Contribution ID: **122** Contribution code: **110** Type: **Poster**

Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models

In large part due to their implicit semantic modeling, self-supervised learning (SSL) methods have significantly increased valence recognition performance in speech emotion recognition (SER) systems. Yet, their large size may often hinder implementation in applications such as virtual assistants and digital customer service. In this work, we analyze the relevance for SER of each of HuBERT's layers, showing that shallower/deeper layers are more important for arousal/valence recognition, respectively. This motivates the use of additional textual information for accurate valence recognition, as the distilled model lacks the depth of its teacher. Thus, we propose a framework that, while having only ~20% of the trainable parameters of a large SSL model, achieves on par SER performance on the MSP-Podcast dataset.

Find me @ my poster

1, 2

Keywords

speech emotion recognition
self-supervised learning
knowledge distillation
paralinguistics
semantics

Authors: DE OLIVEIRA, Danilo (Signal Processing (SP), Universität Hamburg); RAJ PRABHU, Navin (Signal Processing); GERKMANN, Timo