Speech Enhancement Based on Deep Neural Networks

Chin-Hui Lee
School of ECE, Georgia Tech
chl@ece.gatech.edu

Joint work with Yong Xu and Jun Du at USTC

Outline and Talk Agenda

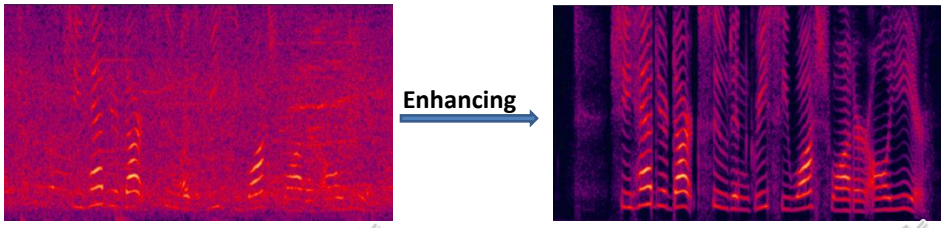
- In Signal Processing Letter, Jan. 2014
- Speech enhancement
  - Background
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural network s
  - SE-DNN: background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

Outline

- Speech enhancement
  - Background
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural network s
  - SE-DNN: background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

Speech Enhancement

- Speech enhancement aims at improving the intelligibility and/or overall perceptual quality of degraded speech signals using audio signal processing techniques
- One of the most addressed classical SP problems in recent years



Noisy speech, Exhibition, SNR=5dB

Clean speech, 8kHz

Speech Enhancement Applications

Mobile phone/ communication

Hearing aids

Security monitoring/ intelligence

Robust speech /speaker/language recognition, etc.

Noise in Speech Enhancement

1. Additive noise:

$$y(t) = x(t) + n(t) \xrightarrow{\text{STFT}} Y(n,d) = X(n,d) + N(n,d)$$

Focused

2. Convolutional noise:

$$y(t) = x(t) * h(t)$$

3. Mixed noise:

$$y(t) = x(t) * h(t) + n(t)$$

$$y(t) = [x(t) + v(t)] * h(t) + n(t)$$

Outline

- Speech enhancement
  - Background
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural network s
  - SE-DNN: background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

Conventional Speech Enhancement

- **Classified by the number of microphones**

1. **Single channel speech enhancement methods**
   - Time and frequency information

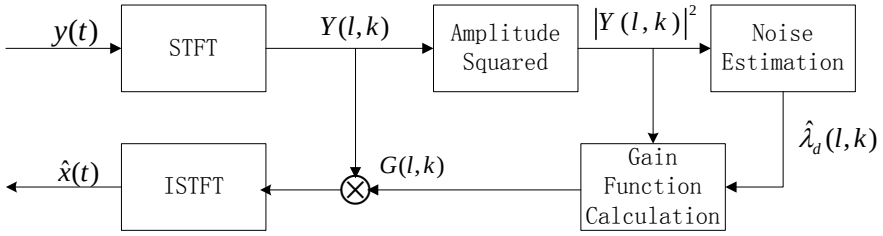
Focused

2. **Microphone based speech enhancement methods**
   - Time and frequency information
   - Spatial information
   - Microphone arrays

Conventional Single Channel SE

1. Spectrum Subtraction, **SS**
2. Wiener Filtering
3. Minimum Mean Square Error Short-Time Spectral Amplitude, **MMSE-STSA**
4. MMSE Log Spectral Amplitude, **MMSE-LSA**
5. Optimally Modified LSA, **OM-LSA**
6. ……

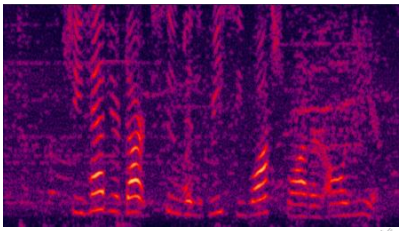
Conventional Single Channel SE



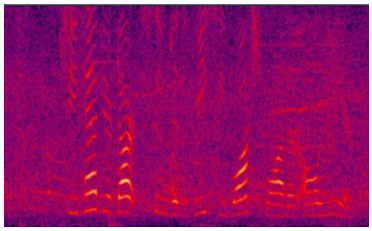
1. STFT on the noisy signal y, get the time-frequency signal Y
2. estimate the variance of noise $\hat{\lambda}_d$
3. estimate all of the parameters (prior SNR γ, posterior SNR ξ and the speech presence probability, etc.) needed by the gain function
4. calculate the gain function G
5. multiply Y with G, then ISTFT to obtain the enhanced signal (using the phase of noisy speech)

