Chapter 2

Synthesis of dialectal speech: A Survey

This chapter presents a brief survey of speech synthesis methods, especially those related to the generation of dialectal speech. The scope of speech synthesis is very wide and therefore only those topics which were studied during the course of this doctoral research work are presented here in brief. The main purpose of this survey is to review various works in the area of dialectal speech synthesis in order to find out the lapses or limitations and gather necessary information required to carry out the research work in a smooth and consistent manner.

The rest of the chapter is organised as follows. Section 2.1 briefly introduces the technique of speech synthesis, with various sub-sections briefly describing the prevalent speech synthesis methods. Section 2.2 highlights on the various ways of incorporating naturalness to speech, with sub-sections 2.2.1 describing the modelling of prosody, 2.2.2 describing the modelling of the voice source (VS), 2.2.3 describing how Voice Conversion (VC) techniques can be used to produce expressive speech and 2.2.4 describing the modelling of formant frequencies. Section 2.3 gives a brief introduction to the synthesis of dialectal speech and Section 2.4 introduces various possible differences or variations among the dialects of a language. Finally, limitations of the current approaches towards the generation of dialectal speech, based on literature review, are presented in Section 2.5.

2.1 Synthesis of Speech

Speech synthesis systems have been in existence since the 18th century. Wolfgang von Kempelen, developed a speaking machine [1] at Haskins Laboratories way back in 1791, based upon his observations of human speech production. This talking machine was reported to speak complete phrases in both French and Italian. Since then, a deeper understanding of the human speech production system together with technological advancements led to the development of electronics-based speech synthesis. Speech synthesis refers to the artificial production of human voice or speech. The ultimate goal of a speech synthesis system is to produce an intelligible, natural sounding voice in order to convey information to the user in the desired accent, language and voice. The output quality of a speech synthesiser is measured based on mainly two factors, naturalness and intelligibility. A speech synthesis system attempts to maximize both these characteristics. These two characteristics of speech, are usually very difficult to measure. Various methods, both subjective such as Mean Opinion Score (MOS), Differential Mean Opinion Score (DMOS), Diagnostic Rhyme Test (DRT), ABX, etc., and objective such as Mel Cepstral Distortion (MCD), Root Mean Square Error (RMSE) etc, are used for evaluating the quality of synthesised speech with varied levels of accuracy. Similarly, synthesised speech can be produced by several different methods. Formant synthesis models the pole frequencies of the speech signal or the transfer function of the vocal tract, based on the source-filter model of speech production. Concatenative synthesis, concatenates variable length prerecorded samples extracted from natural speech. Articulatory synthesis attempts to model the human speech production system directly. Parametric synthesis constructs a model of the acoustic properties of the human vocal tract, and then analyses speech by determining the values of the parameters of the model. The following subsections present brief descriptions of these techniques used for the synthesis of speech.

2.1.1 Formant Synthesis

Also referred to as Synthesis-by-Rule, formant synthesis is the first truly crystallised method of synthesising speech[2]. It dominated synthesis implementations for a long period, i.e., until the early 1980s. Based on the well known source-filter theory

[1] http://www.haskins.yale.edu/featured/heads/SIMULACRA/kempelen.html

[2] http://hcewiki.zcu.cz/hcewiki/index.php/Speech_synthesisers

of speech production, this method is flexible and relatively easy to implement. It can be used to produce almost all sounds; however the simplifications made in the modelling of the source and the filter, lead to some what unnatural sounding results.

Spectrograms of speech signals reveal that the frequency content of speech is dynamic in nature. During the production of speech articulatory movements change the geometry of the vocal tract which enhances certain resonances. These regions of enhanced resonances exhibited by the vocal tract are called formants. Formant synthesis seeks to mimic human speech by artificially creating the movements of these resonant frequencies or formants. There can be any number of formants. However at least three formants are usually required to produce intelligible speech and a maximum of five formants are required to produce speech of high quality [3]. At least two formant regions are usually recognised as uniquely characterising the different vowels. A more complete description of front vowels requires F3 as well. F4 and F5 are said to contain speaker specific information [10]. Each formant is usually modelled with a two-pole resonator which enables both the formant frequency and its bandwidth to be specified [32].

A basic model of a Formant Synthesiser is presented in Figure 2.1. Speech sounds are generated from a periodic source for voiced sounds and from an aperiodic source for unvoiced sounds. This source signal is fed into the vocal tract model where the oral and nasal cavities are modelled separately. The signal therefore passes into the component which models the oral cavity, or into the component for modelling the nasal cavity in case of generating a nasalized sound. Finally, the outputs from these components are combined and passed through a radiation component, which simulates the load and propagation characteristics of the lips and nose [139].

In order to build a vocal tract model the formants can be combined in two different ways as shown in Figure 2.2. In the parallel formant synthesiser, the excitation signal is applied to all the formants in parallel and the outputs are then summed. This enables individual gains to be specified for each formant. In the cascade formant synthesiser, the output of one formant is applied to the next.

