SELF-SUPERVISED DENOISING AUTOENCODER WITH LINEAR REGRESSION DECODER FOR SPEECH ENHANCEMENT

Ryandhimas E. Zezario, Tassadaq Hussain, Xugang Lu, Hsin-Min Wang, Yu Tsao

Presented by:
Ryandhimas Zezario
ryandhimas@citi.sinica.edu.tw
Outlines

• Introduction
• The Proposed Denoising Autoencoder with Linear Regression Decoder (DAELD) System
• Experiments
• Conclusion
Introduction

• What is speech enhancement?

Speech enhancement aims to retrieve clean speech signals from noisy ones and serves as an important pre-processor in many speech related tasks, such as:

➢ Automatic speech recognition
➢ Assistive listening
➢ Speech coding
➢ Speaker recognition
Introduction

• Trends of speech enhancement
  ➢ Started by statistical based speech enhancement
  ➢ Followed by machine learning based speech enhancement
  ➢ Deep-learning-based methods have caught great attention in recent years, in particular the supervised based approach
Introduction

• Challenge of supervised learning based speech enhancement
  ➢ A pair set of noisy and clean is a must
  ➢ Required a sufficient amount of training data
  ➢ No guarantee when operating under unseen or noise types or speakers
Introduction

• **Unsupervised learning**
  - Unrequired labelled training data
  - It can extract essential representations from the salient structure of the input data
  - Example is Autoencoder
Introduction

• Autoencoder
  ➢ It consists of encoder and decoder
  ➢ Encoder transforms the input physical data into latent features
  ➢ Decoder will reconstruct to the original data
The proposed DAELD

- **Architecture**

![Diagram of the DAELD architecture]

- Encoder
  - Input
  - Non-Linear Transformation
  - Linear Regression
- Decoder
  - Output
- Encoder
  - Input
  - Non-Linear Transformation
The proposed DAELD

• Two types of DAELD
  ➢ DAELD$_{\text{(SAE)}}$ and DAELD$_{\text{(BP)}}$
  ➢ DAELD calculates the weights in the encoder in an unsupervised self-learning training criterion
  ➢ It consists of offline and online stages
The proposed DAELD

• Offline
  ➢ DAELD\textsubscript{(SAE)}
    \[
    \boldsymbol{\beta}_{SAE} = \left( \delta I + H_{SAE}^T H_{SAE} \right)^{-1} H_{SAE}^T Y
    \]
  ➢ DAELD\textsubscript{(BP)}
    \[
    \boldsymbol{\beta}_{BP} = \left( \delta I + H_{BP}^T H_{BP} \right)^{-1} H_{BP}^T Y
    \]
The proposed DAELD

• Online

➢ We obtain hidden layer output $\mathbf{H}$ by the encoder whose parameters are trained in the unsupervised manner

➢ Based on the estimated linear transformation, $\mathbf{\beta}$ (either $\mathbf{\beta}_{SAE}$ or $\mathbf{\beta}_{BP}$) the enhanced speech spectral can be estimated as:

$$\mathbf{\hat{X}} = \mathbf{H} \mathbf{\beta}$$
Experiments

• Experimental setup
  ➢ Aurora-4 dataset
    ✓ 2676 training utterances
    ✓ Six types of noises (babble, car, restaurant, street, airport, and train)
    ✓ SNR levels varying from 10 to 20 dB
    ✓ Noisy utterances (contaminated with babble and car noises) at SNR levels varying from 5 to 15 dB, were used as the test data.
Experiments

• Experimental setup
  ➢ TIMIT
  ✓ 4620 training
  ✓ 90 types of noises at eight SNR levels (from -10 dB to 25 dB with steps of 5 dB) into the clean training
  ✓ Four unseen (two stationary and two non-stationary) noise types under five SNR levels (-12 dB, -6 dB, 0 dB, 6 dB and 12 dB) to test the enhancement performance
Experiments

• **Experimental setup**
  - 80-dimensional Mel frequency power spectrum (MFP)
  - DAELD models were formed by a three-layered architecture with [1000 1000 16000] hidden nodes
Experiments

- **Objective evaluation results**
  
  > Aurora 4

<table>
<thead>
<tr>
<th></th>
<th>Babble</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Babble</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ICASSP 2020
## Experiments