Conventional Single Channel SE: Issues

1. Musical noise:



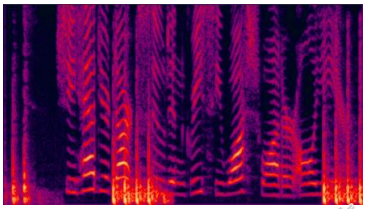
Enhanced by SS



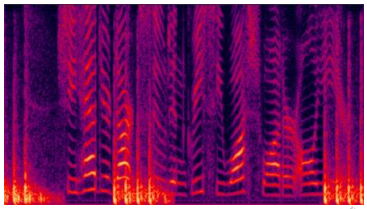
Noisy speech, exhibition noise, SNR=5dB

Conventional Single Channel SE: Issues

2. Difficult to deal with the highly non-stationary noise:



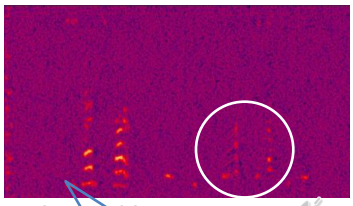
Enhanced by OM-LSA



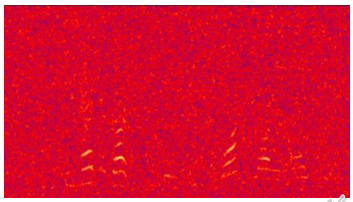
Noisy, Machine Gun, SNR=-5dB

Conventional Single Channel SE: Issues

3. Difficult to deal with the low SNR cases:



Enhanced by OM-LSA



Noisy, AGWN, SNR=-5dB

Easily distorting the speech components and has much residual noise

Conventional Single Channel SE: Issues

4. Introducing some non-linear distortion which is fatal for the back-end recognition, coding, etc.

5. Learning from human listening experience: many years of exposure to clean speech and noise

Conventional Single Channel SE: Issues

- **Analysis of these disadvantages**

$$y(t) = x(t) + n(t) \xrightarrow{\text{STFT}} Y(n,d) = X(n,d) + N(n,d)$$

Gaussian assumptions
Un-correlated assumptions

$$E\{|Y(n,\mathrm{d})|^2\} = E\{|X(\mathrm{n},\mathrm{d})|^2\} + E\{|N(\mathrm{n},\mathrm{d})|^2\}$$

$$\lambda(\mathrm{n},\mathrm{d}) = \lambda_x(\mathrm{n},\mathrm{d}) + \lambda_d(\mathrm{n},\mathrm{d})$$

Binary model assumptions:

$$H_0(\mathrm{n},\mathrm{d}) : Y(\mathrm{n},\mathrm{d}) = N(\mathrm{n},\mathrm{d})$$

$$H_1(\mathrm{n},\mathrm{d}) : Y(\mathrm{n},\mathrm{d}) = X(\mathrm{n},\mathrm{d}) + N(\mathrm{n},\mathrm{d})$$

With these inaccurate assumptions, it is hard for conventional methods to deliver a satisfactory performance!

Outline

- Speech enhancement
  - Background
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural networks
  - 2.1 Background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

DNN-Based Speech Enhancement

- The signal model of the additive noise:

$$y(t) = x(t) + n(t) \xrightarrow{\text{STFT}} Y(n,d) = X(n,d) + N(n,d)$$

➢ Many enhancement methods are derived from this signal model, however, most of them assume that $X(\mathrm{n}, d)$ is described by a Gaussian mixture model (GMM) and $N(\mathrm{n}, \mathrm{d})$ is a single Gaussian. The relationship between the speech and noise is complicated in some non-linear fashion.

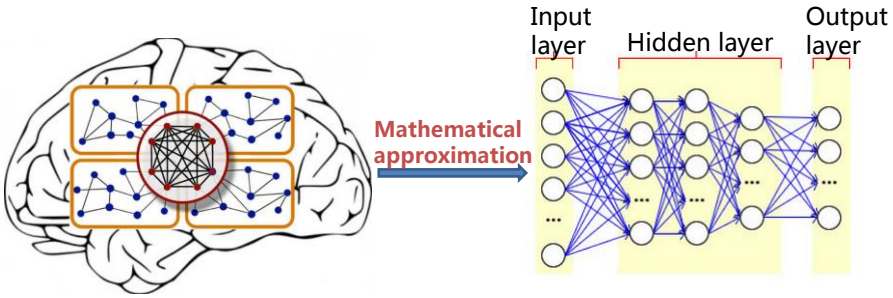
- DNN assumes a nonlinear mapping function *F*:

$$\mathrm{X} = F(Y)$$

➢ Construct the stereo data based on the additive noise model
➢ No special assumptions were made in the DNN based SE method

Deep Neural Network: Overview

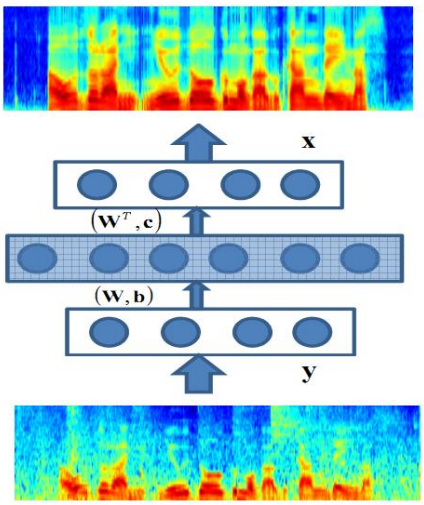
1. Hinton proposed the unsupervised Restricted Boltzmann Machine (RBM) based pre-training in 2006
2. In 2012, MSR, Google and IBM got a great success in large vocabulary continuous speech recognition using DNNs
3. Later, DNNs were adopted in many speech-related tasks



DNN Based SE: Related Work

4. In 2013, Xugang Lu proposed deep de-noising auto-encoder based speech enhancement

$$h(\mathbf{y}_i) = \sigma(\mathbf{W}_1 \mathbf{y}_i + \mathbf{b})$$
$$\hat{\mathbf{x}}_i = \mathbf{W}_2 h(\mathbf{y}_i) + \mathbf{c},$$
$$\mathbf{W}_1 = \mathbf{W}_2{}^T = \mathbf{W}$$
$$L(\Theta) = \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2,$$



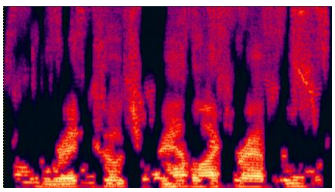
X.-G. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising Auto-Encoder," Proc. Interspeech, pp. 436-440, 2013.