The Klatt synthesiser [76], developed by H.W.Klatt at MIT, Cambridge in

[3] http://piisami.net/dippa/chap5.html

[4] http://svr-www.eng.cam.ac.uk/~pat40/ttsbook_draft_2.pdf

[5] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.7062&rep=rep1&type=pdf

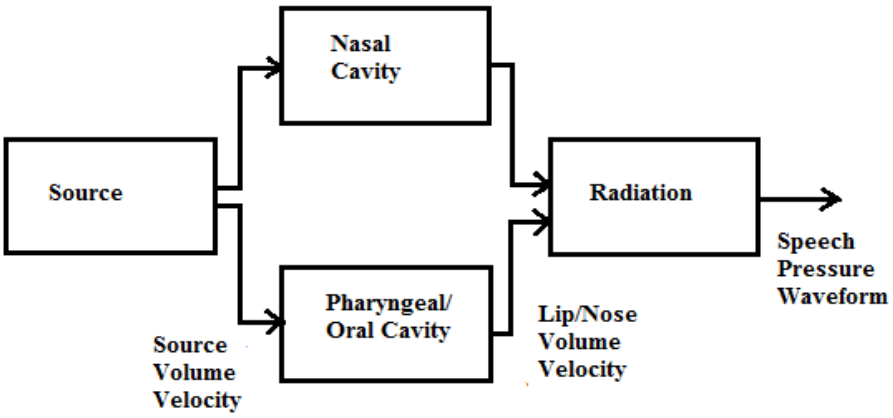
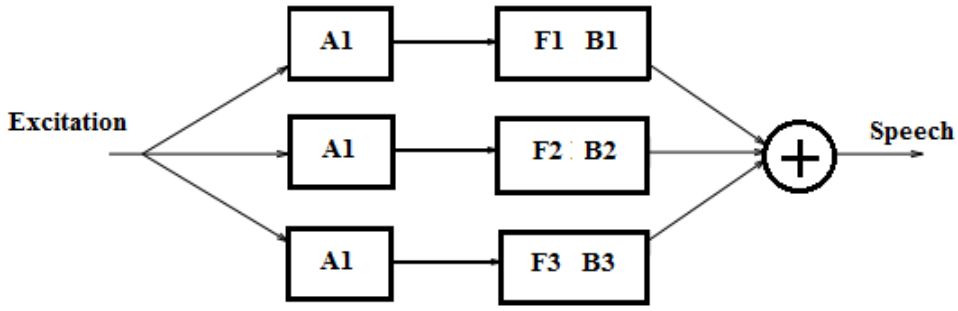
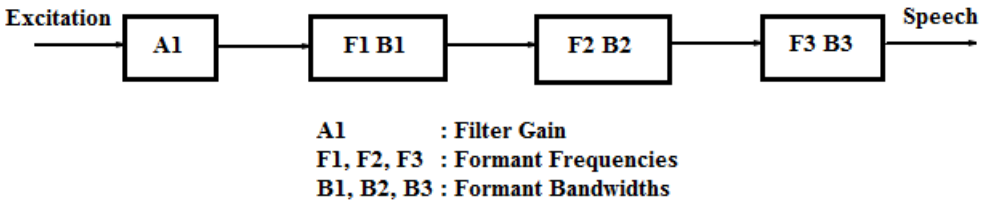


Figure 2.1. Basic Formant Synthesiser [4]



(a)



(b)

Figure 2.2. (a) A Parallel Formant Synthesiser (b) A Cascade Formant Synthesiser [5]

1980, is one of the most sophisticated formant synthesisers developed. A diagram of this system is shown in Figure 2.3. This synthesiser is a hybrid of parallel formant synthesisers and cascade formant synthesisers. This allows the modelling of vowels by using the cascade configuration. The fricatives and stop bursts are modelled by using the parallel configuration [6]. It also used additional resonances

and anti-resonances to help synthesise nasalized sounds. The Klatt synthesiser was set up to work at a sampling rate of 10KHz and 6 formants, F1 through F6, are used. Early literature on formant synthesis uses a sampling rate of 8KHz or 10KHz. This is mainly because space, speed and output requirements at that time did not support the use of higher sampling rates. Higher sampling rate, as required today, is also supported. When generating speech, all the 39 parameters used by the Klatt Synthesizer to generate speech such as formant values (F1-F6), bandwidth (B0-B6), amplitude of voicing (AV), fundamental frequency (F0) etc., are changed on a sample by sample basis usually at a slow rate of parameter update, such as every 5ms. This synthesiser could synthesise very high quality speech and has been incorporated into several TTS systems. It is also highly popular among researchers in the field of speech synthesis.

In general a formant synthesiser is known to produce clean and intelligible speech, but speech lacking naturalness. This may be attributed to the simplicity of the source model. Secondly, the target and transition model is also too simplistic and many of the subtleties actually involved in the dynamics of speech are ignored. Furthermore, the assumptions made about the nature of the vocal tract model do have some effect. Although each of these assumptions is valid on its own, the lack of precision accumulates and affects the overall output of the model .

2.1.1.1 Rule-based Formant Synthesis

Rule-based formant synthesis uses a set of rules to determine the parameters necessary for synthesising an utterance using a formant synthesiser. These rules are generally used together with the set of phoneme string specification of the utterance to be synthesised. They are used to determine which allophones are to be used in what context and also specify how these allophones as well as their transitions should be produced. Kelly and Gerstman [72] attempted to construct the first rule-based formant synthesis system. They used a three formant synthesiser with rules derived from their own experimental results. Rule-based formant synthesis has resulted in the development of a number of TTS systems such as MITalk [6], the Prose-2000 [46] introduced by Speech Plus Inc. and Klattalk [75] which was licensed to Digital Equipment Corporation (DEC) and later became DECtalk.

[6] `http://linguistics.berkeley.edu/plab/guestwiki/index.php?title=Klatt_Synthesizer`

