Computation and Memory Efficient Noise Adaptation of Wav2Vec2.0 for Noisy Speech Emotion Recognition with Skip Connection Adapters

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

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Motivation

Background:

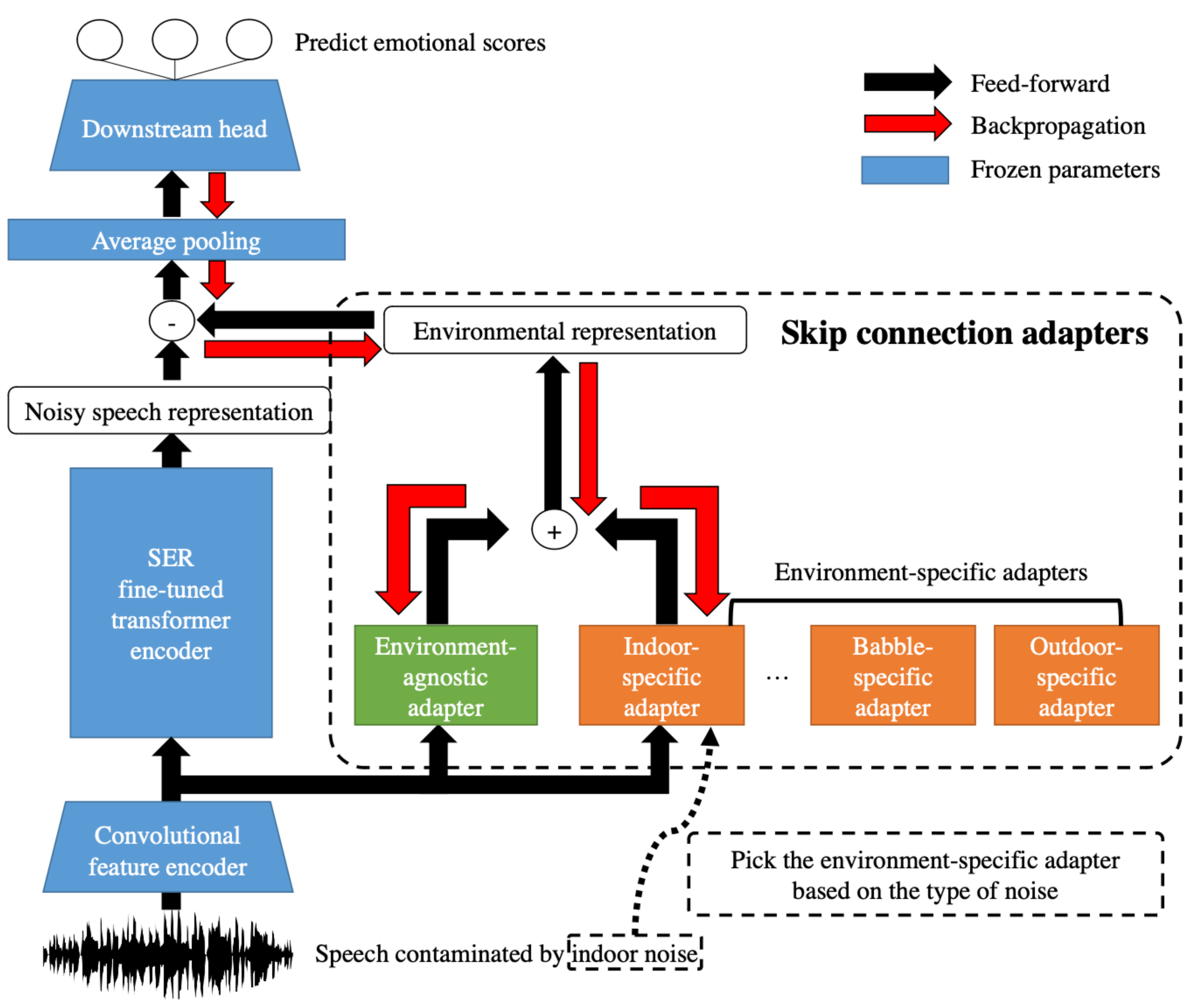
- Fine-tuning a large pre-trained transformer model (Wav2Vec2.0, HuBERT) performs well in SER tasks
- Fine-tuning the model under multiple noisy environments requires considerable resources (adaptation time and parameter space)