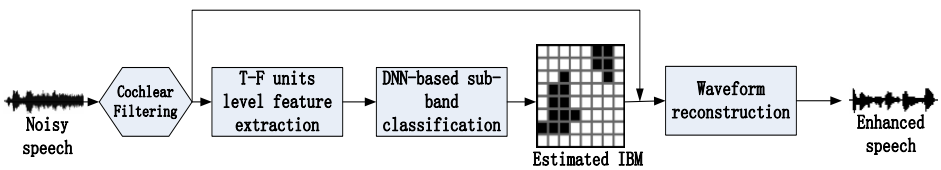
DNN Based SE: Related Work

6. In 2013, Deliang Wang proposed using DNN to classify the time-Frequency bins into 1/0 units (ideal binary mask)

$$IBM(t,f) = \begin{cases} 1, & \text{if } SNR(t,f) > LC \\ 0, & \text{otherwise} \end{cases}$$

IBM-DNN enhanced





Y. X. Wang and D. L. Wang, "Towards scaling up classification based speech separation," IEEE Trans. on Audio, Speech and Language Processing, Vol. 21, No. 7, pp. 1381-1390, 2013.

DNN Based SE: Issues

- **Advantages of SE-DNN**

1. The complicated relationship between noisy and clean speech could be automatically learnt
2. The deep architecture could well fit the non-linear relationship for regression function approximation
3. The highly non-stationary noise could be well suppressed in the off-line learning framework
4. Nearly no Gaussian or independent assumptions
5. Nearly no empirical thresholds to avoid the non-linear distortion in SS-based speech enhancement

DNN Based SE: Issues

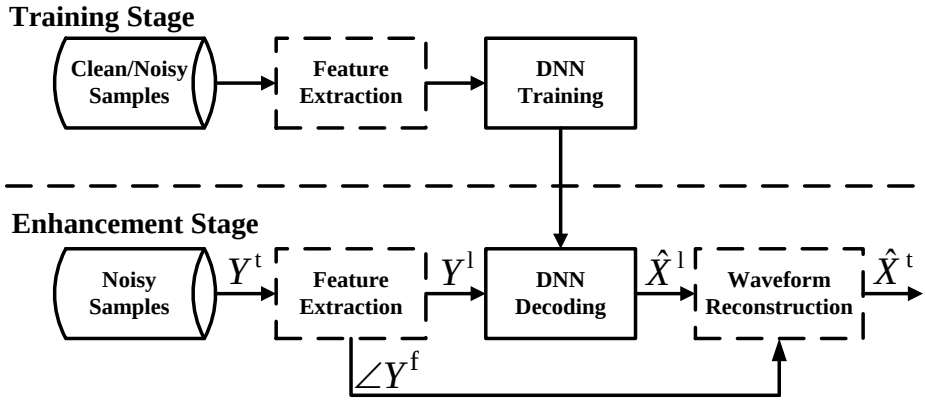
- **Difficulties in using SE-DNN**

1. Which domain is suitable for DNN-based mapping?
2. The generalization capacity to unknown environments, especially for unseen noise types?
3. How to perform noise adaptation? – robustness issue

Outline

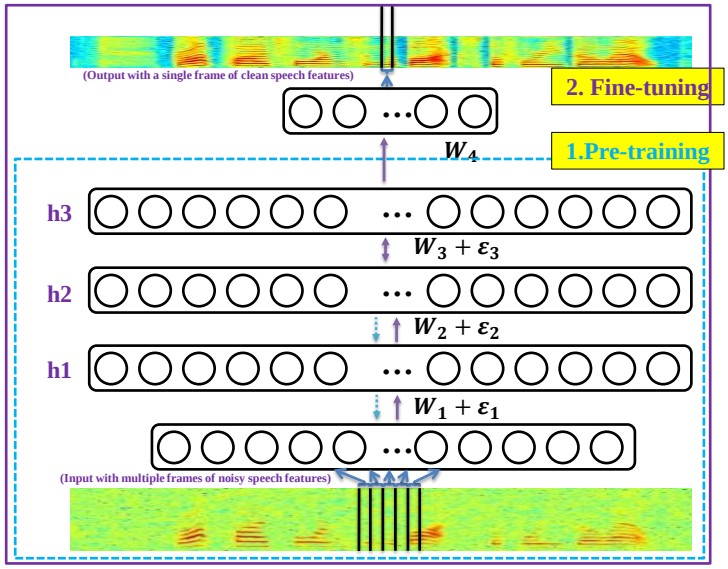
- Speech enhancement
  - Background
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural networks
  - SE-DNN: background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

System Overview



1. Feature extraction: log-power spectra
2. Waveform reconstruction: overlap-add algorithm
3. DNN Training: RBM pre-training + back-propagation fine-tuning