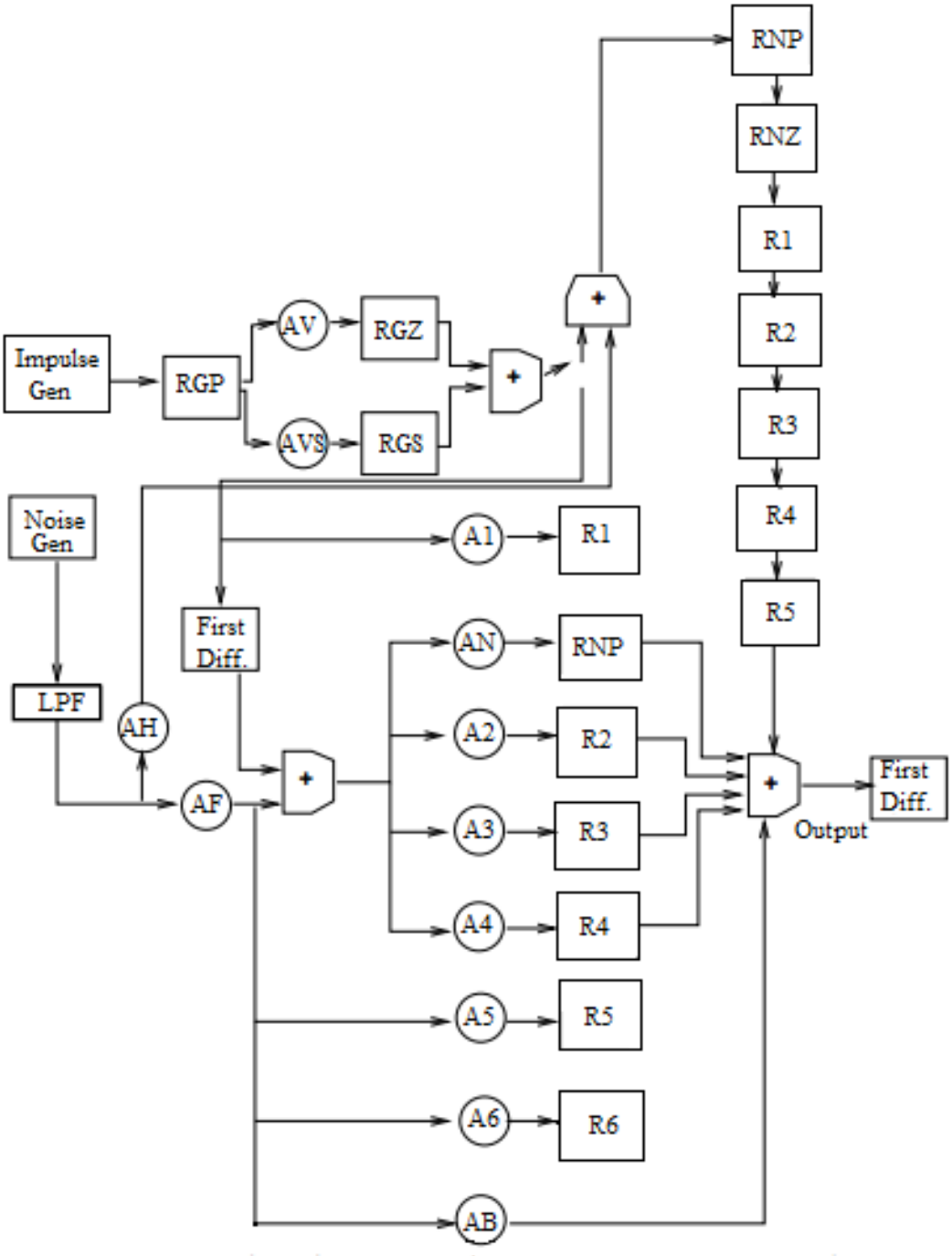


Figure 2.3. The Klatt Synthesiser [76]

2.1.2 Concatenative Synthesis

This method of synthesis is the most popular method because of its ability to produce the most natural sounding output. The implementation is also very simple mainly because of its 'cut and paste' nature where segments of stored speech are concatenated to provide the final output. However concatenative synthesis is

limited to one speaker and one voice in addition to having requirements of high memory capacity. The greatest challenge in this method of synthesis is the discontinuity introduced at the points of concatenation. Another challenge is the selection of the unit size and also the best unit for concatenation. The selection of unit size is a trade-off between longer and shorter units. The longer the selected units are, the fewer will be the concatenation points leading to high naturalness, but at the same time the memory requirements will increase since the number of such units will be more. For example, if the smallest unit, i.e., the phoneme is considered as the basic unit, the number will be around 30-40 for most languages. But if the syllable is considered as the basic unit, the number will be in hundreds. Units such as words, syllables, demisyllables, phonemes, diphones and triphones are mainly used in today's systems. Phonemes are probably the most commonly used, because they are the normal linguistic representation of speech. However since most Indian languages are syllable centric in nature, syllable is the preferred unit for building TTS systems for Indian languages. A new method for concatenating "acoustic inventory elements" of different sizes is described in a paper by Olive [101]. Another system, developed at ATR, is also based on non uniform units [138]. The creation of unit inventories itself is a huge challenge.

Modern concatenative systems utilize extremely large speech corpora from which to draw their speech segments. Such systems are called unit-selection concatenative synthesisers, emphasizing perhaps that the key to synthesis is not just the concatenation of speech segments but the selection of segments with the minimum concatenation cost. Hunt and Black [58] presented a selection model as shown in Figure 2.4, which actually existed previously in ATR vtalk [123]. The basic notion is that of a target cost, i.e., how well a candidate unit from the database matches the required unit, and a concatenation cost, which defines how well two selected units combine. Designing of cost functions is another direction of research. However with the evergrowing size of databases, time dependent voice quality variations have become a serious issue. These extremely large databases require substantial amount of computing resources that limit the use of unit-selection techniques in embedded devices or where multiple voices and multiple languages are required.

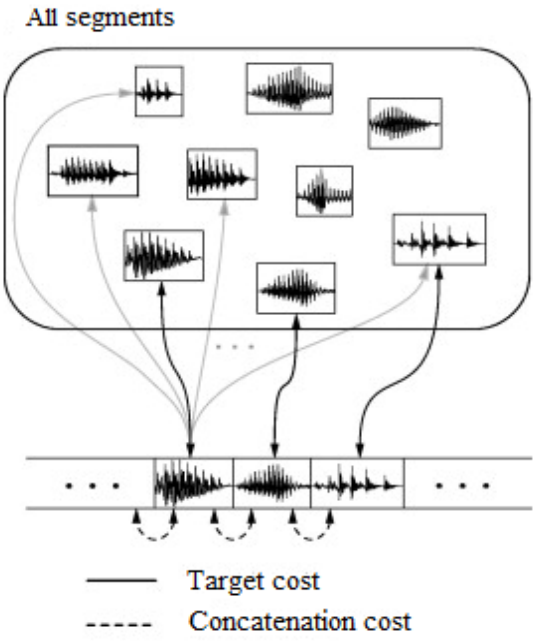


Figure 2.4. Overview of General Unit Selection scheme (Source:[164])

2.1.3 Articulatory Synthesis

This method is by far the most complicated method of synthesis with respect to its model structure and computational burden. It is also the oldest method in the sense that the famous talking machine of von Kempelen can be seen as an articulatory synthesiser. The basic idea of this method is to model the human speech production mechanism as perfectly as possible, which is not an easy task since speech production is a very complex process and not fully understood in every detail. "The speech wave is the response of the vocal tract filter system to one or more sound sources. This simple rule, expressed in the terminology of acoustic and electrical engineering, implies that the speech wave may be uniquely specified in terms of source and filter characteristics". This statement was made in Fant's fundamental book "Acoustic theory of speech production" [36] and since then this has been the foundation for both formant synthesisers and articulatory synthesisers. Articulatory synthesisers determine the characteristics of the vocal tract filter by means of a description of the vocal tract geometry and place the potential sound sources within this geometry. Typically, the vocal tract is divided into many small sections whose dimensions collectively determine its resonant characteristics. This method of synthesis classifies speech in terms of movements of the articulators, i.e., the tongue, lips and velum, and also the vibrations of the vocal cords. The phonetic and prosodic description of the text to be synthesised, is converted into a

sequence of such movements of articulators, and the synchronisations between their movements are calculated. A complex computational model of the human vocal tract is then used to generate a speech signal under control of these movements. The input to the acoustic simulation is usually a piecewise constant area function which corresponds to the vocal tract composed of several cylindrical tube sections as shown in Figure 2.6. The figure shows how the vocal tract is excited by a glottal volume velocity (GVV) function which is the acoustic source, and radiates an acoustic pressure wave at the nostrils and the mouth opening.

