Chapter 1

Introduction

A Text-to-Speech (TTS) system is a computer based system capable of reading out loud any text provided to it as input. Today talking machines are nothing new since they have been in existence since the middle of the 18th century. However all talking machines are not speech synthesisers or Text-to-Speech systems. There is a difference between computers generating speech and computers playing recorded speech; the first is a TTS system, not the second. Systems that simply concatenate isolated words or parts of sentences, denoted as Interactive Voice Response Systems (IVRS), are only applicable when a limited vocabulary is required. For unlimited vocabulary this is not possible since it is impossible to record and store all the words, phrases or sentences of a language. TTS may be defined as the automatic production of speech, through a grapheme-to-phoneme (G2P) transcription of the sentences that are required to be generated. A good example of a speech synthesiser is the *eSpeak*[1] software which is a free open source speech synthesiser for English, German, French, Kannada, Tamil and many other languages.

Essentially there are three stages involved in a typical TTS system, text to words, words (in terms of graphemes) to phonemes, and phonemes to sound. The initial stage, which is generally called pre-processing or normalisation, is all about narrowing down the many different ways a person could read a piece of text into the one that is the most appropriate. It involves the conversion of raw text containing symbols, abbreviations and numbers, to words. The next stage involves breaking written words into their graphemes and then generating phonemes that correspond to them, using a set of simple rules. And finally these phonemes need to be converted to sound that we can hear either by concatenating phoneme sounds recorded from humans (referred to as concatenative synthesis), or by making the

[1] `http://espeak.sourceforge.net`

computer generate the phonemes by generating basic sound frequencies (referred to as formant synthesis) or by making the computer mimic the mechanism of the human voice (referred to as articulatory synthesis). The TTS system may also be said to comprise two basic modules, the Natural Language Processing (NLP) module and the Digital Signal Processing (DSP) module as depicted in Figure 1.1. The first two stages i.e., text to words and words to phonemes, make up the NLP module while the third stage, i.e., phonemes to sound, make up the signal processing module. Technically speaking, the NLP module consists of the text analyser, phonetisation and prosody generation modules while the DSP module deals with the actual machine 'pronunciation' of words, phrases and sentences, analogous to human speech articulation [103].

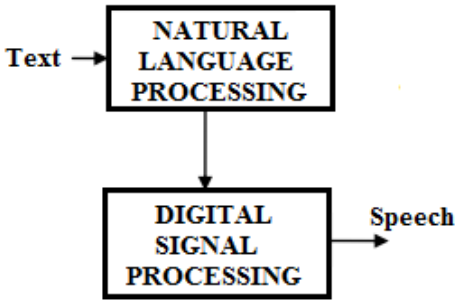

Figure 1.1. A simplified diagram of a Text-to-Speech Synthesiser[2]

The term 'Speech synthesis' refers to the artificial production of human speech. Therefore a TTS system is a speech synthesiser since it produces artificial human speech. Speech synthesis plays an important part in the communication between man and machine. Professor Stephen Hawking, famous for talking in a computerised voice, is an example of being benefitted by the technique of Speech Synthesis. Potential applications of high quality TTS Systems are indeed numerous. Whether it is the Telecommunications service where TTS systems make it possible to access textual information over the telephone or in the field of language education where it can be coupled with a Computer Aided Learning system to provide a helpful tool to learn a new language or as a tool for the visually impaired or voice handicaps, the uses of speech synthesis are profound and many. Today synthesised speech is no longer a luxury. There are automatic techniques to build good quality TTS systems quickly for languages, provided these languages have sufficient data and linguistic resources. Therefore most speech synthesis systems are limited to only

[2] http://www.iaeng.org/publication/WCE2014/WCE2014_pp582-584.pdf

the standard varieties of a language with sufficient data and linguistic resources. Current speech synthesis may have progressed to a very high level, but there are still many inadequacies which are yet to be overcome. One major concern is inadequate prosody or the rhythm of speech at the suprasegmental level while at the segmental level, there is an inadequacy regarding the modelling of dialectal variation. Since the scope of this thesis is limited to the second issue, we note down several reasons as to why dialectal speech generation is important.

