Speech Enhancement Based on Deep Denoising Autoencoder

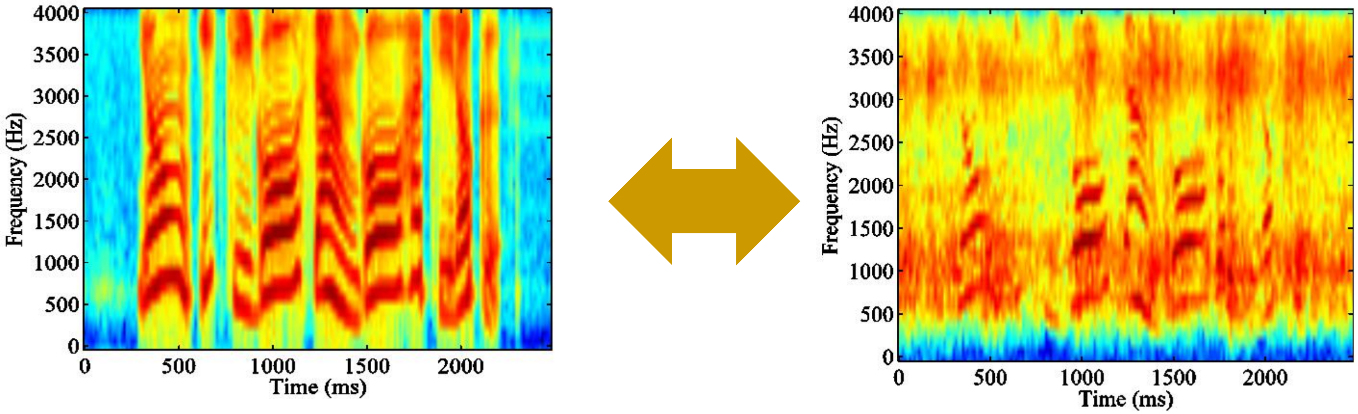
X. Lu*, Y. Tsao, S. Matsuda, C. Hori
xugang.lu@nict.go.jp

National Institute of Information and Communication Technology, Japan.

Interspeech 2013

What is the focus?

- Estimating clean speech spectrum from noisy one



$$f(\cdot):\quad x \;\rightarrow\; y$$

$$\varphi(\cdot):\quad y \;\rightarrow\; x$$

Problem description

Forward problem: $f(\cdot): \; x \rightarrow y$

Clean speech   Noisy speech

Inverse problem: $\varphi(\cdot): \; y \rightarrow x$

Traditional ways

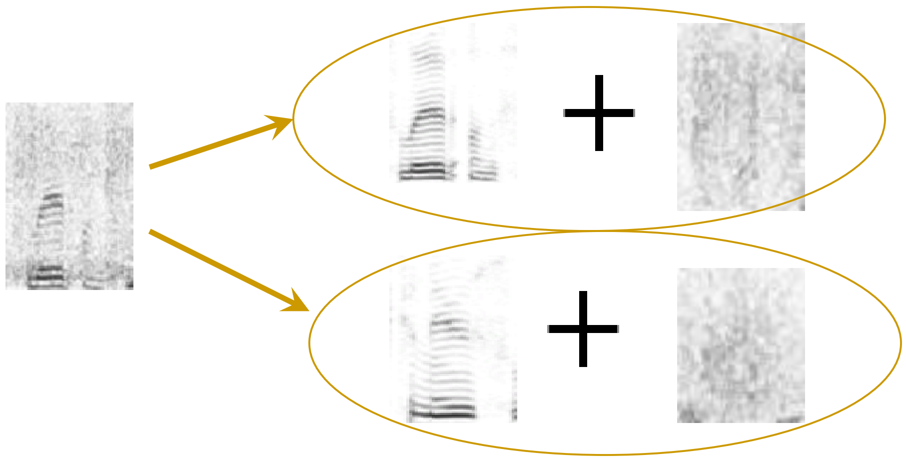
- **Linear or Gaussian assumption**
  - The second order statistic structure

- **Short temporal signal structure**
  - Frame by frame estimation (e.g. 20 ms)

Wiener filtering [P. P. Paliwal et al, 1987];
Signal subspace [P. C. Loizou, 2007];
Minimum mean square error based estimation [Y. Ephraim et al, 1990]

Inverse estimation

Inverse problem: $\varphi(\cdot): \quad y \rightarrow x$



One to many (ill-posed inverse problem)

