Abstract

Text-to-speech synthesis (TTS) deals with the conversion of an input text message to equivalent speech. Nowadays, statistical parametric speech synthesis (SPSS) based on either hidden Markov models (HMMs) or deep neural networks (DNNs) is the most widely used speech synthesis technique. The major advantages of this technique are, (i) flexibility in adapting to different speaking styles and emotions and (ii) decent performance with reduced memory footprint and computational resources. A conventional SPSS system is often viewed as being very language specific. Language dependency can be a limiting factor in using TTS technology for applications requiring synthesis of speech from multi-lingual text. Therefore, polyglot speech synthesis is growing in popularity over the past few years. The term polyglot speech synthesis refers to the process of generating speech in multiple languages with the same voice from a single TTS system. A polyglot synthesizer will find many applications like virtual news reader, telephone services, story telling audio books, e-mail reader, and so on. There are different ways to develop a polyglot TTS system, each having its own merits and demerits. In this work, the polyglot TTS systems are developed using a cross-lingual voice conversion (CLVC) technique. There is a loss of intelligibility and naturalness in the existing CLVC-based polyglot SPSS system due to three factors: (i) errors in pitch extraction, (ii) improper excitation modeling and (iii) sound quality degradation due to voice conversion. Further, the existing polyglot synthesizer is not very flexible in terms of creating new customized voices. In this thesis, we address the aforementioned issues to develop a high-quality and flexible polyglot SPSS system for Indian languages.

The major contributions of this thesis can be summarized as follows :

1. A robust pitch extraction method is proposed for improving the quality of speech synthesized from statistical parametric speech synthesis.

2. A source modeling method is proposed based on epoch parameters and phone-specific natural residual segments.

3. A method based on deep auto-encoder bottleneck features has been proposed for cross-lingual voice conversion.

4. A customizable DNN-based polyglot TTS system is developed for Indian languages by integrating the CLVC technique into multilingual TTS framework.

Keywords: *Statistical Parametric Speech Synthesis, Polyglot Speech Synthesis, Cross-Lingual Voice Conversion, Hidden Markov Models, Deep Neural Networks, Excitation Modeling, Continuous Wavelet Transform, Pitch Extraction, Deep Autoencoder, Gaussian Mixture Models*