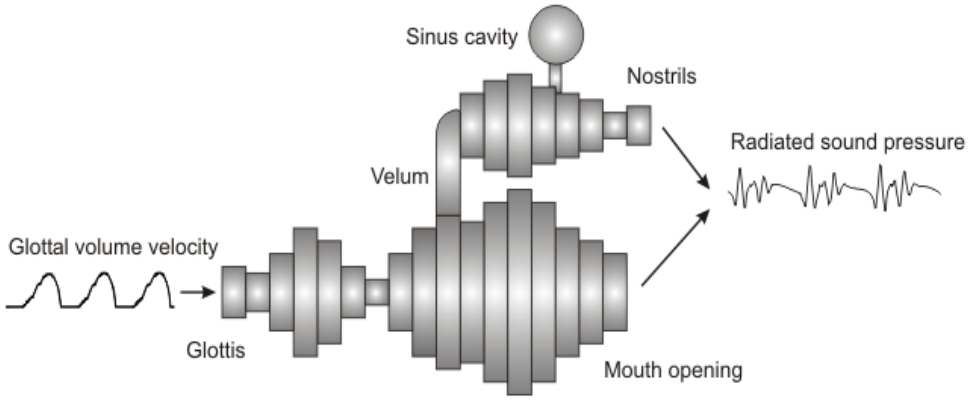


Figure 2.5. Tube Model for Articulatory Speech Synthesis [7]

One major problem with this method of synthesis is parameter generation. Correct articulatory parameters cannot be extracted directly from recordings; rather measures which are intrusive, such as x-ray photography, magnetic radio imaging (MRI) or electro-magnetic acoustic (EMA) imaging are used for parameter extraction. Methods to obtain articulatory parameters from a speech signal remains an on-going research direction and several attempts have been made by researchers. Black et al. [13] describe some of the results from the project entitled "New Parameterization for Emotional Speech Synthesis" held at the Summer 2011 JHU CLSP workshop where methods for deriving articulatory features from speech, predicting articulatory features from text and reconstructing natural sounding speech from the predicted articulatory features have been designed. Mitra [94] presents a deep neural network (DNN) to extract articulatory information from the speech signal and explores different ways to use such information in a continuous speech recognition task. An advantage of articulatory synthesis is that the tract models allow accurate modelling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior.

[7] `http://www.vocaltractlab.de/index.php?page=background-articulatory-synthesis`

2.1.4 Statistical Parametric Synthesis

Statistical parametric synthesis systems use inventories of acoustic features extracted from the speech corpus. Therefore such systems have lesser memory requirements as they need to store the features instead of the actual speech data. These systems also have the advantage of producing varied speech styles by varying the parameters. Simply put, statistical parametric speech synthesis may be described as generating the average of some sets of similarly sounding speech segments [164]. This is in direct contrast to the unit selection method of speech synthesis where actual instances of speech from a huge database is selected. In a typical statistical parametric speech synthesis system, the first step is to extract parametric representations of speech including spectral and excitation parameters from a speech database and then model them by using a set of generative models. Speech parameters for a given word sequence are generated using the estimated models and finally a speech waveform is reconstructed from these parametric representations of speech. Most common technique is the Hidden Markov Model (HMM) based speech synthesis where the generative model is an HMM. Main advantage of HMM based parametric speech synthesis is its flexibility, since speech is stored in the form of parameters which can be very easily modified. Additionally a limited speech corpus is sufficient to train the system. At the same time, it lacks naturalness in the synthesised speech due to over smoothing of the parameters requiring additional signal processing.

HMM based speech synthesis is a statistical parametric model that extracts speech parameters from the speech corpus, trains the system and produces sound equivalent to the input text. Adaptation to new speakers and speaking styles is simple as it involves modification of HMM parameters using various techniques. To develop such a HMM based system, the popular HTS toolkit or HTK [8] is used. A typical HMM-based speech synthesis system [163], as shown in Figure 2.6, consists of two parts, a training part and a synthesis part. The training part extracts both spectrum and excitation parameters from the speech database and models context dependent HMMs taking into account phonetic, linguistic and prosodic contexts. In the synthesis part text to be synthesised is converted to a context dependent label sequence and an utterance HMM is constructed by concatenating context dependent HMMs according to the label sequence. State durations of the utterance HMM are determined based on state duration probability functions. The

[8] `http://htk.eng.cam.ac.uk/`

speech parameter algorithm is then used to generate the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is generated directly from the spectral and excitation parameters using a speech synthesis filter.

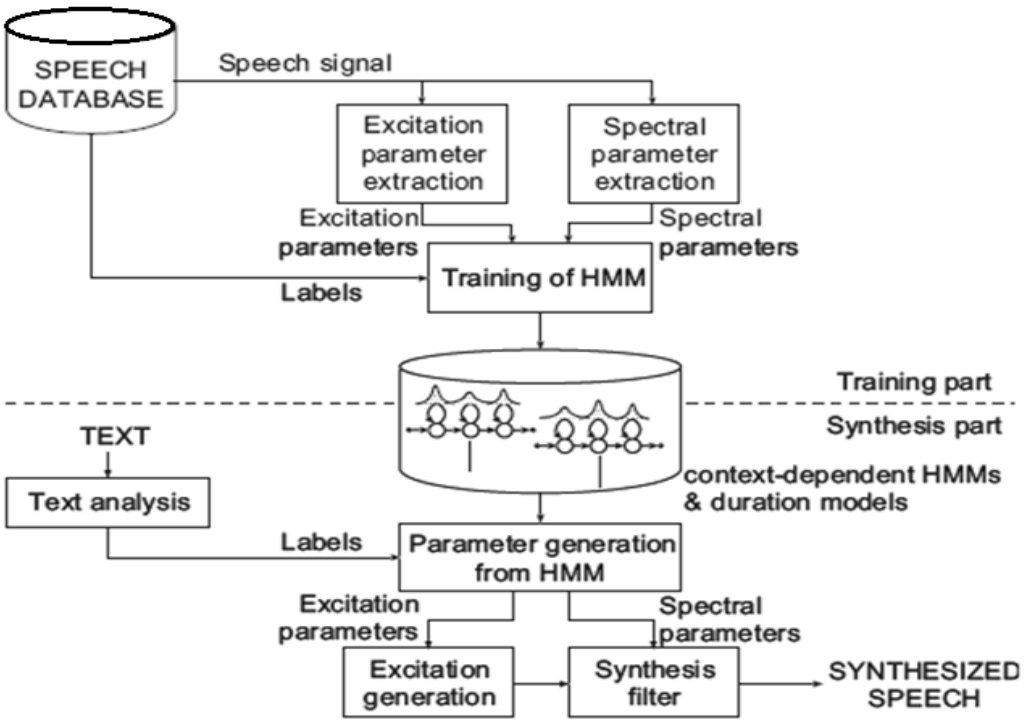


Figure 2.6. Block diagram of a HMM-based speech synthesis system (HTS) (Source:[164])

