POSTER 14

Exploring Generalizability of Automatic Phoneme Recognition Models

Emir Demirel(1), Sven Ahlback(2) and Simon Dixon(1)
(1) Queen Mary University of London, UK
(2) Doremir Music Research AB
Email: e.demirel@qmul.ac.uk

Human speech and singing voice are both produced by the same sound source, the vocal organ. Despite its growing popularity in the last decades, phoneme / word recognition in singing voice has not been widely investigated as it is in the speech domain. According to prior research, one of the major differences between speech and singing is the duration of vowels. This can be interpreted as difference in pronunciation of the voiced phonemes. Phoneme recognition in singing is still not a solved problem due to complex spectral characteristics of the sung vowels. In this study, we tackle this problem using recent and traditional Automatic Speech Recognition (ASR) models that are trained on different speech corpora. To observe the influence of pronunciation, we hold experiments on 'NUS Sung and Read Lyrics Corpus', which consists of lyrics-level utterances both pronounced as speech and singing. We perform the experiments using the Kaldi ASR Toolkit and explore different topologies in the Kaldi PyTorch extension. We further analyze the recognition and the alignment results on both singing and speech, and address the problems to achieve a better recognition result in singing. We have obtained some initial results where we observed a decrease in recognition performance when context dependency is added to the feature space. This indicates that it is necessary to include domain specific information in the phoneme recognition pipeline when applied to singing voice.