Synthesis of dialectal speech is important for use in various local language applications for the visually challenged, IVRS, computer aided learning for rural kiosks etc. Many of the areas, especially in underdeveloped countries like India, which could be highly benefited from community-focused information technology (IT) resource development, have a high rate of illiteracy among their population. Speech-based systems are the most obvious and natural mechanism for such users to connect with computers. Recent studies [77] have suggested, that human beings connect better as listeners to a speaker and voice who sound like them. They find it easier to listen to and understand what is said to them. They also find it easier to assign emotions and judge factors such as authority, honesty, and even intelligibility. Therefore TTS systems capable of generating speech with regional accents are highly desirable. We are thriving for a Digital India. But in a country like India where approximately 70% of the population resides in rural areas, such a campaign can be successful only when IT enabled services are made accessible to the entire population. This makes the implementation of digital services in local languages and dialects, highly important. Accurate regional accents can also be a very sensitive issue, therefore in Telephone Information Systems it might be useful to have the system respond to the caller in the caller's own regional accent. Dialectal speech generation is also considered to be one of the most important aspects of speech synthesis for languages such as Chinese, Vietnamese, Korean, German etc having a large number of dialectal variants.

With the unavailability of high quality speech databases, computer-readable lexicons, and other pre-processed linguistic information that is available for, for example, standard dialects of languages such as French or German, it is costly as well as difficult to build TTS systems for most dialects. In India there are reportedly 22 major languages, written in 13 different scripts, with over 720 dialects[3]. The Linguistic Survey of India (1903-1923), had identified 179 Indian languages and

[3]https://www.justlanded.com/english/India/India-Guide/Language/Languages-in-India

544 dialects[4]. Various projects are currently being carried out for the development of TTS systems for Indian languages and one such project involves the development of TTS systems for 13 of these languages namely Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati, Odia, Assamese, Manipuri, Kannada, Bodo and Rajasthani, by the Indian Language Technology Proliferation and Deployment Centre (ILTPDC)[5]. Likewise TTS for some of these languages are being built by researchers in various academic institutions although resources for these languages are very limited. Unfortunately TTS systems for generating dialectal variations of these languages are almost non-existent. Reason being, resources for dialects are even scarcer. However once a baseline TTS, i.e. TTS of the *standard language* (also referred to as the *standard variety* in this thesis) is built, speech in its dialectal variants may be generated by making the system learn from small amounts of dialectal speech data. This is the basic idea governing or underlying our research work.

1.1 Motivation for current topic of research

Rapid advancements in speech technology have resulted in its widespread use by consumers, especially in mobile applications such as Spoken Dialogue Systems (SDS) like Siri for the iPhone and Voice Search on Android phones. This progress in technology has led to highly intelligible and natural sounding synthesised speech for high resource languages such as English, French, German, Cantonese, Japanese, Mandarin, Italian, Spanish, and Korean. Speech researchers have extensively studied these languages and have built pronunciation rules, dictionaries, part-of-speech taggers (POS), and language models. They have collected and annotated huge amounts of high quality data from professional speakers in order to provide such a high quality of synthesis. However, there are thousands of other languages in the world, many of which are spoken by millions of people, which do not have such resources. There are others which are primarily spoken languages, or languages with large non-literate populations, which could benefit from speech-based systems. Such low resource languages like Telugu, Tok Pisin, Tamil, Vietnamese, Tagalog, Cebuano, Pashto, Hindi and Assamese to name a few, have few NLP and data resources available to TTS researchers and no carefully recorded and

[4] `https://mhrd.gov.in/sites/upload_files/mhrd/files/upload_document/languagebr.pdf`

[5] `http://tdil-dc.in/index.php?option=comvertical&parentid=85&lang-en`

annotated corpora which is the basic requirement for conventional TTS systems. Therefore, speakers of these languages do not have the same access to speech related technologies that allow communication across language barriers, such as SDS or speech-to-speech(S2S) translation, of which TTS is a crucial component.

Building a TTS from scratch is a time consuming and expensive task. State-of-the-art technologies show that of the various methods used for building a TTS, the Unit selection method of speech synthesis gives the most natural sounding speech output. However the drawback is that such systems require a huge corpus containing all phones in all possible contexts preferably by the same speaker. For under resourced languages, more so for dialects of under resourced languages, gathering speech data to build a large enough corpus is a huge challenge, mainly because of the absence of a written form, formal language rules and a transcription system. This would mean that building a TTS for a dialectal variety, compared to that of the standard variety of a language, will be even more difficult. The better option would be to use an existing TTS of the standard variety and adapt it to the target dialect. However this would require an in-depth knowledge of building a TTS. Secondly Voice Conversion (VC) techniques are popular techniques used for the conversion of speaking styles, for example, for converting whispered to natural speech, from sad to happy speech, from child to adult speech and so on. A VC system can be trained using limited training data with minimal annotation, making it suitable for use in the conversion of dialects. The scope of using this technique to convert from one variety of a language to another has not been explored yet. Thirdly, most works on dialectal speech synthesis concentrates mainly on the prosodical aspects of speech. There are other more subtle aspects, such as how different speech sounds are pronounced in a dialect, which may also contribute to the naturalness of synthesised dialectal speech. Although some amount of work [117], [137] and [132], on the recognition/identification of dialects have been carried out for Indian languages, to the best of our knowledge, work on improving the naturalness of synthesised speech in Indian dialects is non-existent, making it all the more challenging.

