

Acoustic Signal Processing Based on Deep Neural Networks

Chin-Hui Lee

School of ECE, Georgia Tech

chl@ece.gatech.edu

Joint work with Yong Xu, Yanhui Tu, Qing Wang, Tian Gao, Jun Du, LiRong Dai

Outline and Talk Agenda

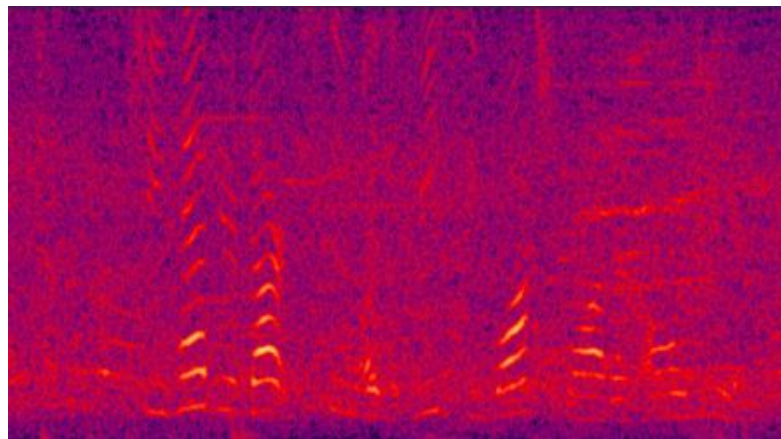
- Speech enhancement (SE)
 - Background
 - Conventional speech enhancement techniques
- Speech enhancement based on deep neural networks
 - SE-DNN: background
 - DNN baseline and enhancement
 - Noise-universal SE-DNN
- Other acoustic signal processing (**ASP**) related effort
 - DNN-based source separation
 - DNN-based bandwidth expansion
 - DNN-based voice conversion
 - DNN-based preprocessing for robust ASR
- Summary

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

Speech Enhancement

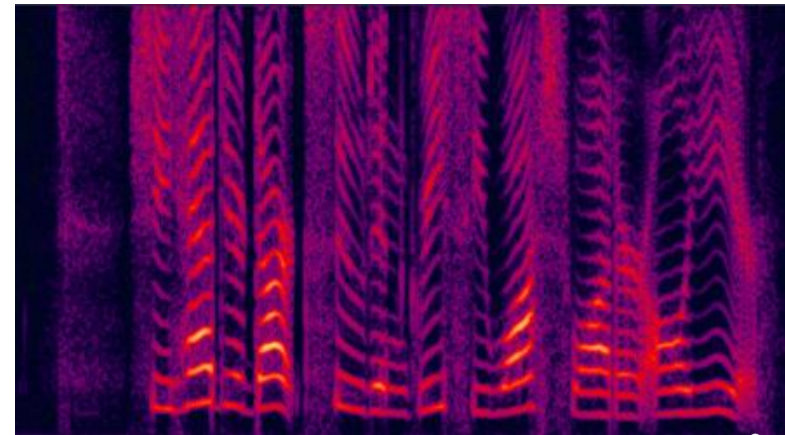
- Speech enhancement aims at improving the intelligibility and/or overall perceptual quality of degraded speech signals using audio signal processing techniques
- One of the most addressed classical SP problems in recent years



Noisy speech,
Exhibition, SNR=5dB



Enhancing
→



Clean speech, 8kHz



Speech Enhancement Applications



**Mobile phone/
communication**



Hearing aids



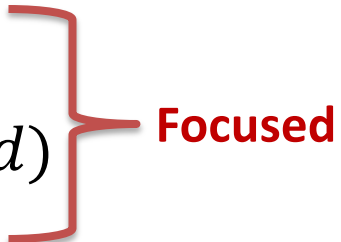
**Security monitoring/
intelligence**



**Robust speech
/speaker/language
recognition, etc.**

Noise in Speech Enhancement

1. Additive noise:

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


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

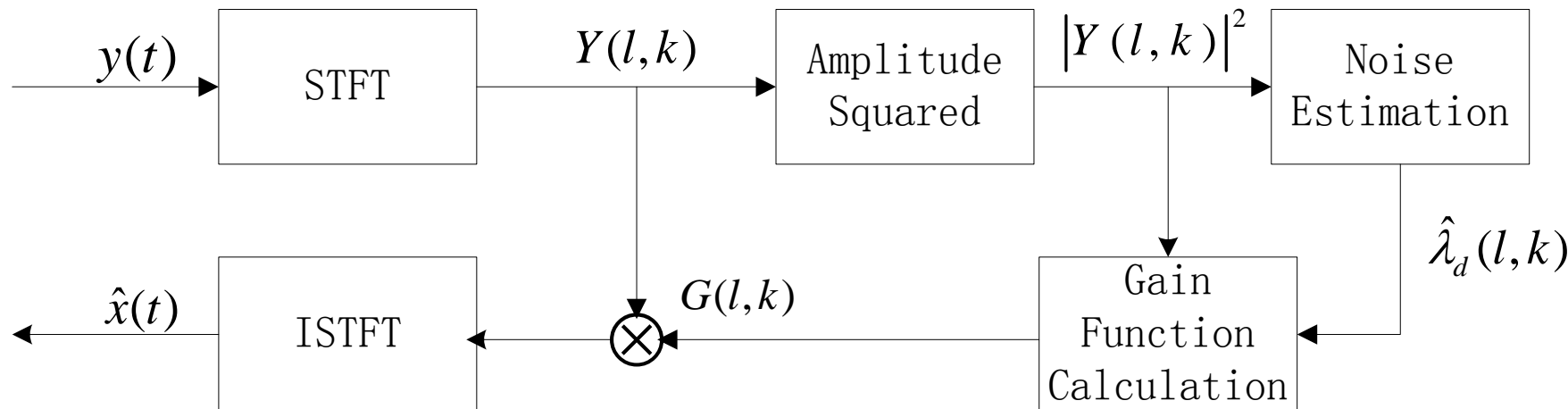
- Speech enhancement
 - Background
 - Conventional speech enhancement methods
- Speech enhancement based on deep neural networks
 - SE-DNN: background
 - DNN baseline and enhancement
 - Noise-universal SE-DNN

Conventional Speech Enhancement

- **Classified by the number of microphones**
 1. **Single channel speech enhancement methods**
 - ☐ Time and frequency information
 2. **Microphone based speech enhancement methods**
 - ☐ Time and frequency information
 - ☐ Spatial information
 - ☐ Microphone arrays
- **Conventional Techniques**
 - Spectral subtraction, Wiener filtering
 - MMSE Log Spectral Amplitude, MMSE-LSA
 - Optimally Modified LSA, OM-LSA
 - others

