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

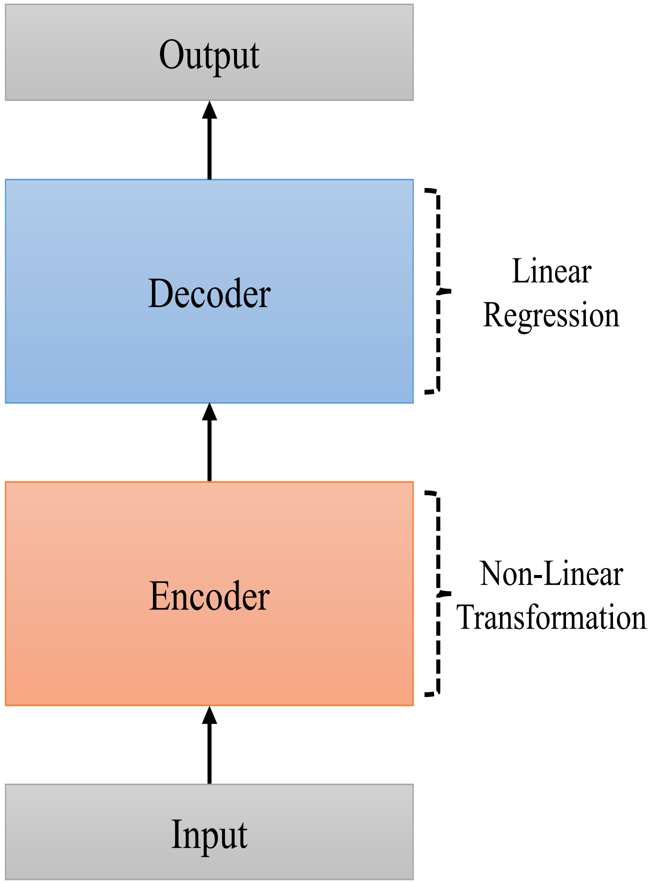
- **Unsupervised learning**
  - Unrequired labelled training data
  - It can extract essential representations from the salient structure of the input data
  - Example is Autoencoder

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**



The proposed DAELD

- **Two types of DAELD**
  - DAELD $_{(SAE)}$ and DAELD $_{(BP)}$
  - 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 (SAE)

$$\boldsymbol{\beta}_{SAE} = \left(\delta\boldsymbol{I} + \boldsymbol{H}_{SAE}{}^{T}\boldsymbol{H}_{SAE}\right)^{-1}\boldsymbol{H}_{SAE}{}^{T}\,\boldsymbol{Y}$$

  - DAELD (BP)

$$\boldsymbol{\beta}_{BP} = \left(\delta\boldsymbol{I} + \boldsymbol{H}_{BP}{}^{T}\boldsymbol{H}_{BP}\right)^{-1}\boldsymbol{H}_{BP}{}^{T}\,\boldsymbol{Y}$$

The proposed DAELD

- **Online**
  - We obtain hidden layer output $\bar{\boldsymbol{H}}$ by the encoder whose parameters are trained in the unsupervised manner

