Abstract

The objective of the phone recognition system (PRS) is to convert a speech signal into a sequence of phones. The PRS can be developed using data from single language or multiple languages. A PRS developed with data from more than one language is known as multilingual PRS (MPRS). In the literature, there are very few studies on multilingual phone recognition. The MPRS trained and tested using data from the same speech mode, is known as mode dependent or mode specific MPRS (MDMPRS). However, according to literature survey, speech can be broadly classified into three modes, namely, read, extempore and conversation. In conversation mode, two or more people will communicate in an unconstrained environment. In the extempore mode, the speaker will speak on a topic continuously without the help of any notes. For example, delivering a lecture to students in a class room. In read mode, the speaker delivers the speech using formal language and in a constrained environment. The performance of MDMPRS will be affected when the test utterance belongs to a different mode of speech. This is mainly due to mismatch in the acoustic characteristics of speech signals across different modes. Therefore, there is a need to develop a multilingual phone recognition systems which can accurately recognize the phonetic units irrespective of the mode of the speech signal. In this thesis, we address some of the above listed challenges to build a high-quality and robust phone recognition system for Indian languages.

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

1. A speech mode classification model is proposed based on excitation source, and vocal tract features to improve the performance of phone recognition system.
2. A two-stage system is proposed for robust and accurate recognition of the phonetic units present in speech utterances from multiple languages spoken in multiple modes.
3. An approach based on continuous wavelet transform coefficients and phone boundaries is proposed to detect the speech events such as vowel onset points (VOPs), vowel end points (VEPs) and vowel regions (VRs) for improving the performance of multilingual speech mode classification model.
4. A CycleGAN based non-parallel speech mode transformation (SMT) model is proposed as the front-end to improve the robustness of MPRS by transforming the characteristics of conversation and extempore modes into read mode of speech.

Keywords: *Multilingual Phone Recognition, Multilingual Speech Mode Classification, Multilingual Speech Mode Transformation, Read Speech Mode, Extempore Speech Mode, Conversation Speech Mode, Vocal Tract Features, Source Features, Feed Forward Neural Networks, Continuous Wavelet Transform, Phone Boundary, Vowel Detection, Cycle-Consistent Generative Adversarial Network*