} **Focused**

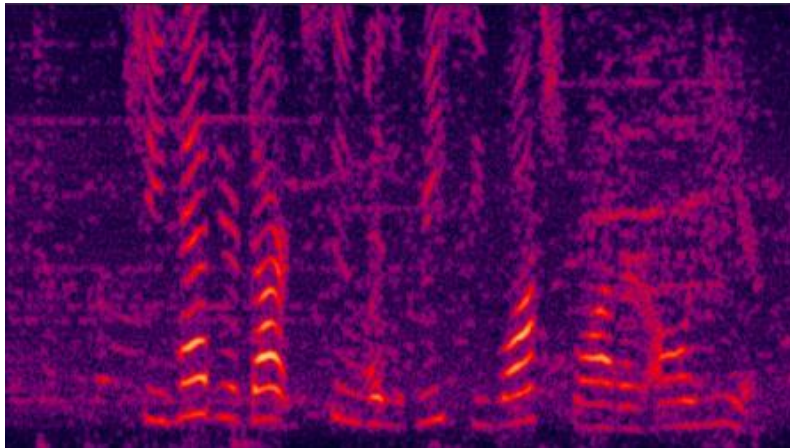
Conventional Single Channel SE



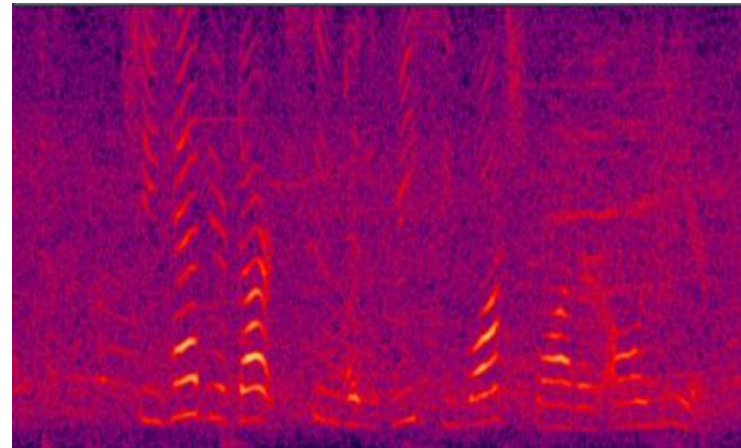
1. STFT on the noisy signal y , get the time-frequency signal Y
2. Estimate the variance of noise $\hat{\lambda}_d$
3. Estimate all of the parameters (prior SNR γ , posterior SNR ξ and the speech presence probability, etc.) needed by the gain function
4. Calculate the gain function G
5. Multiply Y with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

Conventional Single Channel SE: Issues

1. Musical noise:



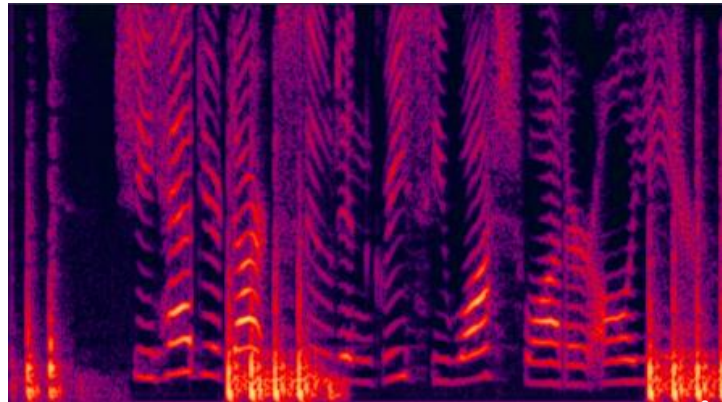
Enhanced by SS



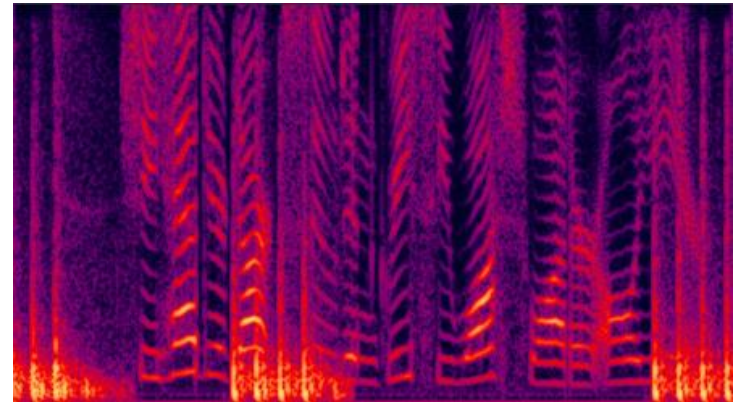
Noisy speech, exhibition noise, SNR=5dB

Conventional Single Channel SE: Issues

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



Enhanced by OM-LSA



Noisy, Machine Gun,
SNR=-5dB



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

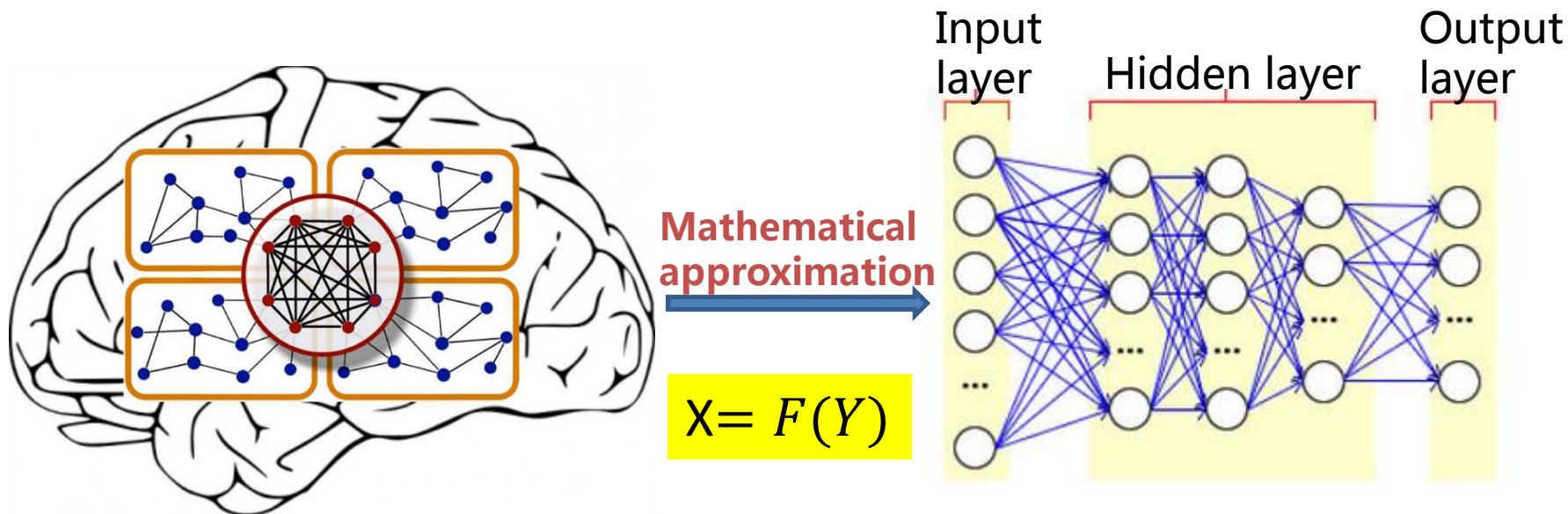
- The signal model of the additive noise:

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

- Many enhancement methods are derived from this signal model, however, most of them assume that $X(n, d)$ is described by a Gaussian mixture model (GMM) independent from $N(n, d)$ that 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 :
$$X = F(Y)$$
 - 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. ANN has been used a great deal in the 80's and 90's
2. Hinton proposed the unsupervised Restricted Boltzmann Machine (RBM) based pre-training in 2006
3. In 2012, MSR, Google and IBM got a great success in large vocabulary continuous speech recognition using DNNs
4. Later, DNNs were adopted in many speech-related tasks



