Adapting a Self-Supervised Speech Representation for Noisy Speech Emotion Recognition by using Contrastive Teacher-Student Learning

Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso

Motivation

Background:
- Self-supervised speech representation with transformer shows good performance for speech emotion recognition (SER) tasks
- It still requires adapting the model to the target noisy environment for real-world applications

Our Work:
- We build an appropriate adaptation algorithm for SER model to compensate for environmental noise
- Our approach is built upon a pre-trained transformer that can:
  - Acquire new knowledge from adverse recording conditions
  - Keep original knowledge acquired during pre-training and fine-tuning

Proposed Method

- Emotion prediction loss ($L_{CCC}$)
  - Train student model to predict accurate emotion from noisy speech
  - Maximize concordance correlation coefficient (CCC)
  - $L_{CCC} = (1 - CCC_{tra}) + (1 - CCC_{Dom}) + (1 - CCC_{Val})$

- Transfer loss ($L_{TL}$)
  - Prevent student model from catastrophic forgetting of teacher model’s knowledge
  - $L_{TL} = \frac{1}{2} \sum_{x \in D_{neg}} \left( L(x_{clean}) - S(x_{noisy}) \right)^2$

- Contrastive loss ($L_{CL}$)
  - Learn the n emotionally discriminative knowledge regardless of environment
  - $L_{CL} = InfoCE(\text{anchor} = S(x_{noisy}), \text{positive} = T(x_{clean}), \text{negative} = S(x_{-j}^{(-)}_{noisy})$}
  - Discard the sample from negatives if it has similar emotion to the anchor

Experiment & Embedding Analysis

- Only maximizing CCC during environment adaptation causes catastrophic forgetting of pre-trained and fine-tuned knowledge for SER model
- Teacher-student learning can keep the original model’s knowledge while acquiring new knowledge from noisy speech
- Contrastive learning can further improve performance by learning emotionally discriminative knowledge regardless of environmental conditions

Future Work
- Environment adaptation for self-supervised representations to various noise types

This study was supported by NIH under grant 1R01MH122367-01