Therefore a speech modification module can be used as a post processing module, after the speech synthesis module (which is an existing TTS system), where variations with respect to the dialect concerned can be incorporated into the synthesised speech to make it sound more like the target dialect. Such a module would be highly desirable if it can be trained using limited dialectal speech data. The main objective of our work is to contribute in some way to the generation of di-

alectal speech. In our work we have used the TTS system built for the standard variety of a language to generate dialectal speech. In other words we use an existing method for synthesising speech in a standard form, and attempt to bring the subtle changes to incorporate the effects of another dialectal variant. This is made possible by using a post processing module to incorporate dialect-distinctive features into the speech generated by the standard TTS. Most work on improving the naturalness of synthesised speech, concentrate on prosody since at the perceptual level prosodic aspects are more prominent. However we have concentrated on the segmental features which mainly relate to how a particular sound is pronounced and though not very prominent they can help render the finer details of naturalness to synthesised speech. Two approaches are taken in this work. In one, VC techniques are used to develop a mapping function to map the spectral and prosodic features from one variety of a language to another. In the other approach, the vowel/diphthong formant space of one variety is transformed to that of the other variety. Results from both the approaches are encouraging and suggest their future applicability in the generation of dialectal speech. Subjective evaluation of results using the Mean Opinion Score (MOS) shows that the output of the post processing module using VC, with an MOS of 3.2, outperforms the standard TTS output having an MOS of 2.3. Likewise the output of the post processing module using Formant Transformation, results in bringing the vowel formant space of the source closer to that of the target. The scope of this work is limited to the Assamese language which is one of the major languages of India and to other languages whose dialects also exhibit similar distinctive characteristics.

1.2 Thesis Objective

The objective of our work is ***to develop an approach to incorporate dialectal features into synthesised speech of the standard variety so that the synthesised speech, sounds natural with respect to the dialect considered and is distinct from the same speech in some other dialect.*** The two varieties of Assamese considered for our study are the AIR variety which is a form of the central group of Assamese, generally spoken by the readers of Assamese news of All India Radio, and the Nalbaria variety (NAL) which is a form of Assamese spoken by the people in around the district of Nalbari in Assam. The AIR variety is considered the standard variety. Speech in the AIR variety is generated from a TTS built for the AIR variety with text in the Nalbaria variety. This speech lacks

naturalness with respect to the target dialect. The naturalness of dialectal speech synthesised by a TTS built for the standard variety of the language can be improved by using a post processing module for incorporating dialect-distinctive features into the synthesised speech. The main objective of this thesis can be broken down into the following sub-objectives:

(i) To carry out an extensive experimental analysis of features, both spectral and prosodic, in the AIR and NAL varieties of Assamese, in order to find out features distinctive to these dialectal variants.

(ii) To use VC techniques to convert spectral and prosodic features from the source (synthesised speech using a TTS for the standard variety) to the target (Nalbaria speech).

(iii) To develop a GMM-based approach for Formant Transformation (FT) in order to transform the vowel formant space from one variety of Assamese (AIR) to another (NAL).

The block diagram of the proposed model to be used as the post processing module is presented in Figure 1.2. Our work concentrates on the design and development of the two modules, the TTS module and the module for incorporating dialectal features using VC/FT techniques and all related experimental analytics.

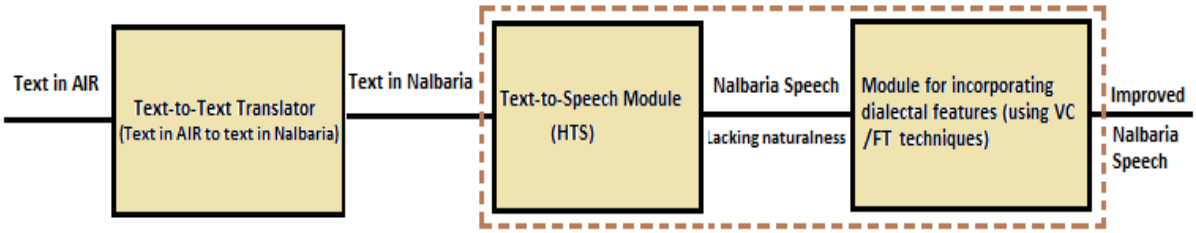

Figure 1.2. Block Diagram of the proposed Model

1.3 Thesis Organisation

The thesis is organised as follows:

Chapter 2 carries out an extensive review of literature relating to the synthesis of dialectal speech. It gives a brief introduction to the different methods used for generating synthetic speech, various existing design methods for a TTS system and various techniques used to incorporate naturalness to synthesised speech. It also gives a brief overview of various methods currently being used for generating dialectal speech. In addition to this it also describes briefly various differences that may exist among the dialects of a language. In other words, it also brings to light various limitations existing in the current methods for generating dialectal speech and briefly states how the approaches used in this research work can help in achieving the thesis objective.