DNN Based SE: Related Work

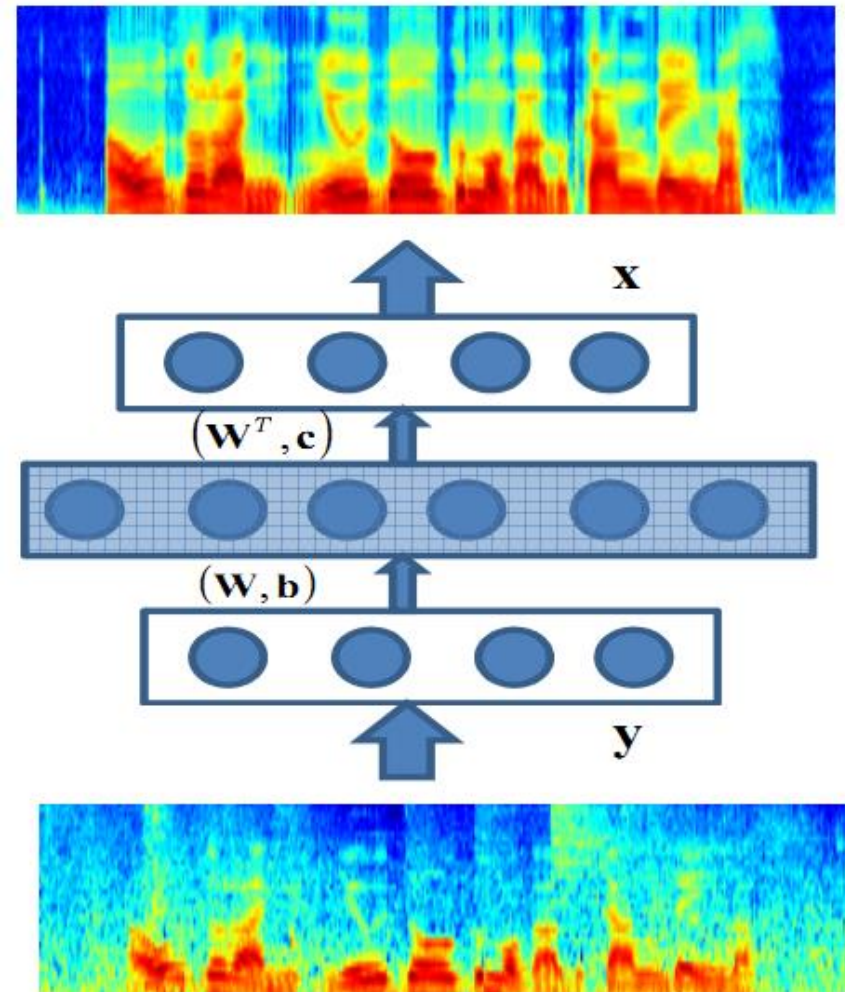
1. In 2013, Xugang Lu proposed deep de-noising auto-encoder based speech enhancement

$$h(\mathbf{y}_i) = \sigma(\mathbf{W}_1 \mathbf{y}_i + \mathbf{b})$$

$$\hat{\mathbf{x}}_i = \mathbf{W}_2 h(\mathbf{y}_i) + \mathbf{c},$$

$$\mathbf{W}_1 = \mathbf{W}_2^T = \mathbf{W}$$

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



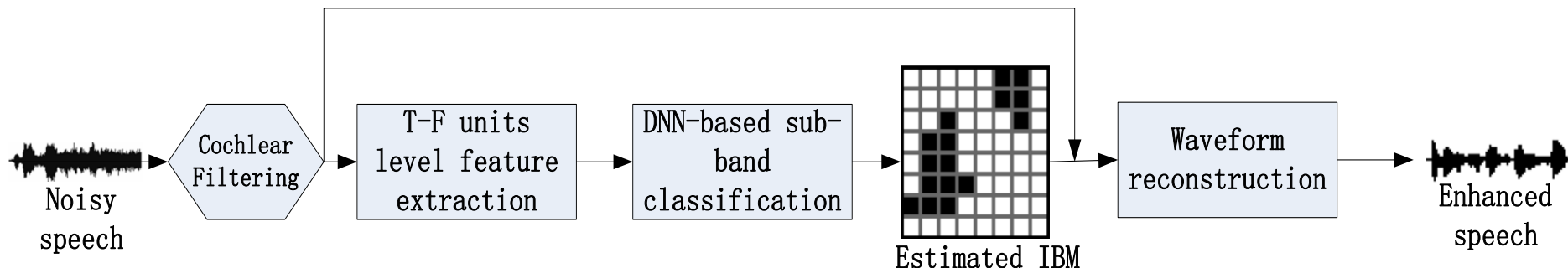
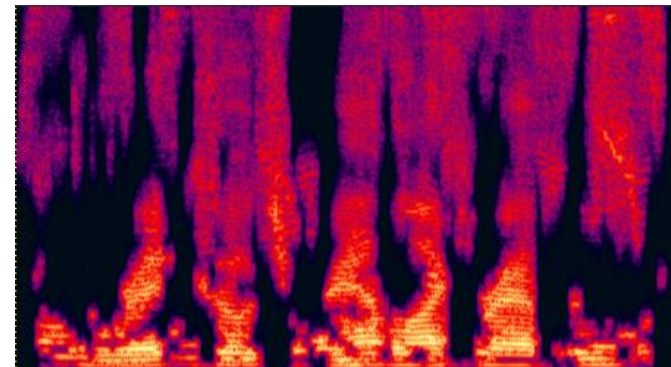
X.-G. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising Auto-Encoder," Proc. Interspeech, pp. 436-440, 2013.

DNN Based SE: Related Work

2. 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: Research Issues

- **Advantages of SE-DNN**

1. Nonlinear relationship between noisy and speech
2. Deep architecture for regression approximation
3. Highly non-stationary noise could be learnt
4. Nearly no Gaussian or independent assumptions
5. Nearly no empirical thresholds to avoid the non-linear distortion in SS-based speech enhancement

- **Challenges of SE-DNN**

1. Which domain is suitable for DNN-based mapping?
2. The generalization capacity to unknown environments
3. 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

