Abstract

Pitch or fundamental frequency estimation is the task of automatically estimating the rate of vocal fold vibrations of a human speech signal. The major categorization of human voice based on prosodic variations includes neutral speech, song, and emotional speech. Initial pitch estimation approaches focused on neutral speech signals, and later they were adapted to songs. Although the fundamental frequency estimation can be adapted to work with speech and song, tuning of input parameters is not a trivial task. This becomes even more difficult when analyzing emotional, conversational and storytelling speech signals, since the prosodic variations resemble the characteristics of both speech and song. The main objective of our work is to develop a single pitch estimation method that works well across neutral speech, song and emotional speech signals.

In this work, the pitch estimation task is analyzed from time-domain, spectral-domain and the data-driven approaches for neutral speech, song and emotional speech signals. In the spectral domain, we develop novel features and use machine-learning algorithms in detection of harmonic partials. Specifically, we exploit the characteristics of resonant frequencies and strong harmonic partials in voiced segments to identify at least three successive harmonic partials. In the time-domain, the glottal excitation regions are identified based the RNN-LSTM model. The fundamental frequency is estimated from the voiced regions by exploiting the quasi-harmonic and quasi-periodic properties of the signal. In the data-driven approach, the fundamental frequency estimation is considered as a classification problem. A novel encoder-decoder model is developed with the deep learning approaches performed in encoder block and signal processing approaches performed in the decoder block. The developed fundamental frequency estimation approaches are deployed for developing pedagogical tools for learning Hindustani music.

The major contributions of the works presented in the thesis can be summarized as follows:

- A robust GCI estimation method based on filtering and non-parametric techniques is proposed.
- A spectrum-based f_0 estimation method is developed for accurate melody extraction from monophonic songs.
- A LSTM model is developed for accurate detection of voiced/unvoiced regions in the speech and the song signals.
- A time-domain filtering approach is proposed for accurate estimation of f_0 from the speech and song signals.
- A hybrid method consisting of deep learning and signal processing techniques is proposed for accurate f_0 estimation from both speech and song signals.
- An automatic SARGAM learning system is developed for guiding the amateur singers to practice and learn singing in the absence of the teacher.
- A system to identify the raga based on the note distribution is developed.

Keywords: *Glottal Closure Instant, Harmonic Partials, Convolutional Neural Networks, Long Short Term Memory, Neutral speech, Monophonic songs, Multi-Layer Perceptron, Melody Extraction, Fundamental frequency estimation, Pitch estimation, Hindustani Music, Indian Classical, SARGAM analysis, Tanpura tuner, Raga Identification.*