Our considerations

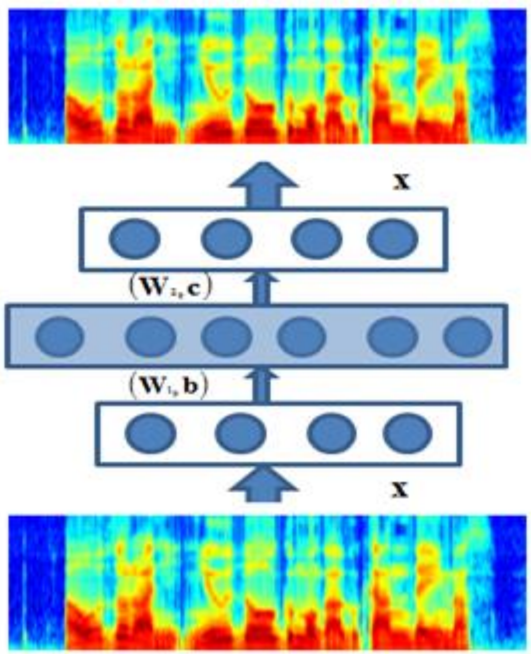
- **One to many (ill-posed inverse problem)**
  - Nonlinear high order statistical structure
  - Long temporal signal structure

Nonlinear mapping

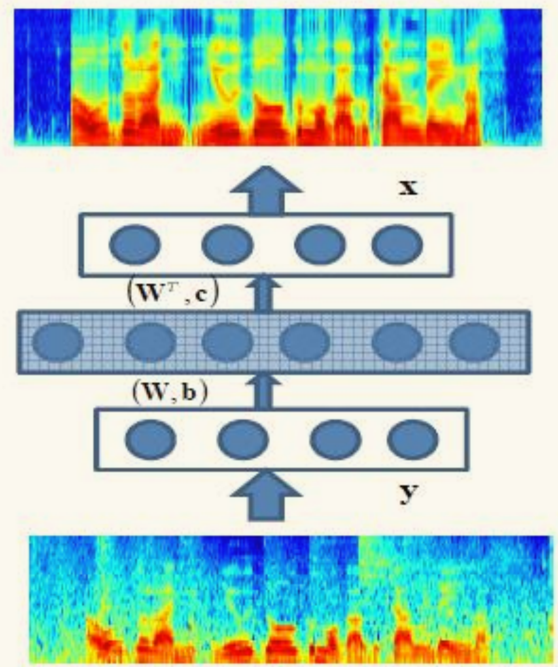
- **Neural network (NN)-Universal approximation**
  - One of the most efficient ways for learning nonlinear mapping functions
- **Deep neural network (DNN)**
  - Better generalization with robust performance than traditional one-hidden-layer NN [Hinton et al, 2006]
  - Successfully used on automatic speech recognition (ASR) [Yu et al, 2009]

→ Learning the inverse denoising function

Autoencoder vs. Denoising autoencoder



Learning speech basis functions
for approximation
(clean-clean speech pairs)



Learning discriminative basis functions
for approximation
(noisy-clean speech pairs)

Problem formulation for denoising AE

Reconstruction error: $$L(\Theta) = \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$$

Objective function: $$J(\Theta) = L(\Theta) + \alpha \|\mathbf{W}\|_2^2 + \beta\rho(h(\mathbf{y}))$$

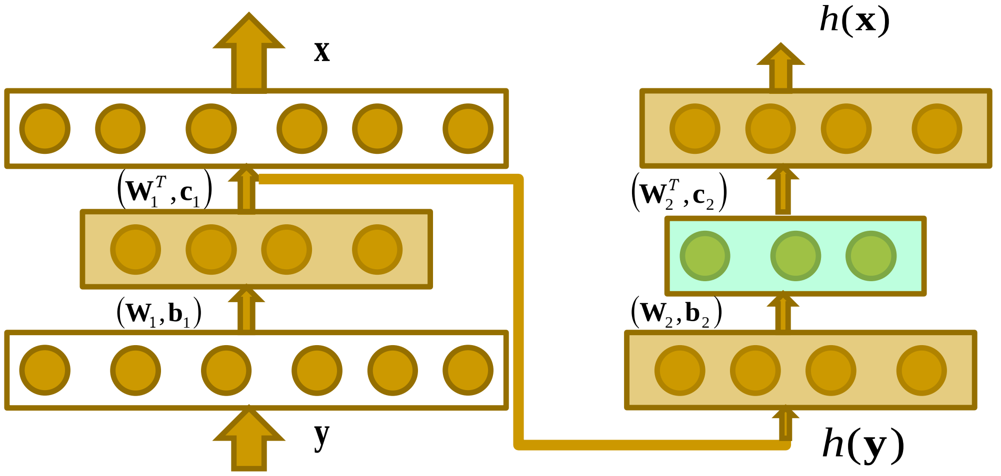
Autoencoder transform: $$h(\mathbf{y}_i) = \sigma(\mathbf{W}_1\mathbf{y}_i + \mathbf{b})$$ $$\hat{\mathbf{x}}_i = \mathbf{W}_2 h(\mathbf{y}_i) + \mathbf{c},$$