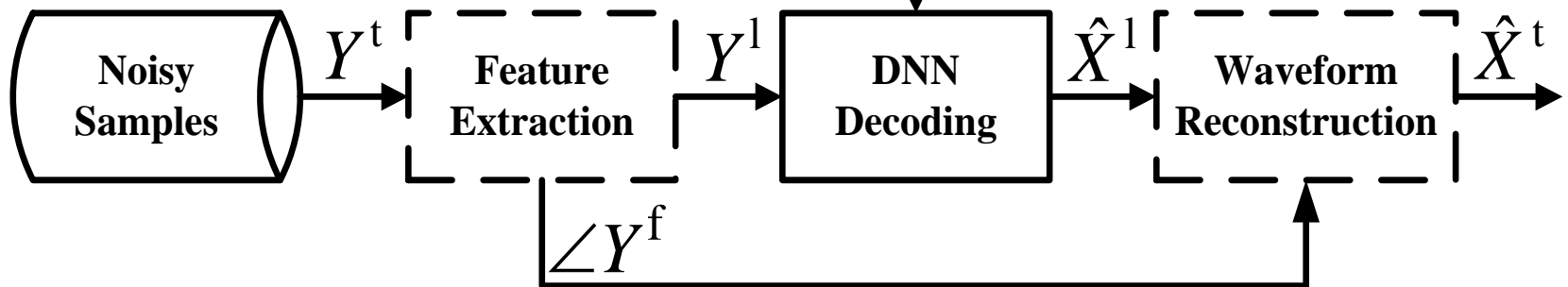
System Overview

Training Stage



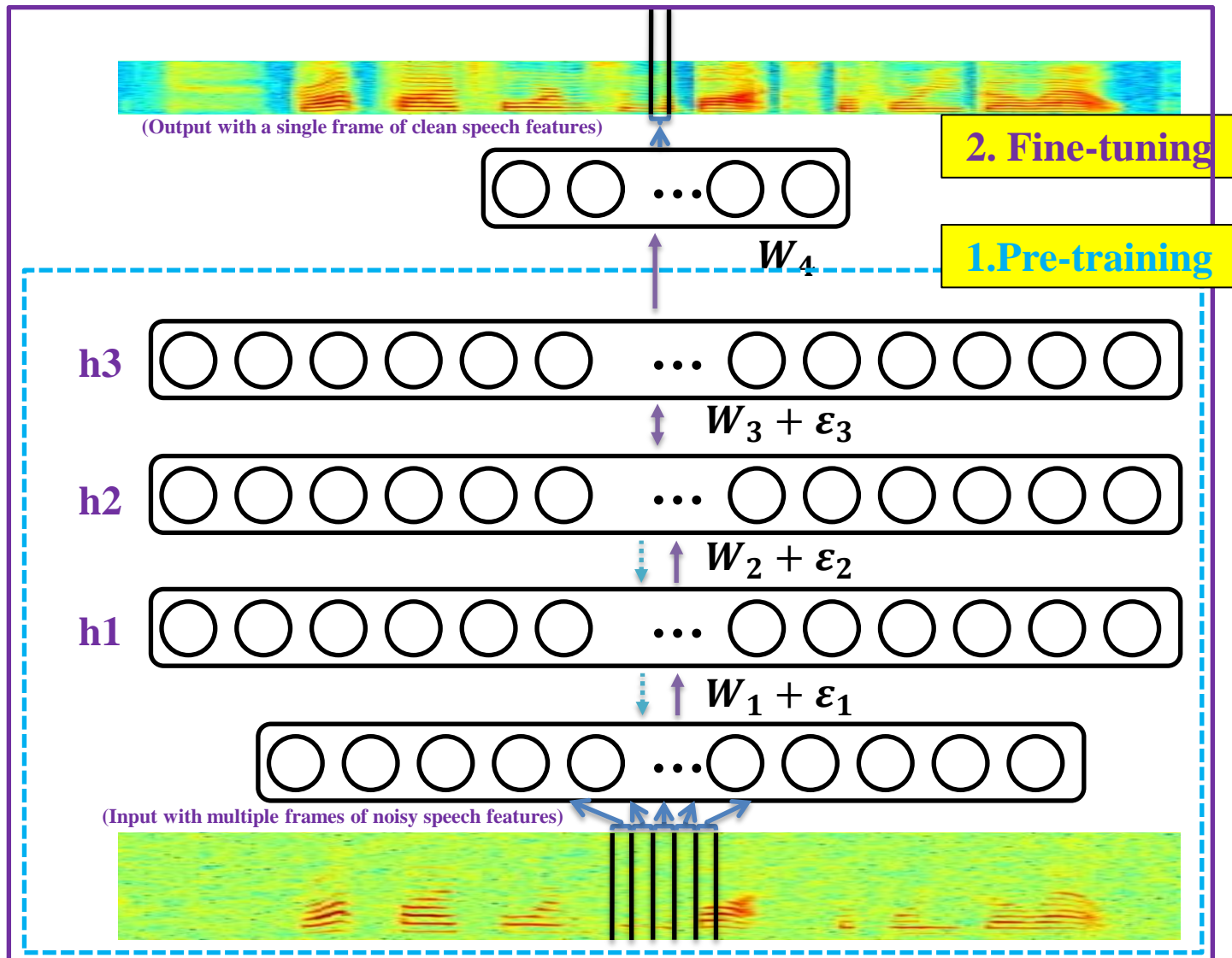
Signal Processing
Letter, Jan. 2014

Enhancement Stage



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

DNN Training



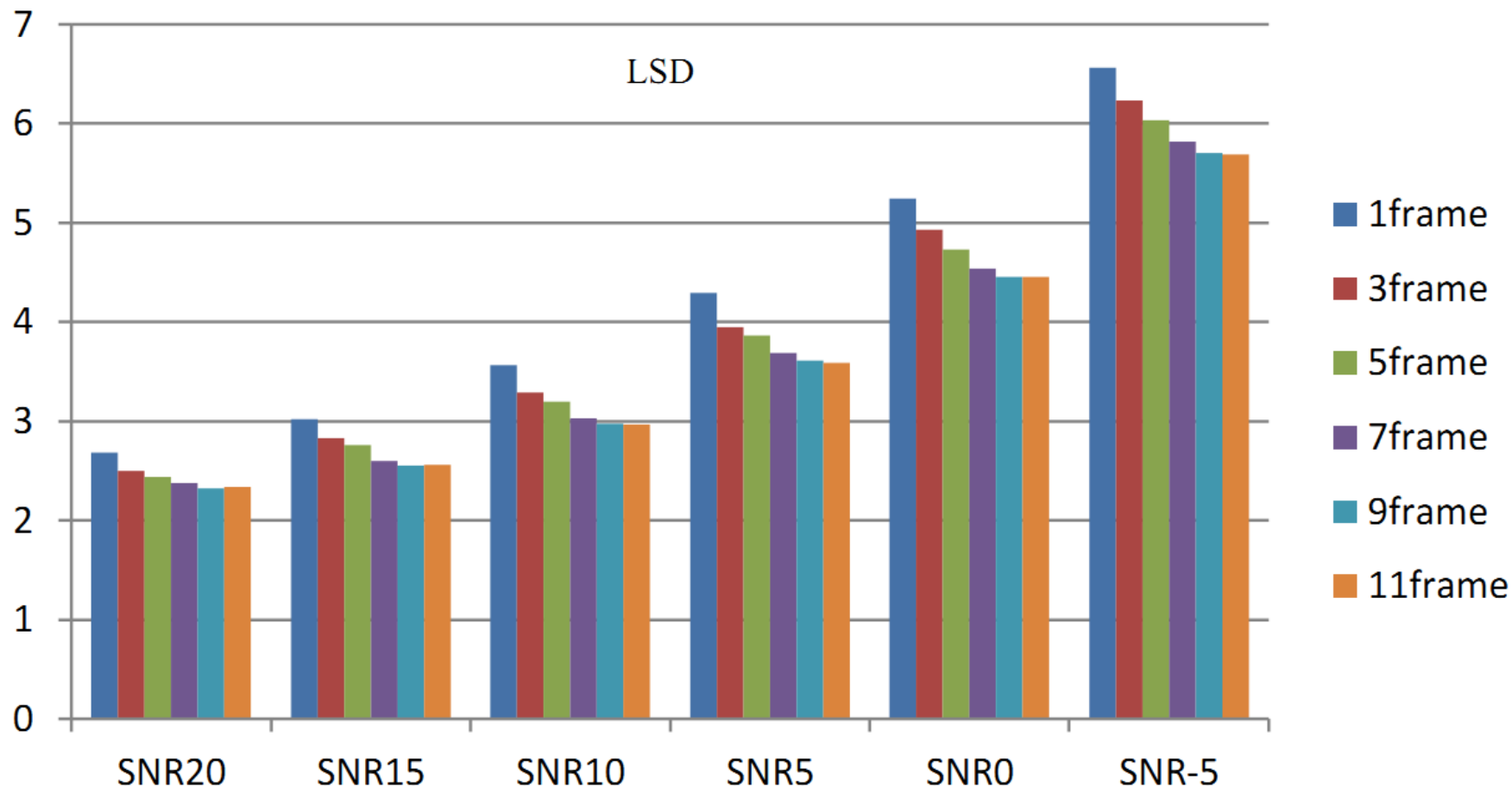
1. MMSE-based object function: $E = \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D (\hat{X}_n^d(\mathbf{W}, \mathbf{b}) - X_n^d)^2 + \lambda \|\mathbf{W}\|_2^2$