DNN Training



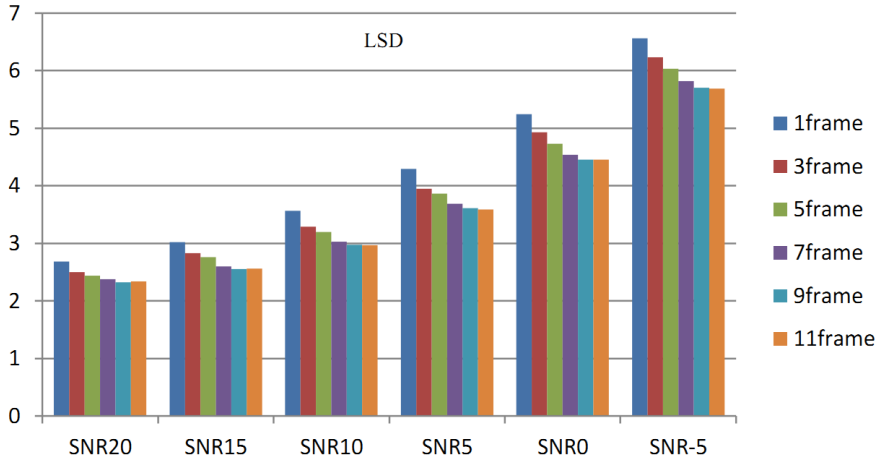
1. MMSE-based object function: $E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(\hat{X}_n^d(\boldsymbol{W},\boldsymbol{b}) - X_n^d)^2 + \lambda\|\boldsymbol{W}\|_2^2$ 30

Experimental Setup

1. Clean speech set: TIMIT corpus, 8kHz
2. Noise set: Additive Gaussian White Noise (AGWN), Babble, Restaurant, Street
3. Signal to Noise ratios: Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB
4. Construct 100 hours multi-condition training data
5. Test set: 200 randomly selected utterances from TIMIT test set, and two unseen noise types: *Car* and *Exhibition*
6. Three objective quality measures: segmental SNR (SegSNR in dB), log-spectral distortion (LSD in dB), perceptual evaluation of speech quality (PESQ)
7. Standard DNN configurations: 11 frames expansion, 3 hidden layers and 2048 hidden units for each
8. Competing methods: improved version of the optimally modified log-spectral amplitude (OM-LSA), denoted as log-MMSE (L-MMSE)

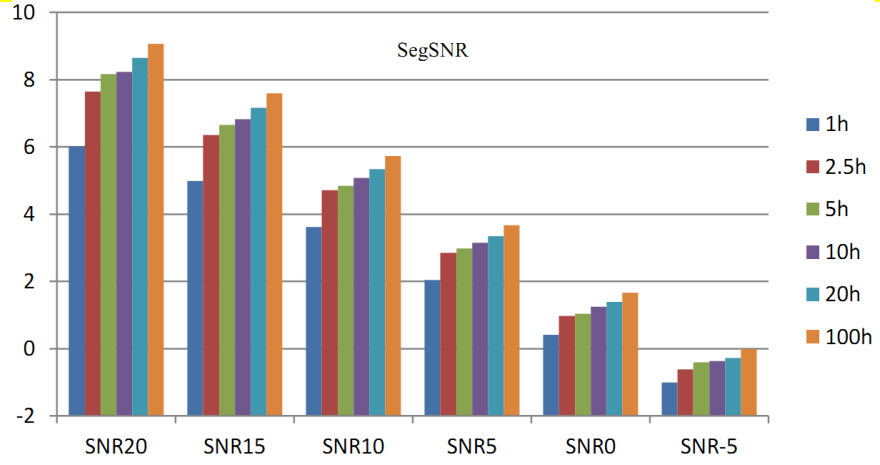
Baseline Experimental Results: I

1. Average **LSD** using input with different acoustic context on the test set at different SNRs across four noise types: A good choice: 11 frames



Baseline Experimental Results: II

2. Average **SegSNR**s using different training set size on the test set at different SNRs across four noise types: still improving with 100 hours



Baseline Experimental Results: III

3. Average **PESQ**s among methods on the test set at different SNRs with four noise types. The subscript of DNN_l represents l hidden layers

	Noisy	L-MMSE	SNN*	DNN_1	DNN_2	DNN_3	DNN_4
SNR20	2.99	3.32	3.48	3.46	3.59	**3.6**	3.59
SNR15	2.65	2.99	3.26	3.24	3.35	**3.36**	**3.36**
SNR10	2.32	2.65	2.99	2.97	3.08	**3.1**	3.09
SNR5	1.98	2.3	2.68	2.65	2.76	**2.78**	**2.78**
SNR0	1.65	1.93	2.32	2.29	2.38	**2.41**	**2.41**
SNR-5	1.38	1.55	1.92	1.89	1.95	**1.97**	**1.97**
Ave	2.16	2.46	2.78	2.75	2.85	**2.87**	**2.87**

*Shallow Neural Network (SNN) has the same computation complexity with DNN_3

- Deep structure can get better performance compared with SNN.
- DNN_3 outperforms the L-MMSE method, especially at low SNRs.