- **Objective evaluation results**

  ➢ **TIMIT**

### PESQ

<table>
<thead>
<tr>
<th></th>
<th>12</th>
<th>6</th>
<th>0</th>
<th>-6</th>
<th>-12</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noisy</strong></td>
<td>2.45</td>
<td>1.95</td>
<td>1.60</td>
<td>1.39</td>
<td>1.30</td>
<td>1.74</td>
</tr>
<tr>
<td><strong>DDAE</strong></td>
<td>2.53</td>
<td>2.18</td>
<td>1.79</td>
<td>1.47</td>
<td>1.32</td>
<td>1.86</td>
</tr>
<tr>
<td><strong>MMSE</strong></td>
<td>2.78</td>
<td>2.24</td>
<td>1.81</td>
<td>1.53</td>
<td>1.36</td>
<td>1.94</td>
</tr>
<tr>
<td><strong>DAELD_{BP}(s)</strong></td>
<td>2.63</td>
<td>2.27</td>
<td>1.89</td>
<td>1.53</td>
<td>1.35</td>
<td>1.93</td>
</tr>
<tr>
<td><strong>DAELD_{SAE}(s)</strong></td>
<td>2.64</td>
<td>2.27</td>
<td>1.87</td>
<td>1.52</td>
<td>1.37</td>
<td>1.94</td>
</tr>
<tr>
<td><strong>DAELD_{BP}(u)</strong></td>
<td>2.78</td>
<td>2.27</td>
<td>1.86</td>
<td>1.57</td>
<td>1.40</td>
<td>1.98</td>
</tr>
<tr>
<td><strong>DAELD_{SAE}(u)</strong></td>
<td>2.80</td>
<td>2.31</td>
<td>1.89</td>
<td>1.58</td>
<td>1.40</td>
<td><strong>2.00</strong></td>
</tr>
</tbody>
</table>

### STOI

<table>
<thead>
<tr>
<th></th>
<th>12</th>
<th>6</th>
<th>0</th>
<th>-6</th>
<th>-12</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noisy</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.82</strong></td>
<td>0.68</td>
<td>0.54</td>
<td><strong>0.43</strong></td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>DDAE</strong></td>
<td>0.81</td>
<td>0.75</td>
<td>0.65</td>
<td>0.51</td>
<td>0.37</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>MMSE</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.82</strong></td>
<td>0.68</td>
<td>0.53</td>
<td>0.40</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>DAELD_{BP}(s)</strong></td>
<td>0.82</td>
<td>0.76</td>
<td>0.67</td>
<td>0.54</td>
<td>0.40</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>DAELD_{SAE}(s)</strong></td>
<td>0.82</td>
<td>0.76</td>
<td>0.67</td>
<td><strong>0.55</strong></td>
<td>0.42</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>DAELD_{BP}(u)</strong></td>
<td>0.89</td>
<td>0.81</td>
<td>0.68</td>
<td>0.54</td>
<td>0.41</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>DAELD_{SAE}(u)</strong></td>
<td>0.88</td>
<td>0.81</td>
<td><strong>0.69</strong></td>
<td>0.54</td>
<td>0.41</td>
<td><strong>0.67</strong></td>
</tr>
</tbody>
</table>

### Stationary Noise (Car and Engine):
- Noisy: 2.50, 2.03, 1.71, 1.48, 1.37, 1.82
- DDAE: 2.61, 2.27, 1.89, 1.58, 1.40, 1.95
- MMSE: 2.61, 2.10, 1.71, 1.46, 1.26, 1.83

### Non-stationary Noise (Babble and Restaurant):
- Noisy: 2.70, 2.35, 1.98, 1.65, 1.46, 2.03
- DDAE: 2.75, 2.40, 2.01, 1.68, 1.48, **2.06**
- MMSE: 2.70, 2.21, 1.85, 1.59, 1.42, 1.95
- DDAE: 2.73, 2.24, 1.87, 1.59, 1.42, 1.97

**ICASSP 2020**
Experiments

• Hidden layer analysis
Experiments

• Spectrogram analysis
Conclusion

• The main contribution of this study is two-fold. First, we investigated to use a linear regression function to form the decoder of the DDAE model (termed DAELD) and tested the DAELD model on two speech enhancement tasks (Aurora-4 and TIMIT).

• Second, we have investigated the performance of the DAELD system trained in a self-supervised learning fashion.

• We will further test DAELD’s capability in other speech-processing tasks, such as dereverberation, or multimodal (audio-visual) speech enhancement tasks.
Thank you