2.1.5 DNN based Speech Synthesis

Ever since deep neural networks have been successful in ASR, they have been applied to various research areas [52]. DNN based speech synthesis is inspired by the human speech production system which is believed to have layered hierarchical structures in transforming the information from the linguistic level to the waveform level [161]. In the field of statistical parametric speech synthesis, DNNs have been used to find mapping rules between linguistic and acoustic features [161], [69]. Recent studies have shown that DNN-based speech synthesis can produce more natural synthesized speech than the conventional HMM-based speech synthesis [53]. In general, DNNs require a large amount of speech to be adequately trained, especially when they have multiple layers and large number of nodes. With the advent of DNN-base speech synthesis systems, studies of speaker-adaptive speech synthesis

using DNNs have also begun [65].

2.2 Bringing naturalness to synthesised speech

Currently available TTS systems are not characterized by a great amount of flexibility, especially when it comes to variations in voice or speaking style. This is mainly because of the wide spread adoption of unit selection synthesis systems. However, there is a very practical need for different speaking styles in TTS systems. The current ambition in speech synthesis research is to model natural speech at a global level, allowing for changes of speaker characteristics and speaking style. Most researchers in the field of speech synthesis would agree that the inclusion of style and expressive content would increase the naturalness of synthesised speech. Various methods in various areas are adopted to improve the naturalness of synthesised speech. Some such methods are presented in the following sub-sections.

2.2.1 Prosody Modelling

Prosody in linguistics is concerned with those elements of speech that are not individual phonetic segments like vowels and consonants, but are properties of syllables and larger units of speech. The term originated from an ancient Greek word that originally meant a song accompanied by music or the particular tone or accent given to an individual syllable[9]. Prosodic features are therefore also called suprasegmental features since they extend beyond a segment. Prosody contributes to linguistic functions such as intonation, tone, stress, and rhythm. Prosody may reflect various features of the speaker or the utterance. For example, the emotional state of the speaker; or the form of the utterance, i.e., whether the utterance is a statement, question, or command; the presence of irony or sarcasm in the speaker's voice; emphasis, contrast, and focus on different segments; or other elements of language that may not be encoded by grammar or by the choice of vocabulary[10].

At the perceptual level, naturalness of speech is attributed to certain properties of the speech signal that are related to the intonation or changes in pitch, intensity or loudness and duration or segmental length, or in other words to what is commonly referred to as the prosody of speech. A good prosody model should be

[9] https://www.britannica.com/topic/prosody

[10] https://en.wikipedia.org/wiki/Prosody_(linguistics)

able to correctly capture the duration, intonation and intensity patterns of natural speech. Control of prosody plays an important role in bringing naturalness to synthesised speech. Bulyko et al. [19] presented perceptual results that showed that naturalness relative to a target speaking style can be significantly improved through prediction of symbolic labels such as phrase breaks, pitch accents and tones and better generation of F0 contours and phone duration. Reddy and Rao [119] proposed features related to the linguistic constraints represented by the positional, contextual and phonological features and production constraints represented by articulatory features. Neural network models are used to capture the implicit duration, F0 and intensity knowledge using these features. Neural networks are also used by Karjalailen et al. [70] to compute the prosodic control parameters of segmental durations, fundamental frequency and gain for the Finnish language. Liao and Shen [86] describe a prosody model wherein a technique named "Soft Template Mark-up Language (Stem-ML)" is employed to improve the smoothness of intonation which has a crucial influence on the naturalness of synthetic speech. Various models for predicting prosody from text are described by Rao [116]. Similarly various models for modelling the prosodic parameters of duration, intonation and intensity have been proposed for improving the naturalness of synthesised speech by Uslu et al. [145], Lazaridis et al. [82], Gopinath et al. [43], and Rao and Reddy [118]. Violante et al. [148] present a method for improving the perceived naturalness of corpus-based speech synthesisers by removing pronounced pitch peaks in the original recordings, which typically produce discontinuities in the synthesised speech. Annotation of speech corpora for prosody modelling is another area where works ([74], [157]) are being carried out for improved modelling of prosody.

2.2.2 Voice Source (VS) Modelling

In contrast to the articulatory synthesis model, where the main objective is to model the complete speech production system involving the various articulators, VS modelling involves modelling of the VS excitation signal. In the production of voiced speech, the oscillation of the vocal folds interrupts the airflow from the lungs periodically and creates changes in air pressure. The airflow escaping the glottis is converted into a train of pulses which is referred to as the "VS excitation signal". The VS is therefore manipulated by vocal fold vibrations and is responsible for controlling "voice quality" (VoQ), which is the perceptual characteristic of a speaker's voice [22]. It can differ from one speaker to another, and vary within a speaker

from occasion to occasion. The VS is known to be the origin for several essential acoustic cues used in spoken communication, such as fundamental frequency, but it is also related to acoustic cues underlying voice quality, speaking style, and speaker identity, which all contribute to the naturalness and expressivity of speech.

VoQ or VS dynamics play an important role in prosody and intonation, signaling prominence or focus [146], [159], [33] in spoken utterances. Raitio [113], in his PhD thesis, presents two new Glottal Inverse Filtering methods that can be used for improved estimation of the VS signal thereby improving the naturalness of synthesised speech. Chen [22] proposes a new VS model where the source signal is estimated using a codebook search approach and uses it for vowel synthesis. Perceptual listening experiments show that his model provides a better perceptual match to the target voice compared to traditional models. The VS project at UCLA includes works where the authors have proposed a new source model based on observations from high speed imaging of the larynx [129], where a codebook based method is proposed for estimating the open quotient (OQ) which is an important parameter for the VS [130], and where a new source model has been proposed to capture perceptually important source shape aspects [23]. Oliveira [102] describes a source model based on the polynomial model for the glottal flow suggested by Rosenberg that has an exact representation in the frequency domain, and an automated procedure to estimate its parameters from natural speech.

