Dialect normalisation with deep learning-based automatic speech recognition

Mahsa Vafaie

Automatic speech recognition has benefited a great deal from advances in machine learning throughout the past decade. Convolutional neural networks have demonstrated the ability to effectively model speech signals when provided with very large amounts of input data. To date, these technological advances have been mainly applied to a small number of languages, such as English, German, Chinese and Spanish.

In this thesis, I explore the feasibility of deep learning-based techniques for automatic speech recognition for an under-resourced language, namely Farsi. To accomplish this, I implement Speech-to-Text-WaveNet, a convolutional neural network architecture, on TFarsDat, a corpus of telephone conversations in Farsi. I also repurpose this pipeline for a dialect normalisation task, which aims at outputting standardised transcriptions of colloquial speech in different dialects of Farsi. While the final results of the system are far below state-of-the-art performance, the approach nonetheless demonstrates the viability of deep learning-based ASR, presupposing the existence of high-quality training data. I suggest that development of large, gold-standard datasets for under-resourced languages such as Farsi could be coupled with the proposed approach to improve the state of ASR systems more generally and for dialect normalisation in Farsi in particular.