Regularization on weights: $$\|\mathbf{W}\|_2^2 = \sum_{i,j} w_{ij}^2.$$

Regularization on neural response: $$\rho(h(\mathbf{y}))$$

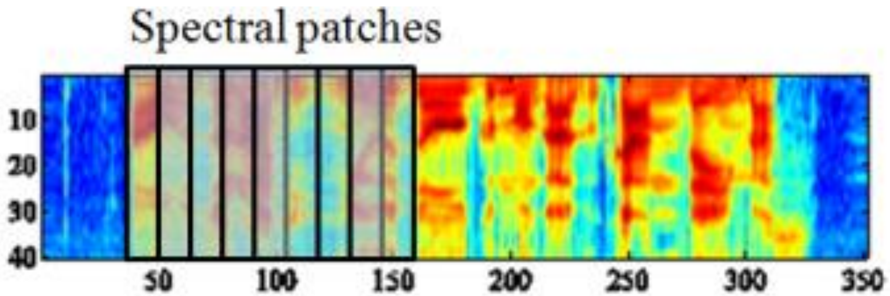
$$\Theta^* \stackrel{\Delta}{=} \arg\min_{\Theta} J(\Theta)$$

How to stack denoising AE to make deep?



Data set and noisy conditions

- Data:
  - Training: 350 clean utterances
  - Testing: 50 utterances
  - Input data: Spectral patches 11 frames (40 Mel bands, 16 ms window size, 8 ms shift)



- Noisy condition
  - Factory and car noise with SNR 0, 5, 10 dB

Evaluation criteria

- Evaluation criteria
  - Noise reduction

$$\text{Reduct} \triangleq \frac{1}{N * d} \sum_{i=1}^{N} |\hat{\mathbf{x}}_i - \mathbf{y}_i|$$

  - Speech distortion

$$\text{Dist} \triangleq \frac{1}{N * d} \sum_{i=1}^{N} |\hat{\mathbf{x}}_i - \mathbf{x}_i|$$

  - Perceptual evaluation of speech quality (PESQ) (0.5-4.5)

Effect of training data set size

Hidden layer size

500

Training set size	10 k	40 k	80 k
Reduct (dB)	**2.01**	1.94	1.93
Dist (dB)	0.62	0.47	**0.43**
PESQ	2.64	3.36	**3.52**

300

Training set size	10 k	40 k	80 k
Reduct (dB)	**2.01**	1.94	1.93
Dist (dB)	0.61	0.47	**0.44**
PESQ	2.77	3.32	**3.44**

→The much the better

100

Training set size	10 k	40 k	80 k
Reduct (dB)	**1.99**	1.94	1.93
Dist (dB)	0.60	0.48	**0.47**
PESQ	2.80	3.30	**3.33**

Training data set size

Effect of hidden layer size

80000 spectral patches for training

hidsize	100	300	500
Reduct (dB)	**1.93**	**1.93**	**1.93**
Dist (dB)	0.47	0.44	**0.43**
PESQ	3.33	3.44	**3.52**

→The larger the better if the training data size is large enough

Effect of hidden depth

Hidden layer size

500

hidsize*layer	500*1	500*2	500*3
Reduct (dB)	**1.93**	1.91	1.92
Dist (dB)	0.43	**0.40**	0.42
PESQ	3.52	**3.61**	3.52

300

hidsize*layer	300*1	300*2	300*3
Reduct (dB)	**1.93**	1.92	1.92
Dist (dB)	0.44	**0.40**	**0.40**
PESQ	3.44	3.52	**3.61**

