Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition

Sefik Emre Eskimez, Zhiyao Duan, Wendi Heinzelman
eeskimez@ur.rochester.edu, {zhiyao.duan,wendi.heinzelman}@rochester.edu
Department of Electrical and Computer Engineering, University of Rochester

Motivation

- **Problem:** Lack of labeled training data
- Recording and annotating emotional speech is a time-consuming process
- **Solution:** Unsupervised feature learning

  - Learn features from widely available general speech
  - Use learned features for automatic speech emotion recognition (ASER)

Method

We follow these steps to build our system:

1. Train an autoencoder
2. Freeze the encoder parameters
3. Add fully connected (FC) layers on top of encoder for classification

Proposed System Overview

- **Figure 1:** Proposed ASER system overview. The dashed red windows represent the sliding window with 50% overlap. From each window, emotion class probabilities \( p_1, p_2, p_3, p_4 \) and \( p_5 \) are predicted and the average of these vectors is calculated over all windows is calculated for each utterance.

**Denoising Autoencoder (DAE)**

**Adversarial Autoencoder (AAE)**

**Variational Autoencoder (VAE)**

**Adversarial Variational Bayes (AVB)**

Results

**Figure 2:** DAE network architecture: reconstructing the clean spectrogram from noisy input

**Figure 3:** AAE network architecture: variational inference on auto-encoder by constraining the latent representation through adversarial training

**Figure 4:** VAE network architecture: variational inference on auto-encoder by constraining the latent representation to follow a normal distribution

**Figure 5:** AVB network architecture: unifying VAE and generative adversarial networks (GANs)

**Figure 6:** The unweighted accuracy rating (UAR) results for the baseline and proposed systems.

**Figure 7:** F1-score results for the baseline systems and the proposed systems. F1-score is calculated for each class, and their unweighted mean is presented.

Conclusions

- Proposed a CNN based ASER system
- Systematically explored the following unsupervised methods for ASER:
  - DAE, VAE, AAE, and AVB
- Showed that these methods performed better than the SVM and CNN baselines