Abstract

Automatic speech recognition (ASR) has established itself as an emerging technology, becoming popular day by day. ASR systems have emerged significantly from training the multiple blocks in an conventional ASR system, to the use of a powerful alternative end-to-end mechanism which replace most modules with a single model. Despite this progress, the performance of ASR system is still limited by several factors such as speaker variations and gender bias. Hence, these variations has to be effectively captured by considering the speech production aspect deriving the appropriate discriminative features. Manner of Articulation (MoA) is related to the speech production mechanism, and it is unbiased with respect to the variations on the speaker and gender characteristics. MoA represents the positioning and movement of various articulators during the production of speech. The MoA knowledge can be integrated for better discrimination among different phonemes in ASR. The acquired MoA knowledge is generally incorporated at the front-end of conventional ASR to improve its performance. However, there are other fundamental blocks in conventional ASR such as decoding graph, lattice structures where the manner knowledge is not explicitly incorporated. Also, the MoA knowledge is not explored in end-to-end ASR system. Hence, this thesis attempts to detect the manners of articulation accurately, and then the acquired MoA knowledge is used in different stages of conventional ASR and end-to-end ASR.

In the first phase of the thesis, the two broad manners of articulation namely sonorants (vowels, semi-vowels, nasals) and obstruents (fricatives, stop consonants, affricates) are detected by exploiting the spectral features. The spectral flatness measure (SFM) is computed on the magnitude linear prediction (LP) spectrum to discriminate sonorants from obstruents. The acquired MoA knowledge is embedded in different stages of conventional ASR such as long short term memory (LSTM) structure, decoding graph and lattice structure to improve the ASR performance. The modified LSTM architecture forces the input to hidden layer-1 to discriminate sonorants from obstruents. As a result, the sonorants which are falsely substituted or inserted as obstruents at the output layer are minimized in the ASR decoded sequence. Later, the decoding graph is restricted to produce either the sonorants or the obstruents at each short time step. As a result, the path in the decoding graph will be according to the manner of articulation. Further, the word lattice is re-scored based on the acquired MoA knowledge to improve ASR performance.

In order to detect the finer levels of MoA using the spectral features, the SFM showed poor discriminative nature within the sonorants and the obstruents. This was because of the overlapping of the spectral characteristics at the finer levels of MoA detection. Therefore, in the second phase of the thesis, a deeper architecture based on end-to-end connectionist temporal classification (CTC) system is explored to detect the finer levels of MoA such as vowels, semi-vowels, nasals, fricatives and stop consonants.

The acquired MoA knowledge is used to modify the decoder of the end-to-end ASR system. The posterior probabilities of the baseline CTC detector are modified by the incorporation of the manner CTC detector. The modified manner based character CTC is evaluated on open source speech datasets such as AN4, LibriSpeech and TEDLIUM-2 and it outperformed over the baseline character CTC by the relative improvement between 4% and 8%. Further, the acoustic model weights of the state-of-the-art end-to-end ASR system are initialized with the manner CTC detected weights. The pre-trained weights improved the performance of the state-of-the-art CTC based end-to-end ASR system on open source datasets such as AN4, LibriSpeech and TEDLIUM-2.

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

- A method based on SFM is proposed to detect the two broad MoA namely the sonorants, and the obstruents in a continuous speech. The acquired MoA knowledge is explicitly incorporated in the LSTM based acoustic models for improving the performance of the ASR system.
- The acquired 2-class MoA knowledge is incorporated in the decoding graph for improving the performance of the conventional ASR system.
- The acquired 2-class MoA knowledge is incorporated to re-score the graph cost of the conventional word lattice for improving the performance of the conventional ASR system.
- A method based on CTC based end-to-end system is explored to detect the finer levels of MoA such as vowels, semi-vowels, nasals, fricatives and stop consonants.
- The acquired 5-class MoA knowledge is incorporated in modifying the decoder of the end-to-end ASR system.
- The knowledge of manner CTC pre-trained weights is used in re-training the end-to-end ASR system.

Keywords: *Automatic Speech Recognition (ASR), Manner of articulation, spectral flatness measure (SFM), long short term memory (LSTM), decoding graph, lattice, end-to-end ASR, connectionist temporal classification (CTC).*