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)

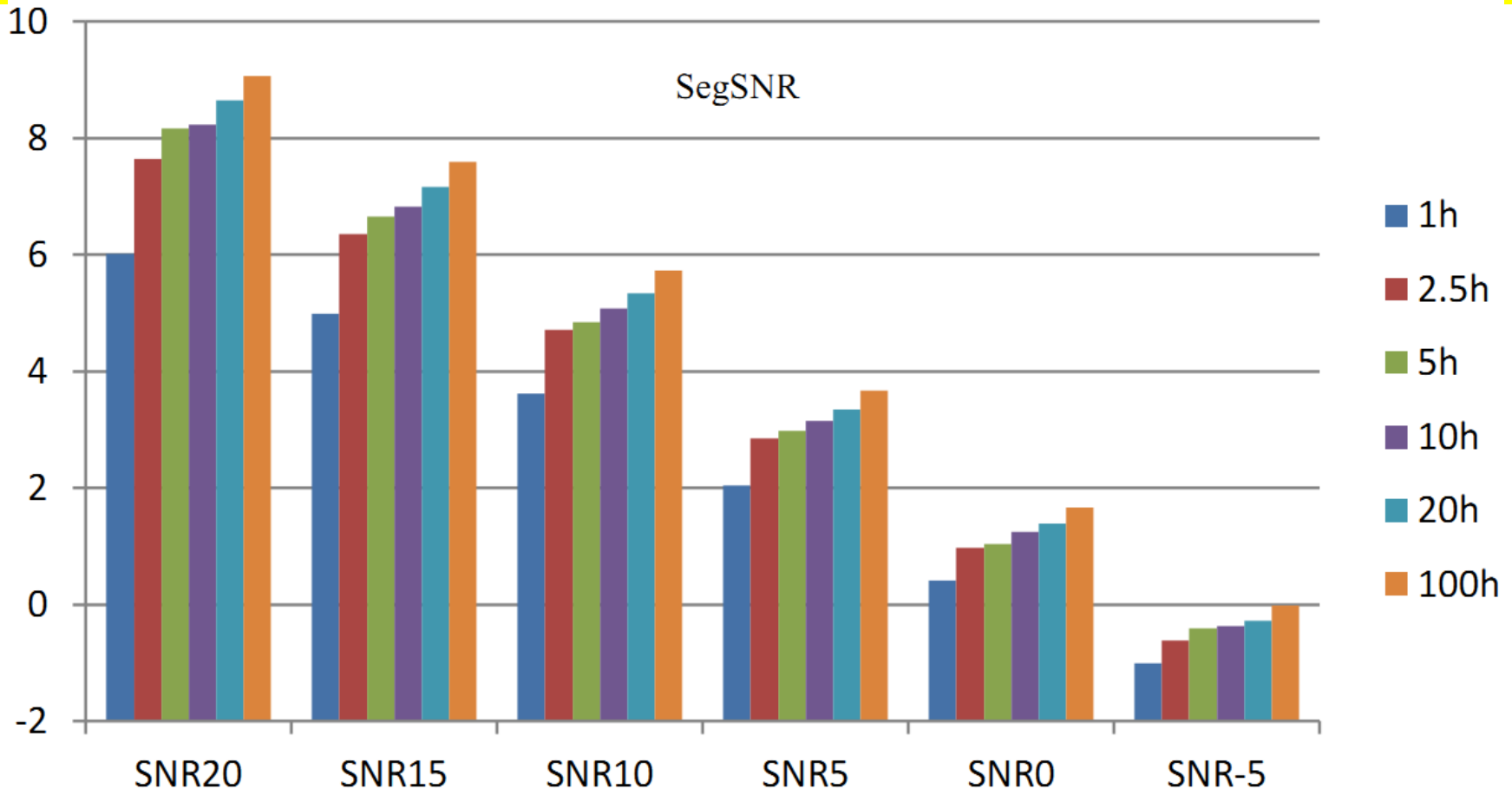
Baseline Experimental Results: I

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 **SegSNRs** 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

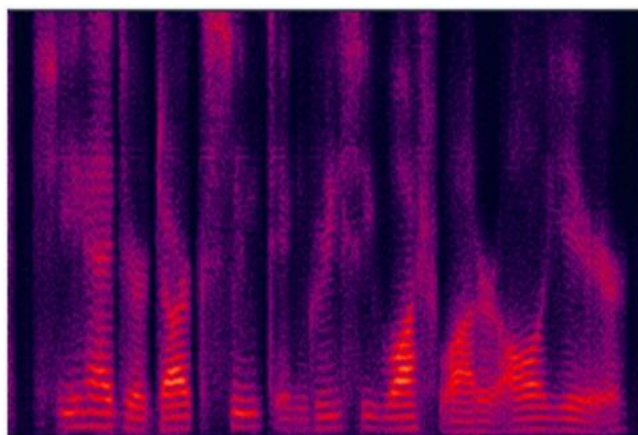
3. Average **PESQs** among methods on the test set at different SNRs with four noise types. The subscript of DNN_l represents l hidden layers

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

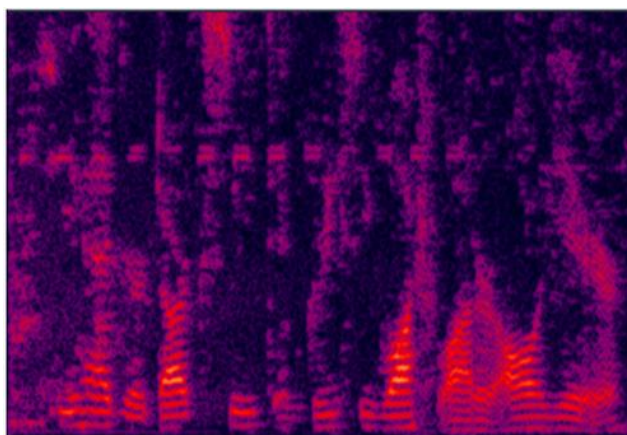
*Shallow Neural Network (SNN) has the same computation complexity with DNN_3

