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

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

Multimodal Signal Processing Lab (MSP)

Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas - 75080, USA

ICASSP 2023, 4-10 JUNE, RHODES ISLAND, GREECE

Motivation

Background:

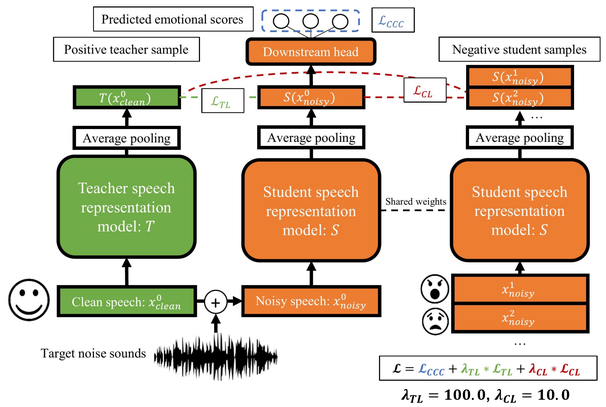
- Self-supervised speech representation with transformer shows good performance for speech emotion recognition (SER) tasks
- It still requires adapting the model to the target noisy environment for real-world applications

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

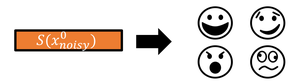
Proposed Method

Contrastive Teacher-Student Learning



- Set fine-tuned model as a teacher model (T)
- Copy T and set it as a student model (S)
- Corrupt train set with target noise to update S

- **Emotion prediction loss ($\mathcal{L}_{CCC}$)**
  - Train student model to predict accurate emotion from noisy speech
  - Maximize concordance correlation coefficient (CCC)
  - $\mathcal{L}_{CCC} = (1 - CCC_{Aro.}) + (1 - CCC_{Dom.}) + (1 - CCC_{Val.})$

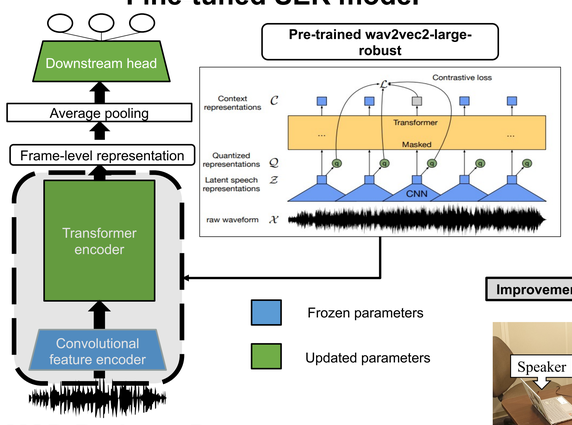


- **Transfer loss ($\mathcal{L}_{TL}$)**
  - Prevent student model from catastrophic forgetting of teacher model's knowledge
  - $\mathcal{L}_{TL} = \frac{1}{N}\sum_{i=0}^{N}\{\sqrt{\left(T(x^i_{clean}) - S(x^i_{noisy})\right)^2}\}$

- **Contrastive loss ($\mathcal{L}_{CL}$)**
  - Learn the n emotionally discriminative knowledge regardless of environment
  - $\mathcal{L}_{CL} = InfoNCE(anchor = S(x^i_{noisy}), positive = T(x^i_{clean}), negative = S(x^{0\ldots j}_{noisy}))$
  - Discard the sample from negatives if it has similar emotion to the anchor

Experiment & Embedding Analysis

Fine-tuned SER model



Test Set Result

Matched condition / Mismatched condition

	0dB			5dB			10dB		
	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence
Original	0.244	0.226	0.227	0.379	0.335	0.308	0.438	0.374	0.351
RH	0.323	0.278	0.164	0.443	0.390	0.236	0.494	0.424	0.278
RM	0.272	0.215	0.330	0.412	0.328	**0.418**	0.459	0.350	0.454
RM+TL	0.345	0.289	**0.337**	0.474	0.403	0.416	0.513	0.430	**0.459**
RM+CL	0.339	0.296	0.314	0.451	0.391	0.399	0.481	0.402	0.433
RM+TL+CL	**0.347**	**0.300**	0.335	**0.477**	**0.410**	0.417	**0.523**	**0.435**	0.452

Baselines
- Fine-tuned SER model (Original)
- Retrain downstream head (RH)
- Retrain all model (RM)

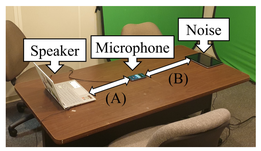
Only uses $\mathcal{L}_{CCC}$

For ablation study
- Contrastive teacher-student learning without contrastive learning (RM+TL)
- Contrastive teacher-student learning without transfer learning (RM+CL)

Proposed model
- Contrastive teacher-student learning (RM+TL+CL)

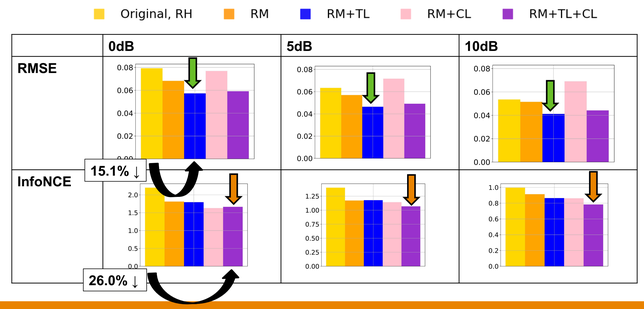
Improvements in all the attributes

MSP-Podcast Dataset



- We use the clean & noisy version of the MSP-Podcast corpus
- Adapt the model with 30 minutes of the target noise sounds
- Test the model with the noisy version of the MSP-Podcast corpus (10dB, 5dB, 0dB)

Embedding Analysis



Conclusions

- **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

NIH National Institutes of Health