Chapter 3 presents a brief introduction to the Assamese language and its dialectal varieties. It then presents a detailed analysis of features, both spectral and prosodic, of two varieties of the Assamese language. Various features such as Voice Onset Time (VOT), Vowel Formants, Formant Trajectories of Diphthongs, Vowel Duration, Spectral Tilt, Cepstral Coefficients and Pitch Contours are analysed and results presented. Based on experimental results, some of these dialect-distinctive features are chosen for transformation of one variety to another.

Chapter 4 starts with an introduction to VC and its applications. It elaborates on the scope of applying VC techniques to incorporate naturalness, with respect to the concerned dialect, into synthesised speech. Three popular techniques for developing mapping functions to be used in VC, are experimented with data from the concerned varieties of Assamese and the best results based on objective evaluation are chosen for synthesis. Subjective evaluation is carried out to check the effects of VC on the resynthesised utterances.

Chapter 5 describes a Gaussian Mixture Model (GMM) based approach for transforming the formants of the vowels/diphthongs of one Assamese variety to another, thereby making the vowels/diphthongs of one variety sound more like the other. Three different transformations are carried out. In the first method, a single GMM is used to model the entire formant data. The second method uses four GMMs to model formants at four equidistant temporal points of the vowel/diphthong, and the third uses independent GMM models for each of the vowel or diphthong. Results of the transformation are evaluated objectively and subjective evaluation is also carried out after test words are resynthesised with transformed formants using Klattworks.

Chapter 6 summarizes the work with important conclusions and directions

for future work. One important direction can be to extend the feature analysis to consonants as well. The place and manner of articulation of consonants may prove to be another feature distinctive to the dialects which may be important for dialectal speech synthesis. Another interesting direction can be to extend this work to the conversion of 'registers' which are known to be situational variations in speech.

1.4 Thesis Statement

The naturalness of dialectal speech synthesised by a TTS built for the standard variety of the language can be improved by incorporating dialect-distinctive features into the synthesised speech.

1.5 Thesis Contributions

The main contributions of the dissertation can be divided into three parts. The following subsections briefly outline the major contributions of the dissertation.

Analysis of dialect-distinctive features In a quest for dialect-distinctive features, various features such as VOT, Vowel Formants, Formant Trajectories of Diphthongs, Vowel Duration, Spectral Tilt, Cepstral Coefficients and Pitch Contours, in the two chosen varieties of Assamese, are analysed and compared. This analysis is the first and most important step towards our goal of incorporating dialect-distinctive features into synthesised speech.

Naturalness using VC techniques VC techniques are used to convert spectral and prosodic features from the source speech data to the target speech data. The first method uses mapping codebooks to carry out the conversion, the second uses a GMM based mapping function and in the third method artificial neural networks (ANN) are trained on parallel source and target data to design the mapping function. Both objective and subjective evaluations are carried out to find out the effects of using VC on dialectal speech synthesised from a TTS for the standard variety of a language.

Naturalness using Formant Transformation One dialect-distinctive feature with respect to the concerned dialectal varieties of Assamese, as pointed out in the next chapter, i.e., Chapter 2, is the vowel formant space (VFS). GMM

based mapping functions are therefore developed to transform the formants of the vowels/diphthongs of one variety to another to make the vowels/diphthongs of one variety sound more like their counterparts in the other variety, thereby contributing to the overall naturalness of synthesised dialectal speech.

In addition to these there are three smaller contributions which are listed below:

Annotated corpora of two varieties of the Assamese language
Speech data of approximately one hour for the standard variety and half hour for the dialectal variety, is collected, cleaned and annotated at the phrase, word, syllable and phoneme level.

A Text To Speech system for the standard variety of Assamese
A HMM based TTS for the standard variety of Assamese is built using the HTS toolkit. The TTS is trained with approximately one hour of speech in the standard variety of Assamese.

Development of a Matlab GUI- based tool named VoiCon
VoiCon presents the researcher with an easy to use interface in order to apply the VC process to various applications. It also allows the researcher to analyze his results both objectively and graphically.