Chapter 6

Conclusions and Future Research Directions

6.1 Research Contributions

Speech synthesis is an area of research which is highly popular with researchers. However it is almost always restricted to the synthesis of speech in the standard varieties of a language. This restriction imposes a strong limitation on various application scenarios and therefore the synthesis of dialects and sociolects enables an important expansion. According to Professor David Crystal, an eminent language expert, a dialect is the use of a distinctive vocabulary and grammar that relates to the identity of a particular group and a person belonging to one dialectal group would not want to be associated in terms of identity, to another dialectal group. In his words "We are talking identity not intelligibility", and therefore he further stresses that synthesising dialects is an obvious next step in the development of artificial speech. However synthesis of dialectal speech has a lot of limitations mainly due to insufficient data and lack of linguistic resources. Today a lot of researchers are engaged in finding new approaches to synthesising dialectal speech using limited data. One such way is to use an existing TTS built for the standard variety of the language and then adapt it to the target dialect with a small amount of target dialectal speech data. Another way would be to analyse the target speech data and find out features distinctive to that dialect so that a post processing module can be developed to incorporate these features into the speech generated by the existing TTS of the standard variety. This thesis takes the second approach and makes three contributions related to the development of the post processing module. The

main contributions made in this thesis are summarized as below.

1. In a quest for dialect-distinctive features, various features such as Voice Onset Time, Vowel Formants, Formant Trajectories of Diphthongs, Vowel Duration, Spectral Tilt, Cepstral Coefficients and Pitch Contours, in the two chosen varieties of Assamese, i.e., AIR and NAL, are experimentally analysed and compared. Features for analysis, are selected based on the observations of a preliminary perceptual study on speech in the two selected varieties of Assamese. Such an analysis, of features covering different aspects of a dialect such as speaking style, tempo and pronunciation, is to the best of our knowledge, non-existent in the literature, especially for dialects of Indian languages. The analysis helps to understand how close or distant the two varieties are from each other and from the synthesis point of view, how manipulating the features of one variety makes it sound closer to the other variety. Results show that the two varieties are significantly different from each other with respect to vowel formant space, extent of diphthongisation and vowel duration. VOT and Cepstral Coefficients are other features distinctive to the two varieties. Another outcome of this analysis is that this set of dialect-distinctive features may be explored further and prove to be useful for works in the recognition/identification of Assamese dialects.

2. The second contribution is an attempt to use VC techniques to bring naturalness to synthesised dialectal speech. VC techniques based on VQ, GMM and ANN, are used to convert spectral (MCEP) and prosodic (F0) features from source data, i.e., speech in NAL synthesised by an AIR TTS, to target data, i.e. speech in NAL. In addition to this, Pitch-Intensity-Duration (PID) values in the test utterances are manipulated to match those in the target utterances. Both objective and subjective evaluation results show that conventional VC involving the conversion of MCEP and F0, helps in making the utterances more natural with respect to the target variety, but notable improvements resulted only when prosodic features such as pitch, intensity and duration, are properly manipulated. For a set of test utterances, results show that there is a significant lowering, approximately 41.7% in MCD and 14.9% in RMSE values after the conversion. At the same time using the VC module after the TTS raises the MOS from 2.3 to 3.2.

3. The third contribution is the development of a GMM based approach to transform the vowel formant space from one variety of Assamese to another,

in order to make the vowels of Nalbaria synthesised from a standard HTS sound more like their counterparts in Nalbaria. Objective evaluation results indicated that in most cases the transformation was successful in bringing the test formants closer to the target formants. Approximately 42.3% lowering of RMSE values is achieved when using a single GMM to model entire training formant data, and approximately 51.8% when using separate GMMs to model the formant data of each vowel/diphthong segment. Furthermore, listening tests performed on the resynthesised vowels/diphthongs using modified formants also indicate closeness to the target dialect.

4. In addition to the above three contributions, the thesis also has three other useful outcomes-

   (a) Development of a HMM based TTS system for the standard variety of the Assamese language. The TTS has been trained with approximately 50 minutes of speech in the AIR variety of Assamese. The quality of speech is of moderate quality however the error ratio is much smaller as compared to the HTS for Assamese developed at IIT-G.

   (b) Speech data of approximately sixty minutes for the standard variety and thirty minutes for the dialectal variety, is collected, cleaned and annotated at the phrase, word, syllable and phoneme level. In addition to this we have also developed a speech corpus containing 250 short sentences from 3 Nalbaria speakers and another containing the same set of 250 sentences in AIR from 3 AIR speakers, resulting in a corpus size of 1500 sentences. This corpus is also annotated at phrase, word, syllable and phoneme level which is a highly manual effort intensive exercise.

   (c) Development of a Matlab GUI- based tool named VoiCon. VoiCon enables a researcher to carry out the various steps required for the VC process in a very easy manner without going into the intricasies of the methods used, and at the same time also allows the researcher to analyze his results both objectively and graphically.