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

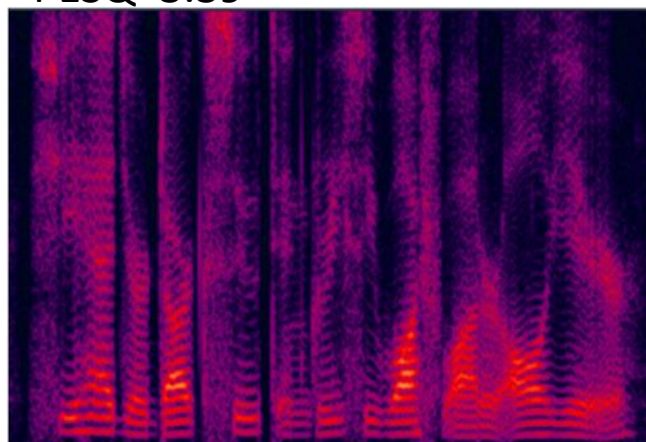
Baseline Experimental Results V



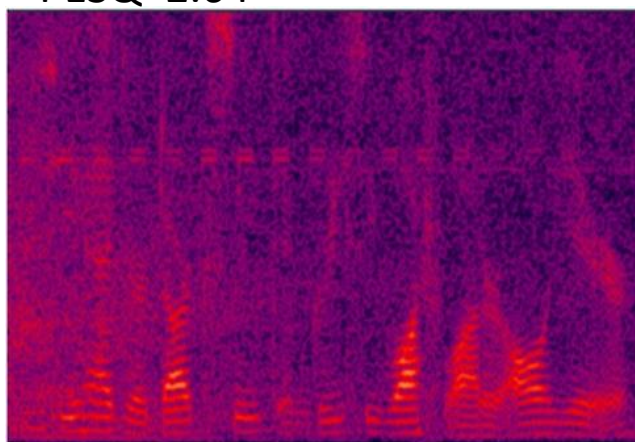
DNN enhanced
PESQ=3.39



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.

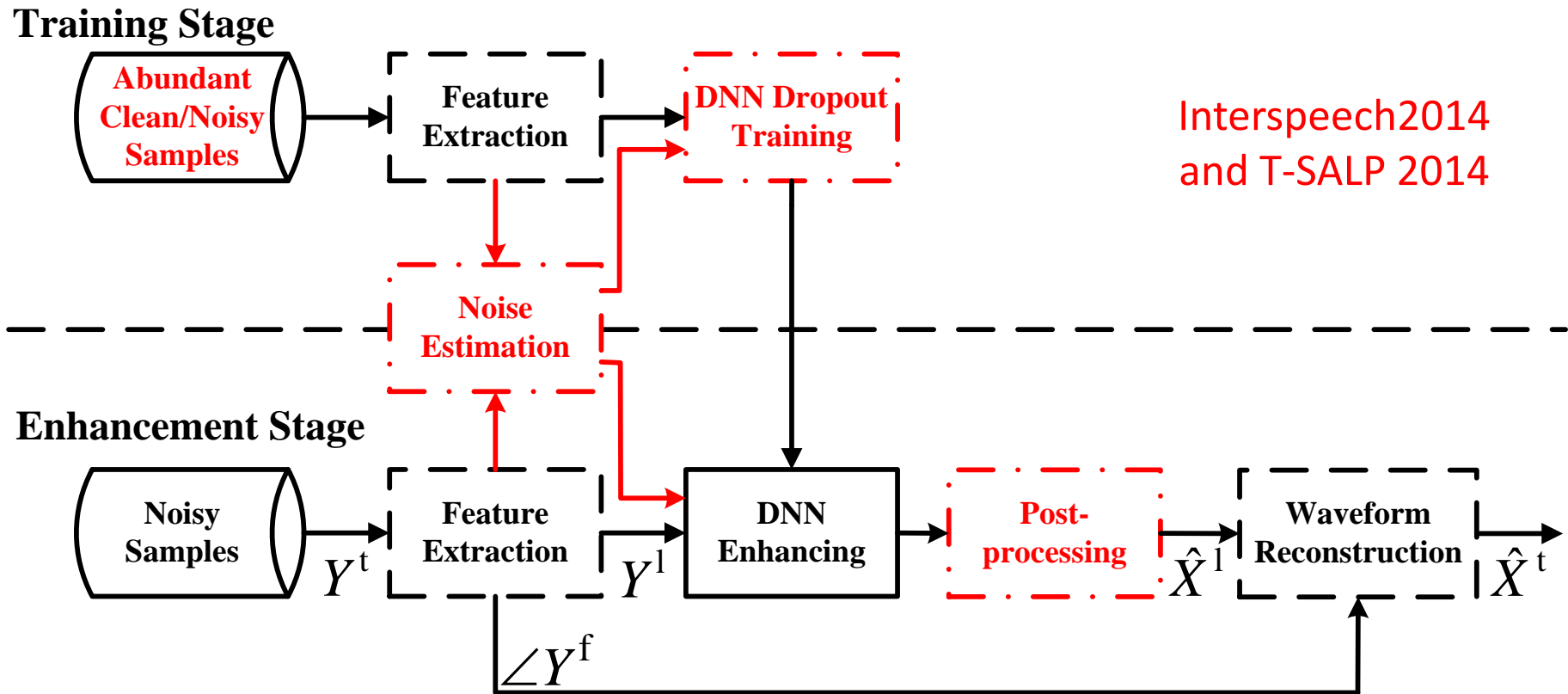
Summary I: DNN-SE Basic Properties

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, with post-training or post-processing

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 (1/3)



- Noise-universal speech enhancement (Interspeech2014)
- Global variance equalization in the post-processing (ChinaSIP2014)

Noise-Universal SE-DNN (2/3)

1. DNN to learn the characteristics of many noise types

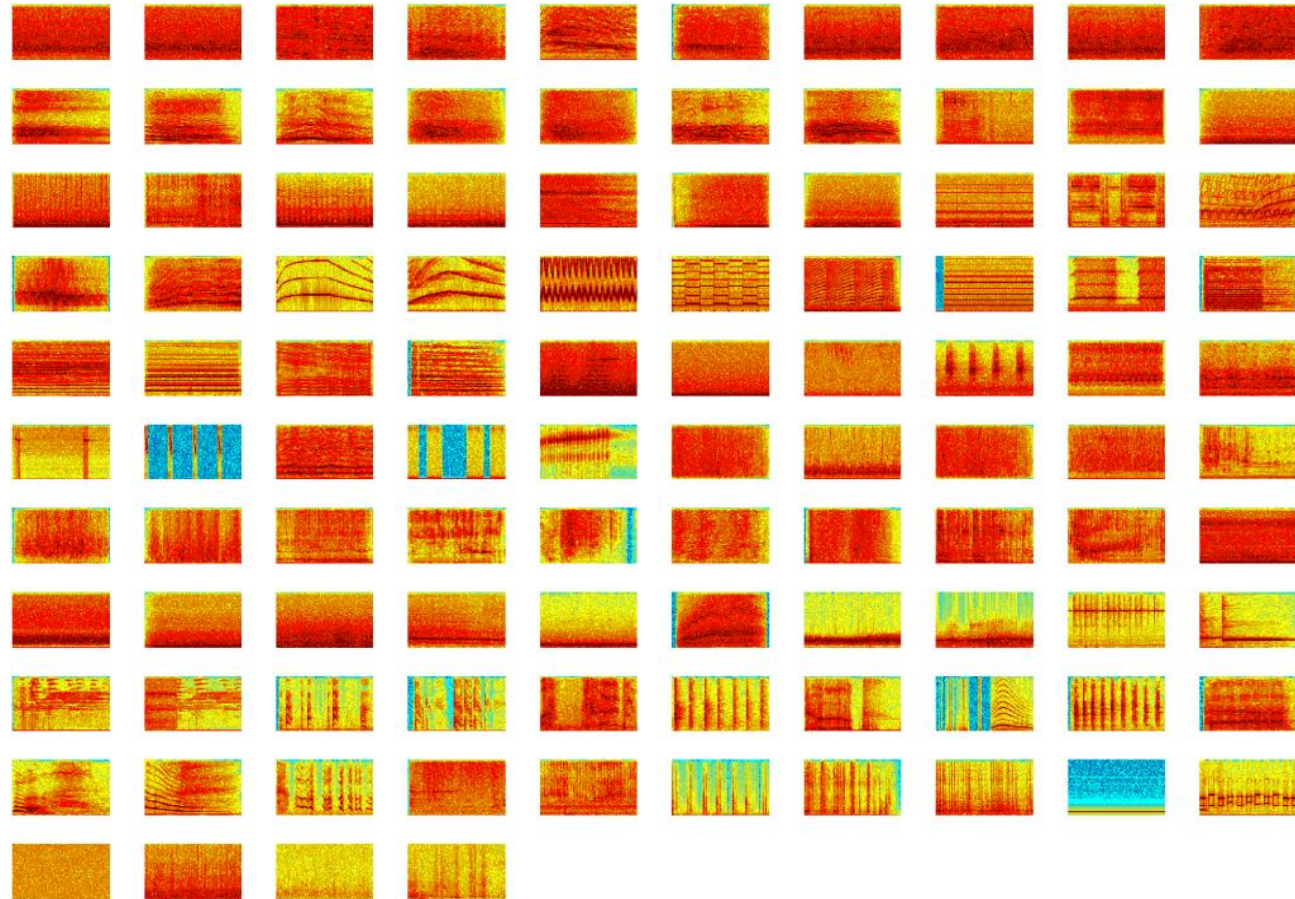
□ Classifications :
Crowding、machine、
transportation、animal、
nature、human, etc.



