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

- Denoising Autoencoder (DAE)
- Adversarial Autoencoder (AAE)
- Variational Autoencoder (VAE)
- Adversarial Variational Bayes (AVB)

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>UAR (%)</th>
<th>F1-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>43.94</td>
<td>41.0</td>
</tr>
<tr>
<td>CNN</td>
<td>44.82</td>
<td>42.0</td>
</tr>
<tr>
<td>DAE</td>
<td>46.02</td>
<td>43.0</td>
</tr>
<tr>
<td>VAE</td>
<td>46.84</td>
<td>44.0</td>
</tr>
<tr>
<td>AAE</td>
<td>48.18</td>
<td>45.0</td>
</tr>
<tr>
<td>AVB</td>
<td>48.54</td>
<td>46.0</td>
</tr>
</tbody>
</table>

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