Abstract

Speech synthesis systems have been in existence since the eighteenth century [136]. A major goal for researchers in the field of speech synthesis, has been to improve the intelligibility and naturalness of synthesized speech to make it sound as close to human speech as possible. Advancements in speech technology has led to very natural and intelligible synthesized speech for high resource languages like English, French, German, Japanese and Korean, having ready availability of proper annotated corpora, pronunciation rules, language models and dictionaries. However work on the synthesis of low resource languages, especially dialects which are even poorer in resources, have been limited and few. In order to keep alive the dying dialects it is necessary that due importance and efforts be also given to research on dialectal speech synthesis inspite of the lack of resources.

One major hurdle to the design of speech synthesis systems for dialects is the unavailability of annotated corpus. However a TTS for the standard language, i.e., the official form of a language, may be easily built since it will be comparatively easier to collect speech data in the standard language than in the dialect. Starting from the standard TTS, two approaches may be taken to produce dialectal speech. With a limited amount of dialectal data, firstly this standard TTS may be adapted to the dialect and secondly, the dialectal data may be used to learn techniques to manipulate the output of the standard TTS in order to produce speech in the required dialect. The second approach implies that starting with a pair of languages, the standard language and its dialect, the features of one may be manipulated to make it sound more like the other. With this purpose in mind, the current research work has been directed towards finding methods to incorporate naturalness, with respect to the dialect concerned, to speech synthesized from a TTS of the standard form of the language.

The first step towards this goal is to analyse the dialectal varieties concerned, in order to find dialect-distinctive features which can be manipulated to bring about the naturalness. In the second step, two approaches are used to modify the chosen features. The first approach uses voice conversion techniques to convert the spectral as well as prosodic features of the source to match that of the target. In our case the source is dialectal speech generated from a TTS of the standard form of th language while the target is speech in the target dialect. The second uses the concept of formant transformation to transform the vowel formant space of the source variety to that of the target variety. Results from both the approaches are encouraging and suggest immense scope in synthesis of dialectal speech. However synthetic output using the first approach is slightly noisy since voice conversion techniques are used on synthetic not real speech. Likewise in the second approach too there are limitations as to the quality of synthesized speech using modified formants.

Keywords — Dialectal Speech Synthesis, Text-to-Speech Systems, Dialect-distinctive features, HMM Based TTS, Voice Conversion, Gaussian Mixture Models, Formant Transformation, Voice Onset Time, Vowel Space, Diphthongization, Spectral Tilt, Mel Cepstral Coefficients, Pitch, Mel Cepstral Distortion, Root Mean Square Error, Mean Opinion Score