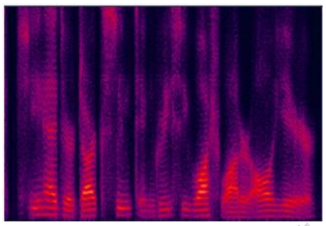
Baseline Experimental Results: IV

4. PESQs among Noisy, L-MMSE, SNN and DNN_3 at different SNRs in mismatch environments under *Car* (A) and *Exhibition* (B) noises,

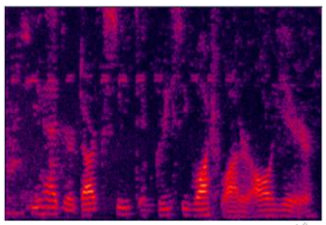
	Noisy		L-MMSE		SNN		DNN_3	
	A	B	A	B	A	B	A	B
SNR20	3.15	2.89	3.52	3.19	3.43	3.24	**3.58**	**3.30**
SNR15	2.81	2.55	3.23	2.85	3.19	2.96	**3.31**	**3.01**
SNR10	2.47	2.21	2.89	2.51	2.93	2.66	**3.03**	**2.69**
SNR5	2.14	1.87	2.57	2.11	2.60	2.30	**2.71**	**2.33**
SNR0	1.81	1.56	2.21	1.72	2.24	1.92	**2.35**	**1.93**
SNR-5	1.52	1.28	1.82	1.34	1.85	1.52	**1.96**	**1.54**
Ave	2.32	2.06	2.70	2.29	2.71	2.43	**2.83**	**2.47**

- SE-DNN has a generalization capacity to unseen noise types. It can be further strengthened by adding more noise types in training

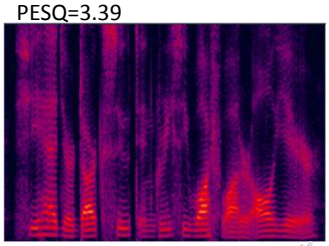
Baseline Experimental Results V



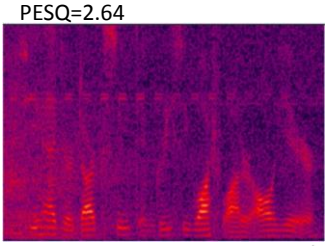
DNN enhanced PESQ=3.39



L-MMSE enhanced PESQ=2.64



Clean, PESQ=4.5

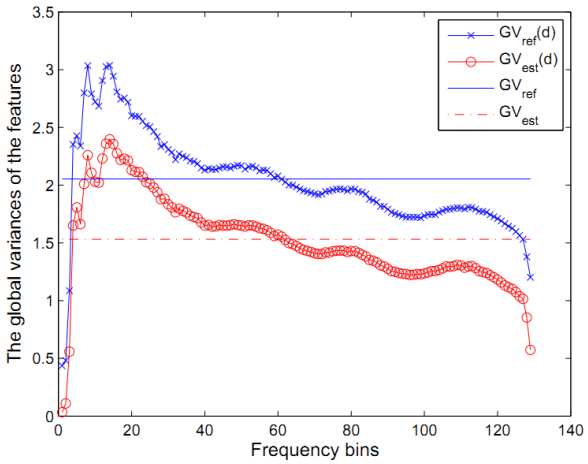


Noisy, street, SNR=10dB, PESQ=2.2

DNN can deal with non-stationary noise given some noise characteristics to learn.

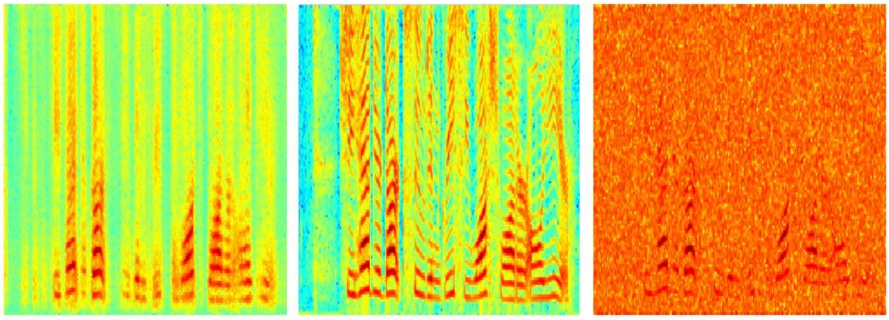
More demos could be found at: http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html

Over-smoothing with SE-DNN (1/2)



1. The global variances of the training set were shown. $GV_{ref}(d)$ and $GV_{est}(d)$ represented the d-th dimension of the global variance of the reference features and the estimation features, respectively. And the corresponding dimension-independent variances were denoted as GV_{ref} and GV_{est}

Over-smoothing with SE-DNN (2/2)



DNN enhanced | Clean | AGWN, SNR=0dB

2. The formant peaks were suppressed, especially in the high frequency band which leads to muffled speech

Methods to Address Over-smoothing

- **The definition of global variance equalization factors:**

$$\beta = \sqrt{\frac{GV_{ref}}{GV_{est}}} \qquad \alpha(d) = \sqrt{\frac{GV_{ref}(d)}{GV_{est}(d)}} \qquad \bar{\alpha} = \frac{1}{D}\sum_{d=1}^{D}\alpha(d)$$

- **Proposed method 1:** post-processing

$$\hat{X}''(d) = \hat{X}(d) * \eta * v(d) + m(d)$$

where $m(d)$ and $v(d)$ are the *d*-th component of the mean and variance of input noisy features, respectively. And η could be β , $\alpha(d)$ or $\bar{\alpha}$

- **Proposed method 2:** post-training

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(\hat{X}_n^d(\boldsymbol{W},\boldsymbol{b}) - \eta * X_n^d)^2 + \lambda\|\boldsymbol{W}\|_2^2$$

GV Experimental Results (1/2)

PESQ results of the L-MMSE method and DNN baseline, compared with different post-processing and post-training schemes using β, $\alpha(d)$ and $\bar{\alpha}$ on the test set at different SNRs across four noise types

