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.

Motivation

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, $A_{spe}$, with respect to the input environment $i$.
- Get denoised representation $z(x^i)$ with $A_{agn}$ and the selected $A_{spe}$.

\[ z(x^i) = T(E(x^i)) - (A_{agn}(E(x^i)) + A_{spe}(E(x^i))) \]

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 environment-agnostic adapter.

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.

Results

Emotion Recognition Performance (CCC)

<table>
<thead>
<tr>
<th></th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aro</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SAC = 0.98%*RT

Efficiency Analysis

Requires less adaptation time than RT while achieving similar performance.

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