alarm



cry



G. Hu, 100 non-speech environmental sounds, 2004.

<http://www.cse.ohiostate.edu/pnl/corpus/HuCorpus.html>.

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

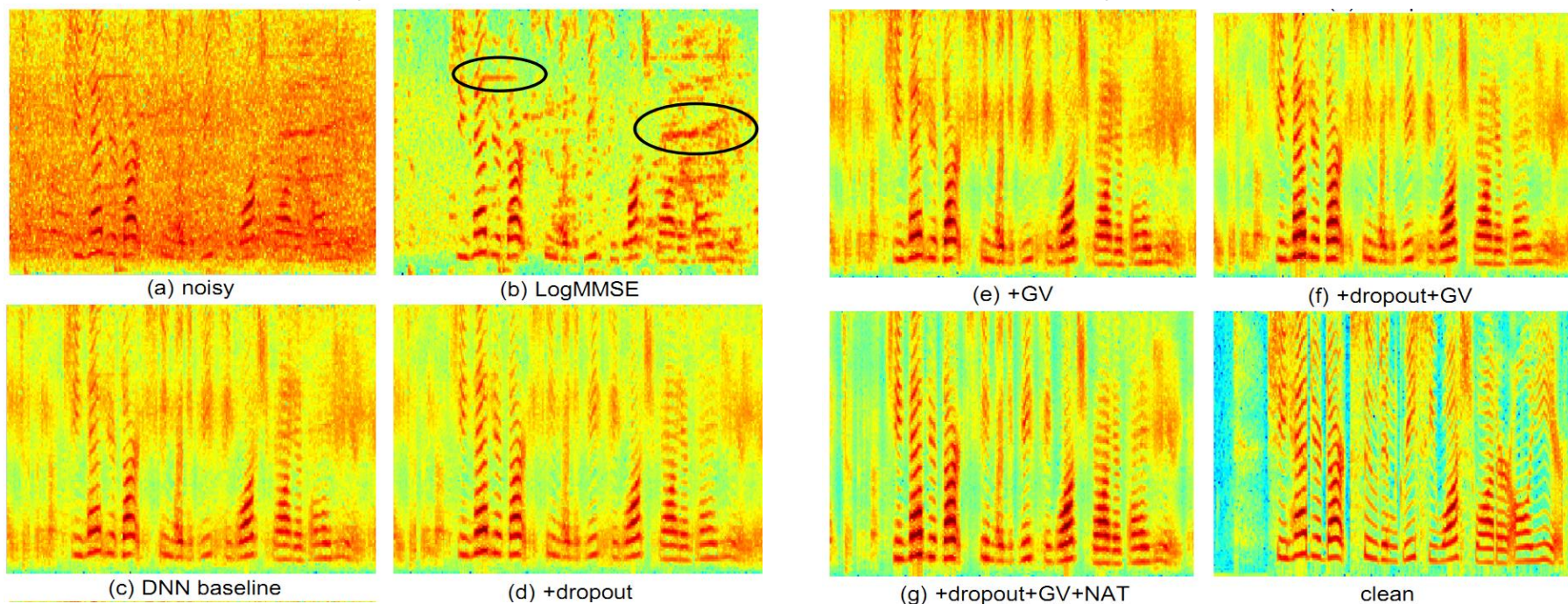
- **LSD** comparison between models trained with four noise types (4NT) and 104 noise types (104NT) on the test set at different SNRs of three **unseen** noise environments :

	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

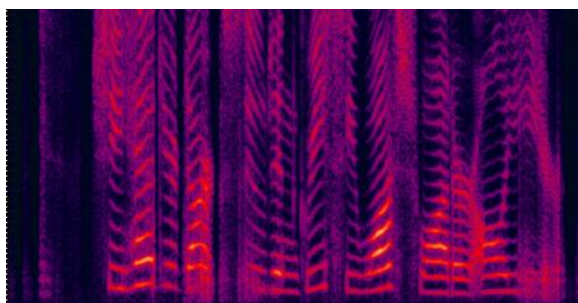
- Spectrograms of an utterance tested with *Exhibition* noise at SNR= -5dB. (a) noisy (PESQ=1.42), (b) LogMMSE (PESQ=1.83), (c) DNN baseline (PESQ=1.87), (d) improved by dropout (PESQ=2.06), (e) improved by GV equalization (PESQ=2.00), (f) improved by dropout and GV (PESQ=2.13), (g) jointly improved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



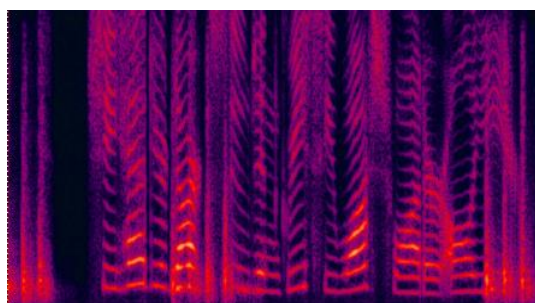
- SE-DNN can suppress the highly non-stationary noise and get less residual noise
- Dropout and NAT can reduce noise while GV equalization can brighten speech

Enhanced Experimental Results: III

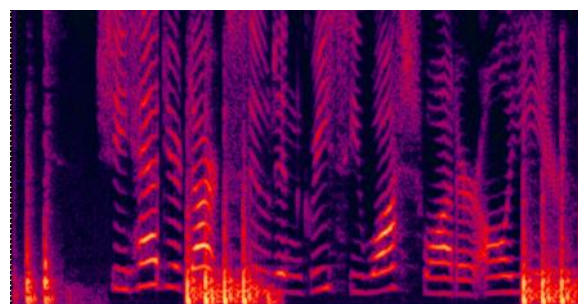
- Spectrograms of an utterance with machine gun noise at SNR= -5dB: with 104-noise DNN enhanced (upper left, PESQ=2.78), Log-MMSE enhanced (lower left, PESQ=1.86), 4-noise DNN enhanced (upper right, PESQ=2.14), and noisy speech (lower right, PESQ=1.85):