	L-MMSE	DNN	Post-processing			Post-training		
			β	$\alpha(d)$	$\bar{\alpha}$	β	$\alpha(d)$	$\bar{\alpha}$
SNR20	3.32	3.60	3.71	3.69	3.71	**3.72**	3.70	**3.72**
SNR15	2.99	3.36	3.47	3.45	3.48	3.48	3.46	**3.49**
SNR10	2.65	3.10	3.18	3.17	3.19	**3.20**	3.18	**3.20**
SNR5	2.30	2.78	2.85	2.84	2.85	**2.86**	2.85	**2.86**
SNR0	1.93	2.41	2.45	2.44	2.45	2.46	2.46	**2.47**
SNR-5	1.55	1.97	1.99	1.99	1.99	2.01	2.00	**2.02**
Ave	2.46	2.87	2.94	2.93	2.94	2.95	2.94	**2.96**

1. $\bar{\alpha}$ is better than β, and $\alpha(d)$ is the worst, indicating that the degree of over-smoothing on different dimensions was similar
2. Equalization operations were much more beneficial for high SNRs
3. Post-training was a little better than Post-processing

GV Experimental Results (2/2)

PESQ results in unseen environments under *Car* and *Exhibition* noises, labeled as case A and B, respectively. The DNN baseline was compared with the L-MMSE method and the proposed two global variance equalization approaches using the factor $\bar{\alpha}$:

	L-MMSE		DNN		Post-processing		Post-training	
	A	B	A	B	A	B	A	B
SNR20	3.52	3.19	3.58	3.30	3.72	**3.46**	**3.73**	**3.46**
SNR15	3.23	2.85	3.31	3.01	3.46	3.15	**3.47**	**3.16**
SNR10	2.89	2.51	3.03	2.69	**3.16**	2.81	**3.16**	**2.82**
SNR5	2.57	2.11	2.71	2.33	2.81	2.42	**2.82**	**2.43**
SNR0	2.21	1.72	2.35	1.93	**2.44**	2.00	**2.44**	**2.01**
SNR-5	1.82	1.34	1.96	1.54	**2.04**	1.59	**2.04**	**1.60**
Ave	2.70	2.29	2.83	2.47	**2.94**	2.57	**2.94**	**2.58**

1. GV equalization is slightly more effective for unseen noise types
2. Post-training was a little better than Post-processing

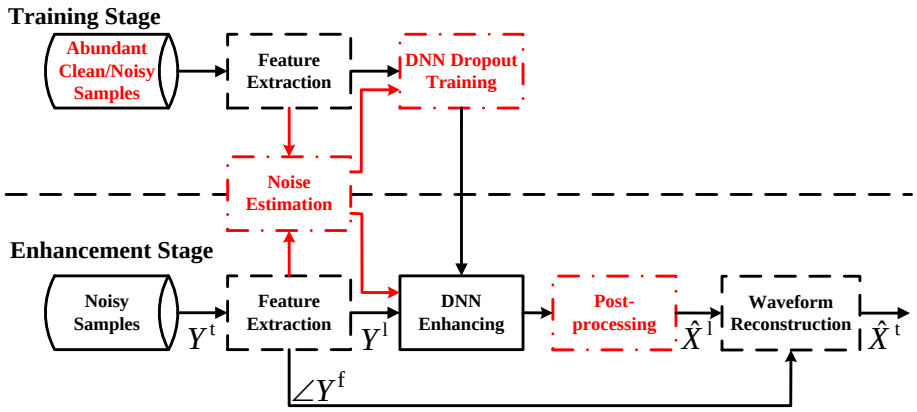
Summary I: DNN-SE Properties

1. SE-DNN achieves better performance than traditional single channel speech enhancement methods (e.g., OM-LSA), especially for low SNRs and non-stationary noise.
2. A large training set is crucial to learn the rich structure of DNN
3. Using more acoustic context information improves performance and makes the enhanced speech less discontinuous
4. Multi-condition training can deal with speech enhancement of new speakers, unseen noise types, various SNR levels under different conditions, and even cross-language generalization.
5. The over-smoothing problem in SE-DNN could be alleviated using two global variance equalization methods, and the equalization factor tends to be independent with the dimension
6. The global variance equalization was much more helpful for unseen noise types

Outline

- Speech enhancement task
  - Backgrounds
  - Conventional speech enhancement methods
- Speech enhancement based on deep neural networks
  - SE-DNN: background
  - DNN baseline and enhancement
  - Noise-universal SE-DNN

Noise Universal SE-DNN



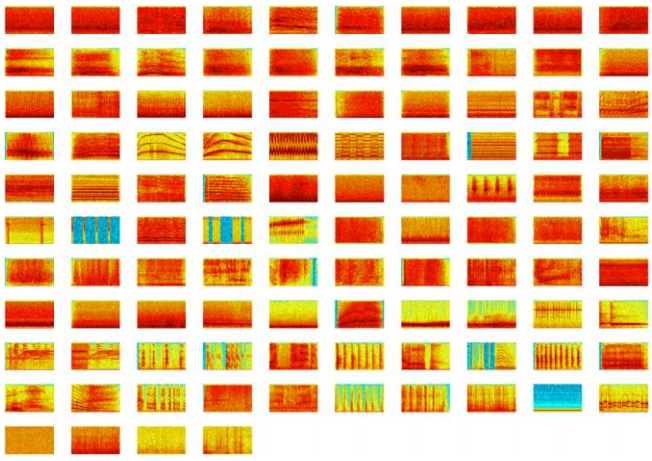
*The global variance equalization was adopted in the post-processing.