6.2 Future Directions

Though the extent of experimentations in this work is limited, but the work is in a way, pioneering for Indian languages' dialects. This thesis clearly outlines

directions for achieving dialectal effects in synthetic speech. Future work may be aimed at addressing some of the shortcomings of the current work and extending the ideas explored here to address several other relevant phenomena. Some of these future work/directions are outlined as follows:

1. The first contribution is an analysis of dialect-distinctive features in the selected two varieties of Assamese, presented in Chapter 3. To the best of our knowledge this is the first time such a work has been reported for Assamese. Since the corpus had to be built from scratch and also due to the lack of resources such as tools for automatic segmentation, annotation, etc for Assamese, there was a limitation to the size of speech data collected and annotated. The number of speakers from whom data is collected is also very limited. Therefore the analysis can be carried out in a more extensive manner for conclusive results by increasing the number of speakers, including female speakers, as well as increasing size of sample data. Analysis results of some of the features such as VOT, Cepstra and Pitch, also do not present a very clear picture and therefore these features can be explored further. Furthermore most of the analysis is limited to the vowels only. This study can be extended to consonants as well since consonants also play a major role in the perception of speech. The place and manner of articulation of consonants may prove to be another feature distinctive to the dialects. Another future work in this direction would be to extend the analysis to features such as Jitter, Shimmer, HNR, etc which are also known to affect the perception of speech sounds. It will also be interesting to study the pause statistics in speech, including features such as pause duration, pause positions, pause context, etc., with respect to different Assamese dialects. Speaking rate and speech rhythm may be other interesting prosodic features which may be explored as well. Another feature that can be explored is glottalisation. Nalbaria speakers are of the opinion that there are many sounds that are produced by constricting the flow of air in the glottis. This is what is referred to as a glottal stop. This is very likely to be another distinctive feature.

2. The second contribution elaborated in Chapter 4, explores the scope of using Voice Conversion to incorporate naturalness to synthesised speech. The conversion process reported here is limited to (i) three basic techniques of developing mapping functions, i.e., VQ-based, GMM-based and ANN-based, and (ii) only two features, i.e., mel-cepstral coefficients and global pitch values. For further improvement in the naturalness of speech with respect to

the dialect concerned, local variations in the contours of pitch and intensity as well as segment durations, need to be converted as well. Therefore, one direction for future work can be to explore techniques which would enable the development of a more appropriate mapping function and techniques which would also lead to conversion of local variations in PID from source to target.

3. As elaborated in Chapter 5, the third contribution is the development of a GMM based process for transforming the formants of vowels as well as diphthongs in the source variety to the target variety. The transformation accuracy can be improved further by including contextual information in the training process since formants of speech segments are influenced by the formants of segments preceding and succeeding it. In most cases the third approach in the process, where a separate mapping function is built for each of the vowels/diphthongs, outperforms the other two. The disadvantage of this approach is the additional requirement of label files (containing vowel/diphthong identity and duration) for each of the utterances. Therefore automatic vowel/diphthong labelling may be considered as a future task. Furthermore, we have transformed the formants of vowels/diphthongs only and have regenerated the formant contours for only the vowels/diphthongs. Placing these formant contours in the formant contours of the complete utterance would lead to discontinuities which has to be smoothed. This is another direction which may be explored further.

4. Recent studies [156], [162], [111] using deep neural networks (DNNs) as acoustic models for statistical parametric speech synthesis (SPSS), illustrating adaptability of DNNs in changing speaker identity and speaking styles, have shown promising results. Therefore, in addition to the above mentioned future directions, another task can be to adapt the TTS for the standard variety of Assamese to Nalbaria, by using the speech corpus built for Nalbaria, probably using DNNs, study the limitations and explore methods to overcome these limitations. Furthermore, a TTS (using both Unit-Selection and HMM methods) can also be developed for the Nalbaria variety and the performance of the proposed VC-based and Formant Transformation-based systems w.r.t these reference systems may be evaluated. A detailed analysis of the phonemes in the major dialects of Assamese and, if required, preparation of separate phonesets for the different varieties may also be taken up as a future task.

5. The design and development of the Text-to-Text translation module to trans-

late text in one variety to text in some other variety, is another direction that needs immediate attention. This would result in the automatic translation of dialects, both at the text level and at the speech level.

6. An important direction would be to extend this work to the other dialectal varieties of Assamese and build a module which, given text in one dialect, would be able to produce speech in some other chosen dialect. This in fact, is the larger goal of this thesis.

7. An interesting future direction can be to extend this work for the conversion of 'registers' within a language. A 'register' may be defined as a variety of a language used in a particular social setting[1] such as a baby talk register, a joking register, a formal register. Registers are situational and therefore may change within a conversation itself that too for the same speaker. Therefore conversion of registers in speech will also add to the naturalness of synthesised speech.

[1] https://en.wikipedia.org/wiki/Variety_(linguistics)