Our Work:

- Propose **environment-agnostic and environment–specific adapters** to adapt a pre-trained transformer model to multiple environments
  - **Decrease the parameter space requirements** for each noisy environment
  - **Reduce the adaptation time** by avoiding the gradient backpropagation through the transformer encoder

Proposed Method

Environment adaptation with skip connection adapters



Skip connection adapters

- Environment-agnostic adapter ($A_{agn.}$): updated for **all the environments**
- Environment-specific adapter ($A_{spe.}$): updated for **each target environment**

Denoised speech representation

- Select the environment-specific adapter, $A^i_{spe.}$, with respect to the input environment i
- Get denoised representation, $z(\widehat{x^i})$, with $A_{agn.}$ and the selected $A^i_{spe.}$

$$z(\widehat{x^i}) = T\left(E(\widehat{x^i})\right) - \{A_{agn.}\left(E(\widehat{x^i})\right) + A^i_{spe.}\left(E(\widehat{x^i})\right)\}$$

- $\widehat{x^i}$: *Input noisy speech*
- *E: Convolutional feature encoder*
- *T: Transformer encoder*

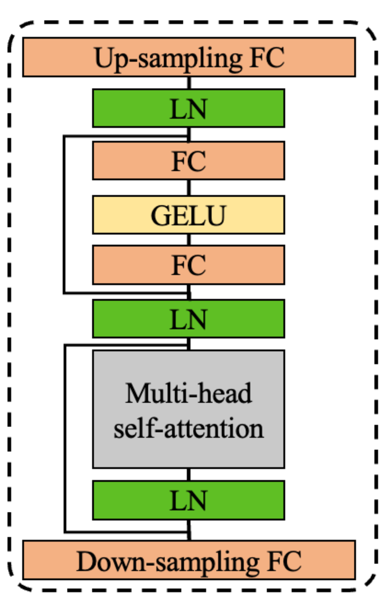
Experiment Settings

Data preparation

- Fine-tune the pre-trained wav2vec2-large robust with the clean version of the MSP-Podcast corpus (v1.8) [Wagner, 2022]
- Contaminate the clean version of the MSP-Podcast corpus to simulate six noisy environments
  - **Radio, Babble, Indoor, outdoor, house, and vehicle**

Adapter architecture

- The architecture of the adapter is the same as a single transformer layer of wav2vec2.0
  - LN: layer normalization
  - FC: fully connected layer
- Shrink dimension size from 1,024 to 256
- Use the same architecture for each environment-specific and –agnostic adapters



Results

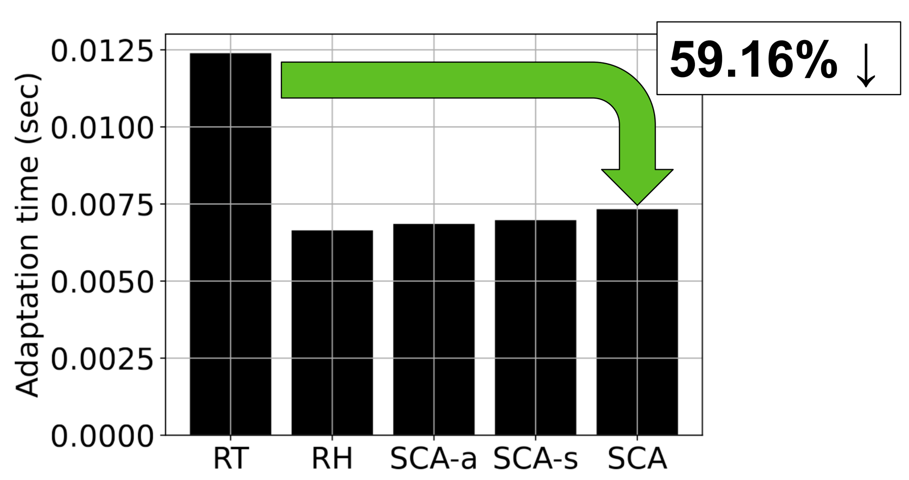
Emotion Recognition Performance (CCC)