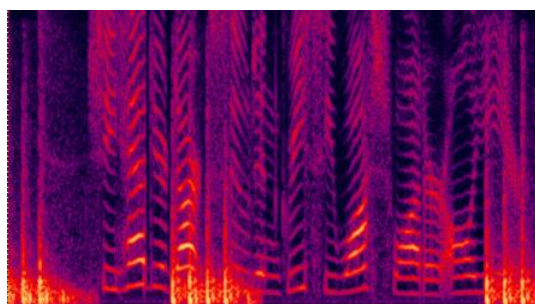
104NT-DNN
enhanced PESQ=2.78



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

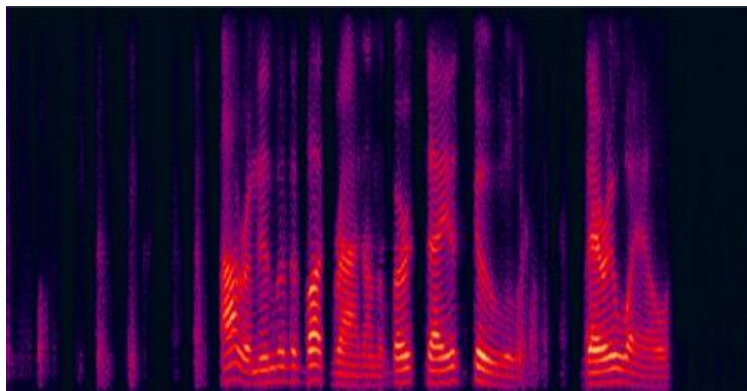
- Average **PESQ** among LogMMSE, DNN baseline with 100 hours data, improved DNN with 100 hours data and improved DNN with 625 hours data on the test set at different SNRs across the whole 15 **unseen** noise types in the NOISEX-92 database:

	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

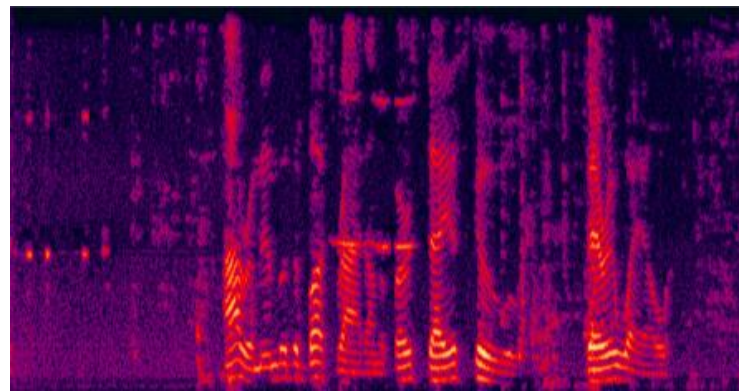
- A good generalization capacity to unseen noise can be obtained.
- SE-DNN outperformed the Log-MMSE, especially at low SNRs

Enhanced Experimental Results: V

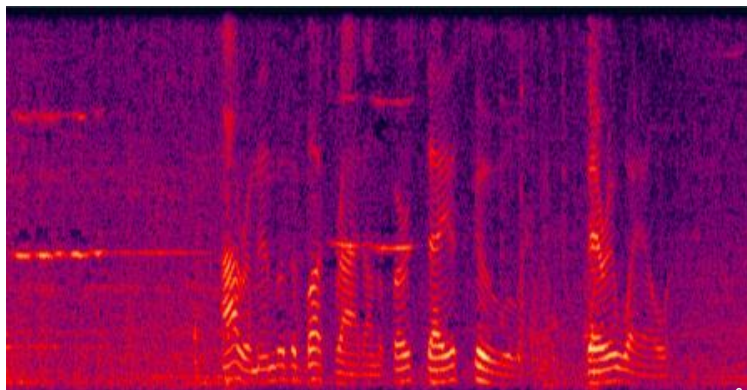
- Spectrograms of a noisy utterance extracted from the movie **Forrest Gump** with: improved DNN (upper left), Log-MMSE (upper right), and noisy speech (bottom left): with **real-world noise never seen**



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. SPL2014 demo: http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html
2. Speech separation: DNN-based semi-supervised separation works better than state-of-the-art supervised source separation (Interspeech2014)
3. Global variance equalization: enhanced contrast (ChinaSIP2014)
4. Noise-Universal SE (Interspeech2014 & T-SALP)
5. Robust ASR: better results with only DNN-based pre-processing, additional compensation can be added later (Interspeech2014)
6. Robust ASR: best results with some post-processing (ICASSP2015)
7. Transfer language learning for DNN (ISCSLP2014)
8. Dual-Output DNN for separation (ICSP2014)
9. CHIME challenge: best results with only DNN-based separation, more post-processing can be added later (ICSP2014 , ICASSP2015)
10. DNN-based bandwidth expansion (better than GMM, ICASSP2015)
11. DNN-based voice conversion (better than GMM, ICASSP2015)

DNN Preprocessing for Robust ASR: III

3. Aurora-4 WER: SE DNN-HMM, LMFB ([Interspeech2014](#))
- After some post-processing: 10.3% (best, [ICASSP2015](#))

System	A	B	C	D	Avg.
Clean-condition Training					
Noisy	4.2	30.8	22.5	47.6	35.5
DNN-PP	4.2	10.9	10.0	27.6	17.5
Multi-condition Training					
Noisy	4.6	8.4	7.8	18.6	12.5
DNN-PP	4.5	7.5	7.4	19.3	12.3














DNN Preprocessing for Robust ASR: V

5. Speech separation challenge (SSC): Chime (ICASSP2015)

- 18 females and 16 males, 500 GRID utterances each
- Mixed speakers, target-to-masker ratio: -9dB to 6dB
- Training 2 DNNs, -10dB to 0dB and 0dB to 10dB
- Using models from Chime (34 for 16 KHz, 1 for 25 KHz)

Word Accuracy	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
16KHz waveform							
Baseline	49.1	34.2	22.9	13.7	10.2	8.0	23.0
DNN	92.6	89.7	86.7	81.3	75.1	69.9	82.6
SND-DNNs	93.1	90.9	89.3	87.6	84.7	75.9	86.9
25KHz waveform							
Baseline	63.3	47.5	35.2	24.0	17.0	12.0	33.2
SND-DNNs	94.9	93.6	92.4	90.6	87.0	81.9	90.1
IBM	93.0	92.5	91.5	89.5	87.0	79.0	88.8

Summary: ASP Demos

1. SPL Jan. 2014, http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html
2. Source separation: DNN-based semi-supervised speech separation
 - Mixed speech:  Target speech: 
 - Log-MMSE enhanced speech: 
 - DNN-separated speech 
3. DNN-based bandwidth expansion works better than all other state-of-the-art techniques + phase info
 - Wideband speech: 
 - Bandwidth-expanded speech with phase imaged: 
 - Bandwidth-expanded speech with phase estimated: 
4. DNN-based enhancement on Samsung benchmark data
 - Noisy speech from main mike:   
 - XYZ 2-mic enhanced:   
 - Single-channel DNN enhanced: 