Comparing data-driven and handcrafted features for dimensional emotion recognition

Bogdan Vlasenko, Sargam Vyas, Mathew Magimai.-Doss

Speech Emotion Recognition (SER) has garnered significant attention over the past two decades. In the early stages of SER technology, 'bruteforce'-based techniques led to a significant expansion in knowledge-based acoustic feature representation (FR) for modeling sparse emotional data. However, as deep learning techniques have become more powerful, their direct application has been limited by the scarcity of well-annotated emotional data. As a result, pretrained neural embeddings on large speech corpora have gained popularity for SER tasks. These embeddings leverage existing transfer learning methods suitable for general-purpose self-supervised learning (SSL) representations. Recent studies on downstream SSL techniques for dimensional SER have shown promising results. In this research, we aim to evaluate the emotion-discriminative characteristics of neural embeddings in general cases (out-of-domain) and when fine-tuned for SER (in-domain). Given that most SSL techniques are pre-trained primarily on English speech, we plan to use speech emotion corpora in both language-matched and mismatched conditions. We will assess the discriminative characteristics of both handcrafted and standalone neural embeddings as FRs.