	10dB			5dB			0dB		
	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.
Original	0.596	0.562	0.473	0.526	0.506	0.424	0.432	0.418	0.338
RT	0.634*	**0.587***	**0.507***	0.581*	**0.541***	0.453*	0.492*	0.465*	0.369*
RH	**0.637***	0.553	0.484	**0.590***	0.535*	0.433	0.497*	0.458*	0.349
SCA-a	0.613	0.571	0.499*	0.561	0.534	0.450*	0.446	0.445	**0.371***
SCA-s	0.629*	0.580*	0.464	0.581*	0.536*	0.410	0.473*	0.460*	0.307
SCA	0.633*	0.573*	0.506*	0.583*	0.540*	**0.461***	**0.502***	**0.469***	0.370*

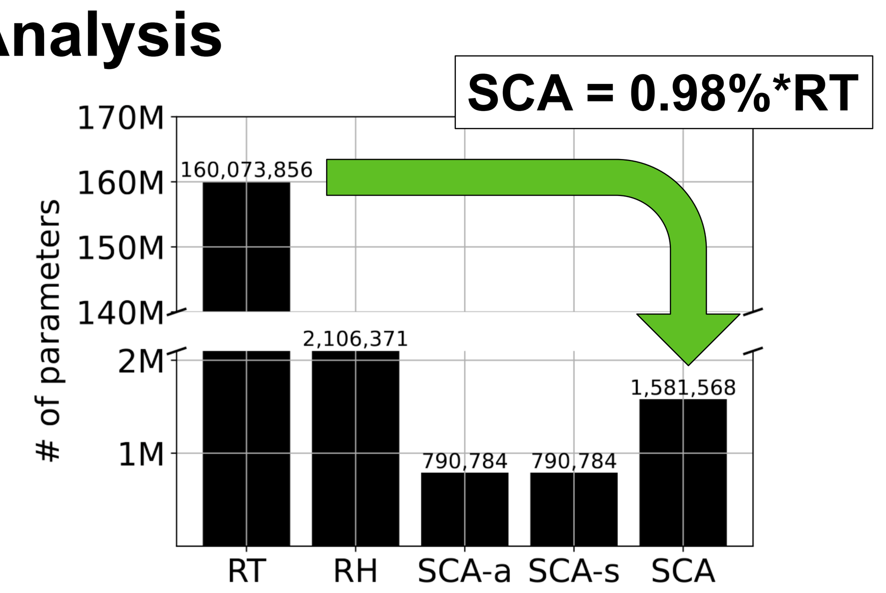
- RT: retrain transformer layers
- RH: retrain downstream head
- SCA-a: use only environment-agnostic skip connection adapter
- SCA-s: use only environment-specific skip connection adapters

Skip connection adapter (SCA) leads to improvements for all the attributes

Efficiency Analysis



Requires less adaptation time than RT while achieving similar performance



SCA is memory efficient method for multiple noisy environments

Conclusions

- **Combining environment-agnostic and environment-specific adapters can improve SER performance** under multiple noisy environments
- **Our proposed adaptation method can decrease the time and memory requirements** to adapt the model to a new environment

Future Work

- Understand why environment-agnostic adapter helps valence prediction, and environment-specific adapter helps arousal and dominance predictions

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