllable based speech synthesis techniques achieves better results compared to the other techniques [41]. The researchers in [122] and [47] proposed different models for syllable based speech synthesis in different Indian languages. Thomas in [123] discusses about a syllable like units for speech synthesis in Tamil language where the units are automatically generated using a group delay based segmentation algorithm. However, the syllable units used for speech synthesis are different for different languages. Therefore, adding a new language requires analysis of the syllable units in that language and adding the respective recorded sound units in its database which increases the database size as the number of considered languages increases.

There are fewer studies that investigated the relationship between the speech database and the naturalness of the produced speech. Narendra's study [122] argues that, even though the syllable based concatenative technique requires around 800 to 1000 number of speech units in the form of syllables to be stored in its database, the sound to sound transition is well maintained and better naturalness may be achieved. [124] presented a global syllable set for speech synthesis in three Indian languages: Hindi, Tamil, and Telugu. Talesara [125] and Rama [126] provide another model where the use of diphone or triphones units may be more useful in producing natural speech. The dhvani- TTS system for Indian languages [127], uses a syllable based concatenative technique to produce intelligible speech segments in 11 official languages of India requiring a large number of speech units to be stored in its database. However, in applications where storage and computational resources are limited like for human computer interactive systems in small handheld devices, these models may further be optimized to achieve better results.

As naturalness is not the only objective to be achieved for better user satisfaction, aspects such as storage and computation overhead must be considered while designing

a TTS system for human computer interactive systems. The TTS systems must be capable of achieving highly intelligible and natural sounding speech with less storage and computation overhead. Based on the previous studies, it may be observed that the existing techniques are somehow lacking in providing the requirements for creating a generic model for speech synthesis in different Indian languages.

2.7 Chapter Summary

In this chapter, the different speech synthesis techniques are discussed along with their advantages and deficiencies. Broadly they may be categorized into two types, rule based techniques or corpus based techniques. While, the rule based techniques produces robotic sounding speech, the corpus based techniques requires creation of a speech corpus in the specific languages requiring a large speech corpus for designing a generic model for different languages. Also, the existing speech synthesis techniques are somehow lacking in achieving the desired objectives for its use in small hand held or human computer interactive systems with limited storage and computation resources. Development of a generic model for all the official Indian languages is still an open area of research. A comparative study of the existing speech synthesis techniques with respect to the considered parameters are presented in Table 2.2.

Table 2.2: Comparison of speech synthesis techniques with respect to considered parameters

Technique	**Corpus Size**	**Complexity**	**Intelligibility**	**Naturalness**
Articulatory	Corpus independent	Very high	High	Low
Formant	Corpus independent	High	High	Low
Concatenative	Large	Low	High	High
Syllable-based	Large	Low	High	High
SPS	Large	Low	High	High
HMM-based	Large	Low	High	High