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

$$\hat{\boldsymbol{X}} = \bar{\boldsymbol{H}}\boldsymbol{\beta}$$

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

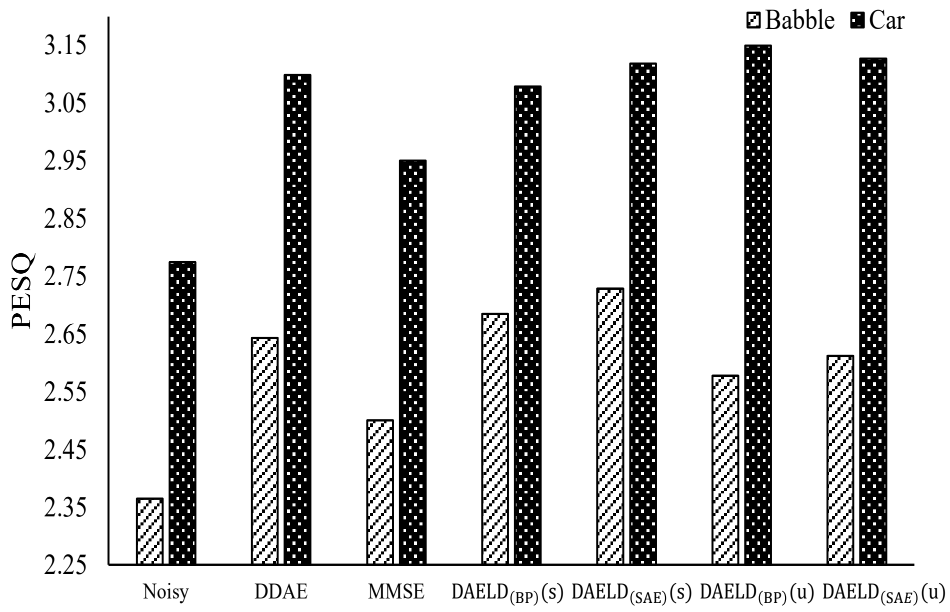
- **Experimental setup**
  - TIMIT
    - ✓4620 training
    - ✓90 types of noises at eight SNR levels (from -10 dB to 25 dB with steps of 5 dB) into the clean training
    - ✓Four unseen (two stationary and two non-stationary) noise types under five SNR levels (-12 dB, -6 dB, 0dB, 6dB and 12 dB) to test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