2.2.3 Voice Conversion (VC) for expressive speech

Work on emotional expressivity in synthetic speech goes back to the end of the 1980s. The functionality to add and control expressiveness in synthetic speech has become an important research objective since it leads to improvement in the naturalness of synthetic speech. Traditional methods of generating expressive synthetic voices require carefully designed databases that contain sufficient amount of expressive speech material. But of late VC techniques are gaining popularity in making synthetic speech more expressive by adapting neutral synthesised output to match the expressive target styles. VC techniques aim to convert the speech from one speaker and make it sound like another speaker, at the same time keeping its linguistic contents intact. In short VC modifies speaker-dependent characteristics of the speech signal, such as the spectral and prosodic aspects, in order to modify the perceived speaker identity while keeping the speaker-independent information i.e., the linguistic contents the same [96]. By transforming the overall spectral

characteristics, VC also realises the corresponding voice quality changes implicitly in the spectral conversion function [143].

Turk and Schroder [144], investigate VC and modification techniques to reduce the collection of database and processing efforts, at the same time maintaining acceptable quality and naturalness. It is obvious that incorporating emotional features into synthesised speech would make it more expressive and natural. Wang et al. [149] use the GMM based VC technique, introduced first by Styllianou [135], to carry out spectral conversion for neutral to emotional speech conversion. However it is seen that spectral conversion alone is insufficient for conveying the target emotion, and F0 modelling is equally important. Similarly, Aihara et al. [2] use a GMM based VC method to carry out both spectrum and prosody conversion. Laskar et.al [81] present a comparative analysis of artificial neural networks (ANNs) and Gaussian mixture models (GMMs) for the design of a VC system using line spectral frequencies (LSFs) as feature vectors. Mapping of the pitch contour is carried out using a codebook based model at segmental level while the energy profile of the signal is modified using a fixed scaling factor defined between the source and target speakers at the segmental level. Results of their work indicate that their proposed ANN-based model may be used as an alternative to state-of-the-art GMM-based models used to design a VC system. VC techniques not only convert the organic properties of speech, i.e., VoQ, but also the linguistic cues i.e., regional accents, to some extent, from the source to the target speaker. This makes it ill-suited for accent conversion (AC) where the goal is to capture the native accent of the source speaker and VoQ of the target speaker. The conversion technique converts both the accent as well the VoQ of the source speaker (native) to that of the target speaker (non-native) because of which the native accent cannot be achieved. Aryal and Osuna [8] however illustrate that with a modified training process, VC can yield noticeable reductions in the perceived foreign accent of the target voice. Their approach consists of pairing source and target vectors based not on their ordering within the corpus as is done in a normal VC setup, but on their linguistic similarity. In another work of VC, Anumanchipalli et al. [7] use a new approach to F0 transformation that can capture aspects of speaking style. A statistical phrase accent model is used to represent the F0 contour and a GMM transform is then applied on the TILT [11] parameters of the accents over each metrical foot (a stress group spread over multiple syllables), to predict the possible accent shapes of the target speaker. Rao [115] in his work uses LSFs to represent the vocal tract charac-

[11] `https://www:cs:cmu:edu/~awb/papers/ESCA-int97/node2:html`

teristics, and to develop its associated mapping function. LP residual of the speech signal is viewed as excitation source, and residual samples around the instant of glottal closure are used for mapping. Prosodic parameters at both syllable and phrase levels are used for deriving the mapping function. This approach of mapping and modifying parameters using a pitch synchronous approach used for VC is seen to perform better compared to the author's earlier method which involved mapping vocal tract parameters using block processing. In another application of VC towards improvement of synthesised speech, Jiao et al. [64] propose to use VC to transform synthetic speech towards the original so as to improve its quality using local linear transformation (LLT) combined with temporal decomposition (TD) as the conversion method.

2.2.4 Formant modelling

Another direction that can be considered to improve naturalness of synthetic speech is the correct modelling of formants. One of the oldest works on natural synthesised speech [91] reports on the importance of formant transitions as constituents of speech naturalness. The vocal folds vibrate in a quasi-periodic manner during the production of voiced speech sounds. The resonator system of the vocal tract, which includes the pharyngeal, nasal and oral cavity, modulates the excitation signal so produced. Harmonics near the resonant frequency are boosted, other harmonics are attenuated. These vocal tract resonances are referred to as formants. In other words, regions of frequency space where speech sounds carry a lot of energy are known as "formants". Speakers can change the resonance frequencies by moving their "articulators" (lips, jaws, tongue, soft palate), and thereby changing the dimensions of the resonance cavities in the vocal tract. The information that humans require to distinguish between speech sounds can be represented purely quantitatively by specifying peaks in the amplitude/frequency spectrum and these peaks are the formants. The first two formants (F1, F2) are specially important in determining the quality of vowels. Vowel quality in phonetics refers to the property that makes one vowel sound different from another. It is determined by the position of the tongue, lips, and lower jaw, and the resulting size and shape of the mouth and pharynx.

Accurate tracking and modification of formant trajectories is essential for research purposes, for specific applications such as voice character modification for dialect transformation, speech correction, timbre modification, and for smoothen-

ing of formant trajectories of waveforms generated by concatenative text-to-speech systems. One such method for formant tracking, modification and resynthesis is presented in the works of Bohm and Nemeth [16]. A novel approach to speech synthesis based on waveform segments is proposed by Mizuno et al. [95] which uses a new formant frequency modification algorithm which makes it possible to change formant frequency flexibly and so reproduce the desired speech quality. The proposed method was found to increase significantly the naturalness of speech and to clearly increase speech quality. Another aspect of the importance of formant modelling is that in unit selection synthesis, since units cannot appear in all possible contexts, adjoining units may not have smooth transition in formants at the joining points. It has been observed that smooth changes in frequency are perceived as changes within a single speaker, whereas sudden changes are perceived as being a change in speaker. Formant re-synthesis[23] at the joining points of the units can be used to attain smooth transition of formants thereby improving the naturalness of the synthesised speech. Another application of formant control is illustrated by Lei et al. [83] in their work where they have proposed a framework for formant control and manipulation in an HMM based synthesis system. Results show that the proposed method can control vowels in the synthesised speech by manipulating F1 and F2 without any degradation in the quality of synthesis. Rule based formant synthesis allows control of both glottal and supraglottal parameters of speech, many of which are potentially relevant in the modelling of expressive speech. The first emotionally expressive speech synthesis systems were created based on the commercial formant synthesiser DECtalk [49]. Burkhardt [20] also uses formant synthesis to systematically vary acoustic settings in order to find perceptually optimal values for a number of emotion categories.