Noise Universal SE-DNN

1. DNN to learn the characteristics of many noise types

❑ Classifications :
Crowding、machine、transportation、animal、nature、human, etc.

alarm cry



G. Hu, 100 non-speech environmental sounds, 2004.
http://www.cse.ohiostate.edu/pnl/corpus/HuCorpus.html.

Noise Universal SE-DNN

2. Noise aware training

- Using the average feature of the first *T* frames of the current utterance to help DNN to learn a "noise code"

$$\mathbf{V}_n = [\mathbf{Y}_{n-\tau}, ..., \mathbf{Y}_{n-1}, \mathbf{Y}_n, \mathbf{Y}_{n+1}, ..., \mathbf{Y}_{n+\tau}, \hat{\mathbf{Z}}_n]$$

$$\hat{\mathbf{Z}}_n = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Y}_t$$

3. Dropout learning

- Randomly disable some units of the input layer and the hidden layers to improve generalization capacity (0.1 for the input layer and 0.2 for the hidden layers)
- Regulation technology and avoid over-fitting

Experimental Setup

1. Clean speech training set: TIMIT corpus, 8kHz
2. Noise training set: 104 noise types
3. Signal to Noise ratios: Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB
4. Construct 100/625 hours multi-condition training data
5. Test set: 200 randomly selected utterances from the TIMIT test set corrupted by the noises from the NOISEX-92 database
6. Three objective quality measures: segmental SNR (SegSNR in dB), log-spectral distortion (LSD in dB), perceptual evaluation of speech quality (PESQ)
7. Standard DNN configurations: 11 frames context expansion, 3 hidden layers and 2048 hidden units for each hidden layer
8. Competing state-of-the-art methods: improved version of the optimally modified log-spectral amplitude (OM-LSA), denoted as log-MMSE (L-MMSE)

Enhanced Experimental Results: I

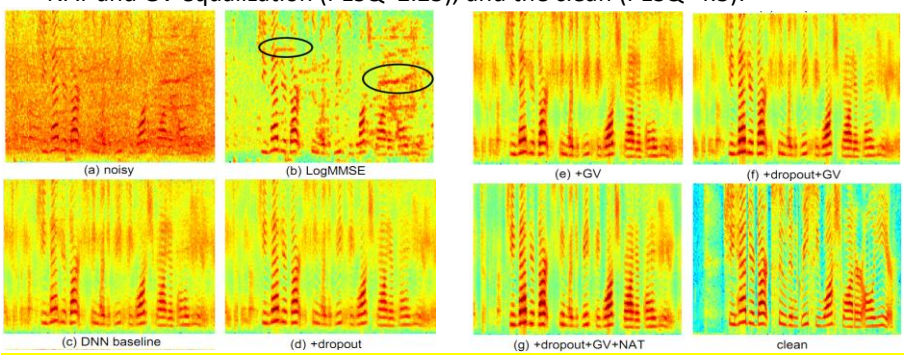
- **LSD** comparison between models trained with four noise types (4NT) and 104 noise types (104NT) on the test set at different SNRs of three **unseen** noise environments :

	Exhibition		Destroyer engine		HF channel	
	4NT	104NT	4NT	104NT	4NT	104NT
SNR20	2.55	**2.24**	2.51	**2.25**	3.09	**2.39**
SNR15	3.14	**2.73**	2.91	**2.73**	4.53	**3.26**
SNR10	4.42	**3.70**	3.68	**3.58**	6.85	**4.96**
SNR5	6.53	**5.28**	4.97	**4.90**	9.85	**7.44**
SNR0	9.44	**7.60**	6.91	**6.65**	13.32	**10.43**
SNR-5	12.96	**10.62**	9.48	**8.75**	16.98	**13.64**
Ave	6.51	**5.36**	5.08	**4.81**	9.11	**7.02**

Abundance of noise types is important to predict unseen noise types

Enhanced Experimental Results: II

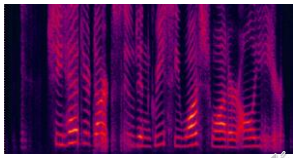
- Spectrograms of an utterance tested with *Exhibition* noise at SNR= 5dB. (a) noisy (PESQ=1.42), (b) LogMMSE (PESQ=1.83), (c) DNN baseline (PESQ=1.87), (d) improved by dropout (PESQ=2.06), (e) improved by GV equalization (PESQ=2.00), (f) improved by dropout and GV (PESQ=2.13), (g) jointly improved by dropout, NAT and GV equalization (PESQ=2.25), and the clean (PESQ=4.5):



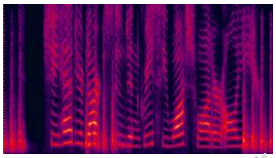
1. SE-DNN can suppress the highly non-stationary noise and get less residual noise
2. Dropout and NAT can reduce noise while GV equalization can brighten speech

Enhanced Experimental Results: III

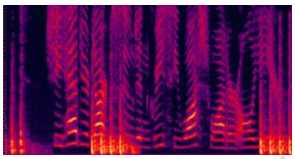
- Spectrograms of an utterance with machine gun noise at SNR= -5dB: with 104-noise DNN enhanced (upper left, PESQ=2.78), Log-MMSE enhanced (lower left, PESQ=1.86), 4-noise DNN enhanced (upper right, PESQ=2.14), and noisy speech (lower right, PESQ=1.85):



