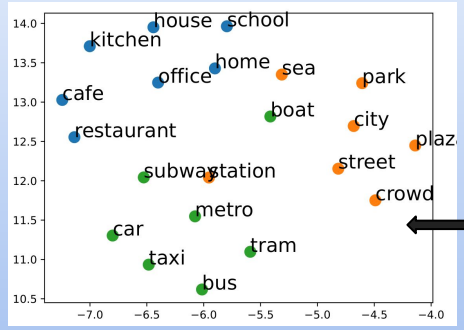
Describe Where You Are: Improving Noise-Robustness for Speech Emotion Recognition with Text Description of the Environment

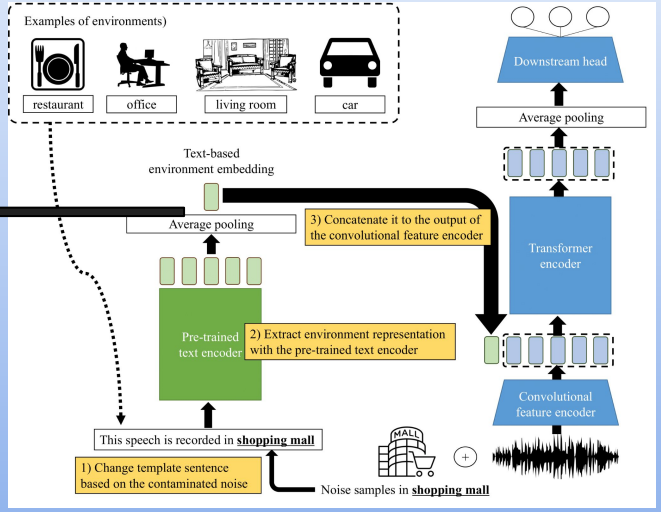
Motivation

Use a **pre-trained text encoder** to leverage **unseen target environment information** for SER under noisy environment

Semantically similar environments are clustered together in the embedding space



Proposed Framework



Observations

1. Proposed approach **yields better noise-robustness for SER than re-training the model only with noisy speech**
2. Text-guided environment embedding **outperforms than inferring an environment condition**, especially in low SNR-level conditions