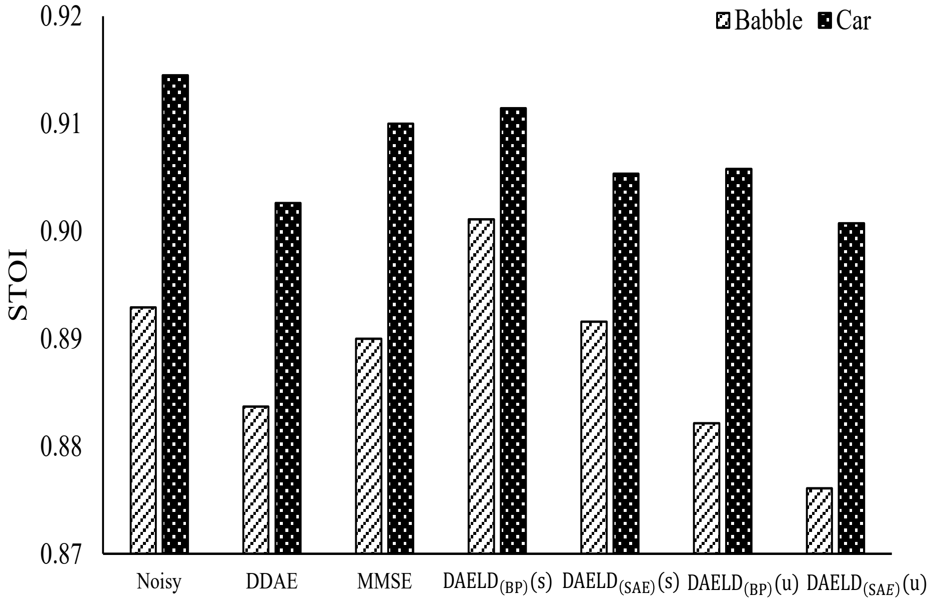
Experiments

- **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





Experiments

- **Objective evaluation results**
  - TIMIT

PESQ

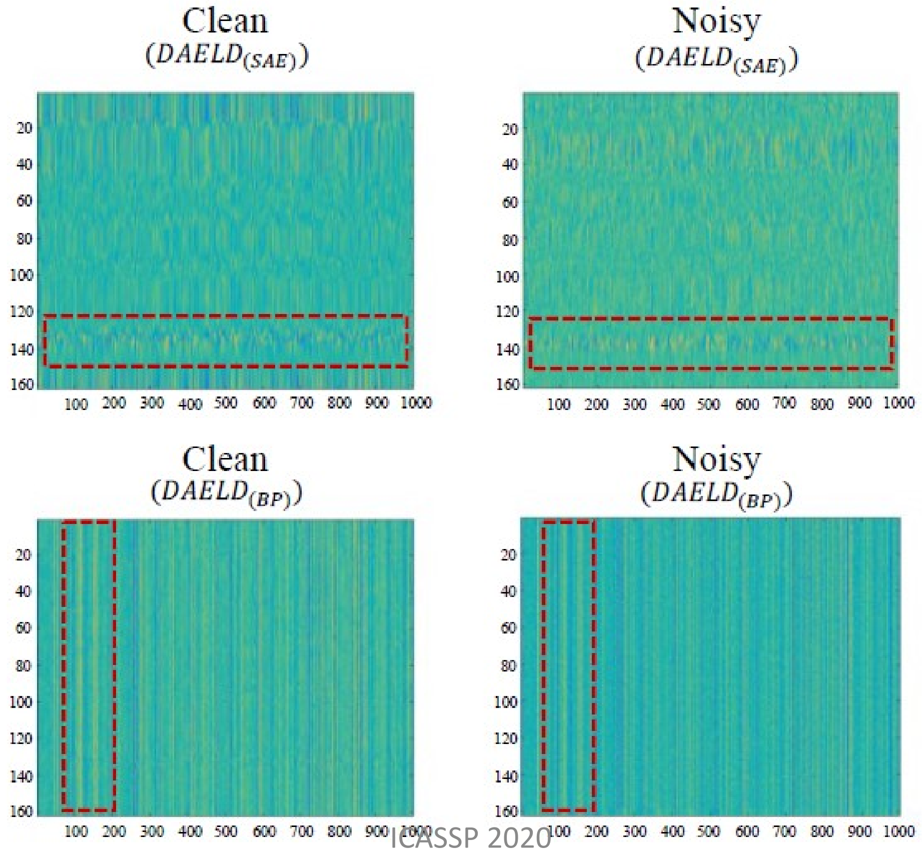
	12	6	0	-6	-12	Ave
Stationary Noises (Car and Engine)						
Noisy	2.45	1.95	1.60	1.39	1.30	1.74
DDAE	2.53	2.18	1.79	1.47	1.32	1.86
MMSE	2.78	2.24	1.81	1.53	1.36	1.94
$DAELD_{(BP)}$(s)	2.63	2.27	**1.89**	1.53	1.35	1.93
$DAELD_{(SAE)}$(s)	2.64	2.27	1.87	1.52	1.37	1.94
$DAELD_{(BP)}$(u)	2.78	2.27	1.86	1.57	1.40	1.98
$DAELD_{(SAE)}$(u)	**2.80**	**2.31**	**1.89**	**1.58**	**1.40**	**2.00**
Non-stationary Noises (Babble and Restaurant)						
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
$DAELD_{(BP)}$(s)	2.70	2.35	1.98	1.65	1.46	2.03
$DAELD_{(SAE)}$(s)	**2.75**	**2.40**	**2.01**	**1.68**	**1.48**	**2.06**
$DAELD_{(BP)}$(u)	2.70	2.21	1.85	1.59	1.42	1.95
$DAELD_{(SAE)}$(u)	2.73	2.24	1.87	1.59	1.42	1.97