2.3 Synthesis of Dialectal Speech

Present speech synthesis systems are mostly restricted to the synthesis of standard varieties of languages. However research on synthesis of dialectal speech is currently gaining popularity and importance. One reason maybe, as recent research suggests, that human beings are able to connect better as listeners, to a speaker and voice who sound like them, which means a person speaking a dialect will be able to relate better to a person talking to him in that dialect. Authentic dialect synthesis requires a high-quality speech corpus of phonetically transcribed dialect utterances. But the collection of such a corpus is a time consuming task due to the non-standard

nature of dialects. The biggest hurdles in the synthesis of dialects in fact, are the unavailability of high quality speech databases, computer readable lexicons and other preprocessed linguistic information necessary for building a TTS. Pucher et al. [108], describe the steps necessary for the construction of a speech synthesis system for dialects; how dialects for which neither a corpus nor a sufficient linguistic description exist are modelled and synthesised by means of Hidden Markov Models (HMMs) on the basis of a comparatively small corpus. In another work [107], Pucher et al. describe a method for selecting an appropriate phone set to be used during the development of a HMM-based TTS System for an undescribed dialect by applying HMM based training and clustering methods. In yet another work by Pucher et al. [110], the authors describe a method which uses transcriptions of original dialect data to map the phones from the source to the target and also uses prosodic transfer of F0 and duration to improve naturalness.

In addition to building a synthesis system for a dialect from scratch, a number of works exist ([109], [9] and [73]) where the TTS for the standard variety of a language is developed as a baseline, and the system uses a limited amount of dialect data and learns from it. Variations in dialectal speech data may exist in the phonetic inventory, in the set of pronunciation rules and in the collection of stress and intonational patterns which help provide structure and syntactic/semantic context to the overall produced utterances. Therefore for cross-dialectal adaptation, not only the unit-selection database , but also those components which assign phonetic realisations to the given text, i.e., the letter-to-sound rules and the pronunciation dictionary or lexicon, may also need alteration. Craig Olinsky in his work[12], describes two types of dialect adaption, phone unit adaption where the phonetic inventory is changed or modified to adapt to the target dialect, and pronunciation adaptation wherein the set of pronunciation rules which dictate how the phonetic units are put together to assign a pronunciation to an orthographic form, is modified. Pucher et al. [109] develop speaker dependent voices for the Tosk and Gheg dialect and adapt models for the Gheg dialect from the Tosk models, Beskow and Gustafson [12] adapt models for Swedish dialectal varieties from standard Swedish while Khaw and Tan [73] adapt a model for Kelantanese Malay from standard Malay.

Grapheme-to-phoneme(G2P) rules which contribute to the pronunciation of a speech synthesis system, play an important role in building a TTS system. Therefore incorporating variations in these rules with respect to the target dialect would

[12]`http://www.cs.cmu.edu/~colinsky/main.htm#dynamic`

help in the synthesis of the dialect. Sitaram [133] in her PhD thesis improves the pronunciation of grapheme-based voices by using better modelling techniques. The author also disambiguates homographs in lexicons in Arabic dialects in order to improve the pronunciation of TTS systems. Her work also shows that phoneme-like features derived using articulatory features may prove useful for improving grapheme-based voices. Likewise Sarma and Talukdar [127] analyse the different Bodo dialects of Assam and form G2P rules by including the variations due to the dialects. These rules can be used while adapting a TTS system to various dialects. It is worth noting that very few works exist on improving the naturalness of synthesised speech in Indian dialects, although some amount of work on the recognition/identification of dialects have been carried out for Indian languages [117], [137] and [132].

2.4 Cross-Dialectal Differences

A regional dialect is a distinct form of a language spoken in a particular geographical area. A very basic difference between languages and dialects is that languages are written and standardized, while dialects are oral, without codified rules. It would be more appropriate if we define dialect as a particular form of use of a language by a group of users, that may be distinct from the form used by another group of users of the same language. In regional dialects we may look for differences in phonemes, morphemes, semantics (vocabulary and expressions), syntax and prosody. In short, variations exist in a few key areas such as phonetic inventory, pronunciation rules and stress and intonation patterns. Despite speech variability, communication is still possible between speakers speaking different dialects of the same language. Sindhi dialects vary from one to other in terms of phonology, morphology and syntax. Shaikh et al. [126] investigate the accents of three dialects of the Sindhi language, Lassi, Laarri and Vicholi, and draw differences on the grounds of vowel duration, pause usage, F0 peak alignment, F0 excursion size and F0 contour shape. Holliday and Kong [54] investigate dialectal variation of voice onset time (VOT) values for the three-way laryngeal stop categories by Seoul, Daegu and Jeju speakers. The authors also compare the relative weights of VOT, F0 and H1-H2 parameters in differentiating the three categories in these three dialects of Korean. Lengeris et al. [85] examine cross dialectal differences on the perception of Greek vowels in standard modern Greek, and two dialectal areas, Crete and Kozani. They found the organisation of perceived vowel space to

be dialect specific, i.e., vowels in all the three Greek varieties are well separated in the perceptual space. Cross-linguistic research on intonation has a long tradition, but most of these studies are monodialectal. Grabe [41] studies the intonational variation in urban dialects of English spoken in the British Isles. Prosodic differences among American English dialects were explored by Hart [50], where prosodic features such as F0 variation, vowel stress and vowel duration were analysed with respect to American English dialects. Results show that spectral change and pitch contour could produce the melodic variation distinctive of the dialects. Jacewicz and Fox [61] provide evidence that vowel inherent spectral change can vary systematically across dialects of the same language. Jacewicz et al. [63] characterise speech tempo of two distinct varieties of American English taking into account both between-speaker and within-speaker variation. Their results show that Wisconsin speakers had faster articulation rates than did North Carolina speakers. In another study [40], the same authors carried out experiments on the vowels of three distinct regional varieties of American English spoken in Western North Carolina, in Central Ohio and in Southern Wisconsin, and their results revealed variation in formant dynamics as a function of the phonetic factors such as vowel emphasis and consonantal context. They concluded that the dialect specific nature and amount of spectral change can be effectively characterised by position and movement in the F1-F2 space, vowel duration, trajectory length, and spectral rate of change. Most of the works cited above report vowel formant space and prosodic features such as pitch, intensity and duration as cross dialectal variations. However cross dialectal variations may also exist in terms of pharyngealisation, phonation, diphthongisation and tone.

Pharyngealisation: Pharyngealisation is a secondary articulation of consonants or vowels by which the pharynx or epiglottis is constricted during the articulation of the sound. Qutaish [112] examines the pharyngealisation spread in one of the Yemeni Arabic dialects, namely, Ibbi Arabic (IA). He investigates how pharyngealised sounds spread their acoustic features onto the neighboring vowels and change their default features. Results show that this pharyngealisation spread is gradient in nature. In another work on Arabic varieties, Tamimi [4] carries out an analysis of vowels in Moroccan and Jordanian Arabic varieties, results of which establish spectral tilt as an acoustic correlate of pharyngealisation.