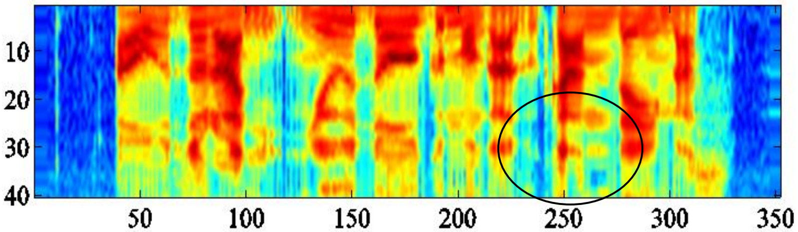
100

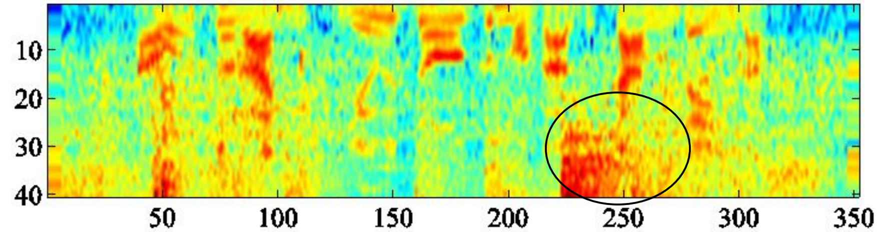
hidsize*layer	100*1	100*2	100*3
Reduct (dB)	**1.93**	**1.93**	**1.93**
Dist (dB)	0.47	0.44	**0.43**
PESQ	3.33	**3.39**	**3.39**

Deep

→The deep the better if the training data size is large enough

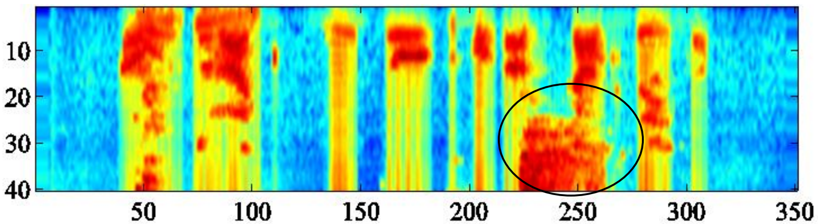
MMSE vs DAE

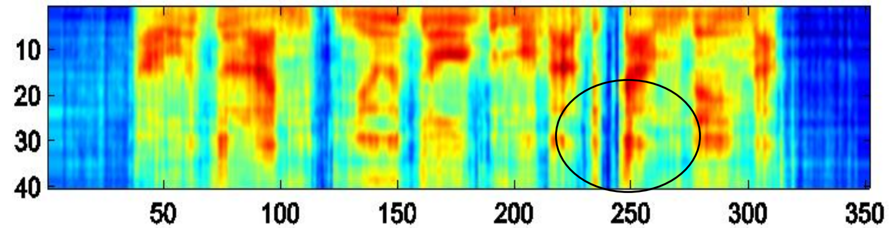




Clean

Factory noise SNR =10dB





MMSE denoising

DAE denoising

The second order statistic based methods try to keep large energy components.

Quantitative evaluations

Evaluations	Noise reduction			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	2.35	**2.72**	**1.05**	0.83
5	2.08	**2.32**	**0.92**	0.63
10	1.84	**1.93**	**0.82**	0.47

Evaluations	Speech distortion			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	1.56	**0.59**	0.63	**0.27**
5	1.28	**0.47**	0.59	**0.24**
10	1.05	**0.43**	0.57	**0.21**

Evaluations	PESQ			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	1.22	**2.82**	2.90	**3.98**
5	1.73	**3.19**	3.05	**4.09**
10	2.15	**3.39**	3.17	**4.18**

DAE outperforms MMSE in almost all conditions (exception For noise reduction in car noise Condition)

Summary and conclusion

- Learning the discriminative mapping function between noisy and clean speech (explores nonlinear high-order statistical structure)
- Long temporal structure is incorporated to implicitly regularize the ill-posed inverse problem
- Deep makes better performance than shallow, but enough training data is required
- →How to incorporate speech temporal hierarchical structure in the network
- →How to regularize the network, e.g., sparse constrain

Last slide

- **Thanks for your attention**