STOI

	12	6	0	-6	-12	Ave
Stationary Noise (Car and Engine)						
Noisy	**0.91**	**0.82**	0.68	0.54	**0.43**	**0.67**
DDAE	0.81	0.75	0.65	0.51	0.37	0.62
MMSE	**0.91**	**0.82**	0.68	0.53	0.40	**0.67**
$DAELD_{(BP)}$(s)	0.82	0.76	0.67	0.54	0.40	0.64
$DAELD_{(SAE)}$(s)	0.82	0.76	0.67	**0.55**	0.42	0.65
$DAELD_{(BP)}$(u)	0.89	0.81	0.68	0.54	0.41	**0.67**
$DAELD_{(SAE)}$(u)	0.88	0.81	**0.69**	0.54	0.41	**0.67**
Non-stationary Noise (Babble and Restaurant)						
Noisy	**0.93**	**0.85**	**0.74**	0.62	0.52	**0.73**
DDAE	0.82	0.78	0.71	0.61	0.50	0.68
MMSE	0.92	0.84	0.73	0.61	0.50	0.72
$DAELD_{(BP)}$(s)	0.83	0.79	0.73	**0.63**	0.52	0.70
$DAELD_{(SAE)}$(s)	0.82	0.79	0.72	**0.63**	**0.53**	0.70
$DAELD_{(BP)}$(u)	0.90	0.83	0.73	0.61	0.50	0.72
$DAELD_{(SAE)}$(u)	0.89	0.83	0.73	0.61	0.50	0.71

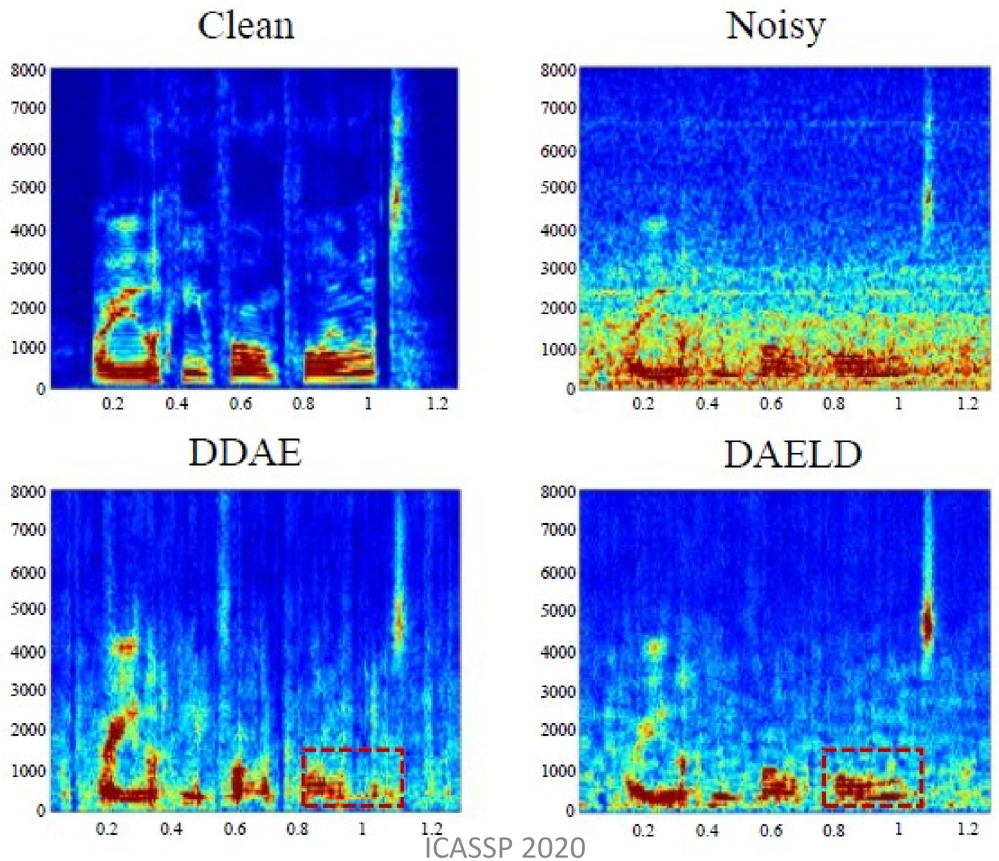
Experiments

- **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