Phonation: Phonation refers to the production or utterance of speech sounds, or simply put phonation is vocalization. It may be defined as the process of producing vocal sound by the vibration of the vocal folds that is in turn modified by

the resonance of the vocal tract[13]. Cross-linguistic phonetic studies have yielded several insights into the possible states of the glottis. The glottis can be controlled to produce speech sounds with not only regular voicing vibrations at a range of different pitches, but also speech sounds in a variety of phonation types such as soft, harsh, creaky and breathy. One of the outstanding problems in the history of Khmer (Cambodian) phonology concerns the existence and development of 'register' or 'phonation types' (i.e., breathy versus clear voice) in the language. One dialect of Khmer that appears to have kept the original 'breathy' versus 'clear' distinction in VoQ is the one spoken in Chanthaburi Province, Thailand. Wayland and Jongman [152], carry out a detailed survey of the acoustic and perceptual correlates of breathy and clear phonation in the vowels of Khmer, results of which suggest that the earlier breathy and clear phonation distinction in Khmer is preserved among female speakers of Chanthaburi Khmer, but this distinction may be disappearing or have become a tense versus lax distinction among male speakers.

Tone: A distinguishing feature across all Chinese dialects is tone or the difference in pitch. Tonal contrasts vary in number between dialects, for example northern dialects tend to have fewer distinctions than southern ones. In the two most commonly spoken dialects of Chinese, i.e., Mandarin and Cantonese, Mandarin is considered to be the standard variety. It has four tones while Cantonese has six tones and may have upto nine tones. Shanghainese or the Shanghai dialect is considered to be one of the Wu varieties and it has five phonetically distinguishable tones for single syllables said in isolation. The Wu varieties may be distinguished by their retention of voiced or murmured obstruent initials (stops, affricates and fricatives). Tone functions as a distinctive feature in the eastern half of Korea consisting of Gyeongsang, Hamgyong and the eastern part of Gangwon. Tonal distinction is not present in the western half where vowel length functions as a distinctive feature [18].

Diphthongisation: Diphthongisation refers to the change of a monophthong to a diphthong. Okati et al. [100] present a study on the phenomenon of 'Diphthongisation' in the different varieties of Iranian Balochi dialects spoken in Sistan, Saravan, Khash, Iranshahr and Chabahar regions of South East Iran. The study revealed that in the Khash dialect, diphthongisation of vowels /e/ and /o/ is predominant and therefore these two vowels should be represented as the diphthongs /ie/ and /ue/ in the vowel inventory of the Khash dialect. The Iranshahr and Chabahar dialects show the second and third highest degrees of diphthongisation

[13] `https://www.wordnik.com/words/phonation`

among the five dialects under study. Data in these two dialects indicate a tendency towards diphthongisation rather than a shift to predominantly diphthongised productions. The Sistan and Saravan dialects show only sporadic tendencies toward diphthongisation. Likewise the extent of diphthongisation varies in the different dialects of American English. Magen [89], carried out experiments on the vowels of the Rhode Island dialect, results of which show that in addition to the usual mid-height /eɪ/ and /oʊ/, diphthongisation of /ɪ/ and /ɔ/ occur, and for diphthongised vowels the perceptually dominant portion of the vowel is variable. Diphthongisation can also be used as a cue for dialect identification as is shown for the British English dialects [37]. The degree to which a given speaker of British English diphthongises her/his vowels has been known for decades to be a good indicator of the speaker's dialectal origin.

2.5 Limitations of current approaches towards the Generation of Dialectal Speech

From the survey, it is observed that currently available TTS systems are not characterised by a great amount of flexibility, especially when it comes to variations in voice or speaking style. This is mainly because of the wide spread use of unit selection speech synthesis systems. Therefore inclusion of naturalness to synthesised speech is the main concern of speech synthesis researchers today. Perceptually, naturalness may be attributed to both spectral and prosodic features of speech. VC techniques are popular techniques whereby both spectral and prosodic features are transformed from source speech to target speech. However these techniques are limited to conversion of speaking styles only, for example, for converting whispered to natural speech, from sad to happy speech, from child to adult speech and so on. Furthermore VC systems can be trained with small amounts of target speech data. This feature also makes it suitable for use dialectal speech conversion. However whether such techniques can be used for converting one dialectal variety of a language to another is yet to be explored. Dialectal varieties of a language may be different from each other in various aspects such as the vocabulary, grammar, prosody as well as pronunciation. Prosody is reflected in features such as pitch, intensity and duration while various spectral features reflect the pronunciation of various speech sounds. Therefore there is a scope for using VC techniques for converting speech from one dialect to another. It is also observed that most works on

dialectal speech generation concentrates mainly on the prosody of speech since at the perceptual level it is observed that prosodic aspects have the greatest impact. However finer details such as how a particular speech sound (a vowel or a consonant) is pronounced in the different varieties of a language can also contribute to the naturalness of synthesised dialectal speech. Our survey shows that such an effort has been made by Sitaram [133] in her works where dialectal variations are incorporated in the G2P rules which contribute to the pronunciation of a speech synthesis system. Pucher et al. [107] also uses a method of selecting a phoneset suitable for a dialect, to be used during the building procedure of a HMM-based TTS for a dialect. However both these methods are used to incorporate necessary changes during the building of the TTS system itself and will not be of much help if the existing TTS of a particular variety is used to generate speech in another variety. It is observed that most works on dialectal speech generation are limited to the design or adaptation of TTS systems for generating dialectal speech, but very few works have been reported for manipulating or modifying synthesised speech for the same purpose. Our survey also shows that most of the dialect-distinctive features are used mainly for dialect identification purposes and there is a limitation regarding the use of such features for the synthesis of dialectal speech.

A point to note is that differences between dialect pairs of different languages may vary. For example, perceptual vowel space is distinctive to the Crete and Kozani dialects of Greece, the number of tonal contrasts is distinctive to the Mandarin and Cantonese dialects of China, while extent of diphthongisation varies in the different dialects of American English. Therefore identifying the differences specific to a pair of dialectal varieties of a language is one distinct task which is essential for the larger goal of incorporating dialectal features into synthesised speech. We therefore carry out an extensive analysis of features in the two varieties of Assamese that we have considered for our study, in the following chapter, i.e., Chapter 3.