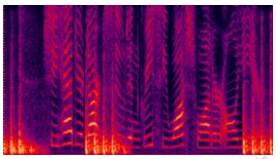
104NT-DNN enhanced **PESQ=2.78**



4NT-DNN enhanced PESQ=2.14



Log-MMSE enhanced PESQ=1.86



noisy，machine gun，SNR=-5dB PESQ=1.85

Even the 4NT-DNN is much better than LogMMSE, SE-DNN can suppress highly non-stationary noise

Enhanced Experimental Results: IV

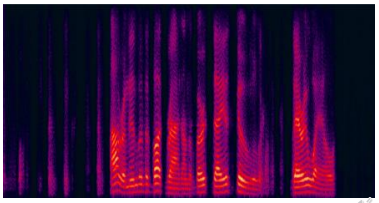
- Average **PESQ** among LogMMSE, DNN baseline with 100 hours data, improved DNN with 100 hours data and improved DNN with 625 hours data on the test set at different SNRs across the whole 15 **unseen** noise types in the NOISEX-92 database:

	Noisy	LogMMSE	100h-baseline	100h-impr	625h-impr
SNR20	3.21	3.60	3.62	3.77	**3.80**
SNR15	2.89	3.33	3.39	3.58	**3.60**
SNR10	2.57	3.02	3.13	3.33	**3.36**
SNR5	2.24	2.66	2.85	3.05	**3.08**
SNR0	1.91	2.25	2.52	2.71	**2.74**
SNR-5	1.61	1.80	2.16	**2.31**	**2.31**
Ave	2.40	2.78	2.94	3.12	**3.15**

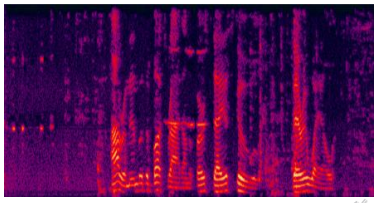
1. A good generalization capacity to unseen noise can be obtained.
2. SE-DNN outperformed the Log-MMSE, especially at low SNRs

Enhanced Experimental Results: V

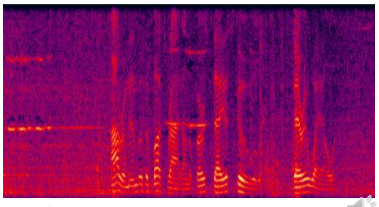
- Spectrograms of a noisy utterance extracted from the movie ***Forrest Gump*** with: improved DNN (upper left), Log-MMSE (upper right), and noisy speech (bottom left): with real-world noise never seen



Universal SE-DNN enhanced



Log-MMSE enhanced



Noisy

- Good generalization capacity to real-world noisy speech
- Could be further improved by adding more varieties of clean data into the training set

Summary II: Noise-Universal DNN

1. Noise aware training (NAT) and dropout learning could suppress more residual noise
2. GV equalization could highlight the speech spectrum to get a better hearing perception
3. The generalization capacity to unseen noise types could be strengthened by adopting more noise types in the training set
4. Noise-universal SE-DNN was also effective in dealing with noisy speech recorded in the real world
5. The generalization capacity could be further improved by adding clean speech data (encompassing different languages, various speaking styles, etc.) into the training set
6. Future work: DNN adaption and other objective functions

Other Recent Efforts

1. Demos: http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html
2. Speech separation: DNN-based semi-supervised speech separation works better than state-of-the-art supervised speech separation (paper submitted to Interspeech2014
3. Dual-Output DNN for separation (submitted to ISCSLP2014)
4. Robust speech recognition: better results than state-of-the-art with only DNN-based pre-processing, additional compensation can be added later (paper submitted to Interspeech2014)
5. Transfer language learning for DNN (submitted to ISCSLP2014)
6. DNN-based bandwidth expansion works better than all other state-of-the-art techniques (submitted to publication)

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 32, No.6, pp. 1109-1121, 1984.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,"IEEE Trans. on Speech and Audio Processing, Vol. 11, No. 5, pp. 466-475, 2003.

[3] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in Handbook of Neural Networks for Speech Processing, Edited by Shigeru Katagiri, Artech House, Boston, 1998.

[4] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "An experimental study on speech enhancement based on deep neural net-works," IEEE Signal Processing Letters, Vol. 21, No. 1, pp. 65-68, 2014.

[5] B.-Y. Xia and C.-C. Bao, "Speech enhancement with weighted denoising Auto-Encoder," Proc. Interspeech, pp. 3444-3448, 2013.

[6] X.-G. Lu and Y. Tsao and S. Matsuda and C. Hori,"Speech enhancement based on deep denoising Auto-Encoder," Proc. Interspeech, pp. 436-440, 2013.

[7] Y. X. Wang and D. L. Wang, "Towards scaling up classification based speech separation," IEEE Trans. on Audio, Speech and Language Processing, Vol. 21, No. 7, pp. 1381-1390, 2013.

[8] G. Hu, 100 nonspeech environmental sounds, 2004. http://www.cse.ohiostate.edu/pnl/corpus/HuCorpus.html.

[9] S. I. Tamura, "An analysis of a noise reduction neural network," Proc. ICASSP, pp. 2001-2004, 1989.

[10] F. Xie and D. V. Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," Proc. ICASSP, pp. 53-56, 1994.

[11] B.-Y. Xia and C.-C. Bao, "Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification," Speech Communication, V. 60, P. 13–29, 2014.