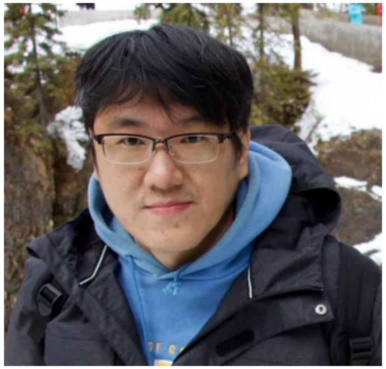
Speech Enhancement based on Deep Learning and Intelligibility Evaluation (Part II)

Yu Tsao
Research Center for Information Technology Innovation
Academia Sinica
yu.tsao@citi.sinica.edu.tw

中央研究院
ACADEMIA SINICA

Dr. Yu Tsao (曹 昱), *Associate Research Fellow*



– Education

- Ph.D. in ECE, Georgia Institute of Technology, 2003-2008
- M.S. in EE, National Taiwan University, 1999-2001
- B.S. in EE, National Taiwan University, 1995-1999

– Work Experience

- Researcher, National Institute of Information and Communications Technology, Spoken Language Communication Group, Japan (2009/4-2011/9)
- Summer Research Associate, Texas Instruments Incorporated, Speech Technologies Laboratory DSP Solutions R&D Center, United States (2004, 2005, 2006 summers)

– Academia Services

- Vice Chair, Speech, Language, and Audio (SLA) Technical Committee, APSIPA
- Distinguished Lecturer, 2019-2020, APSIPA
- Associate Editor of IEICE transactions on Information and Systems
- Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing

– Lab at CITI (Academia Sinica)

Biomedical Acoustic Signal Processing (Bio-ASP) Lab

– Research Interests

Assisitve Speech Communication Technologies, Audio-coding, Deep Neural Networks, Biomedical Signal Processing, Speech Signal Processing

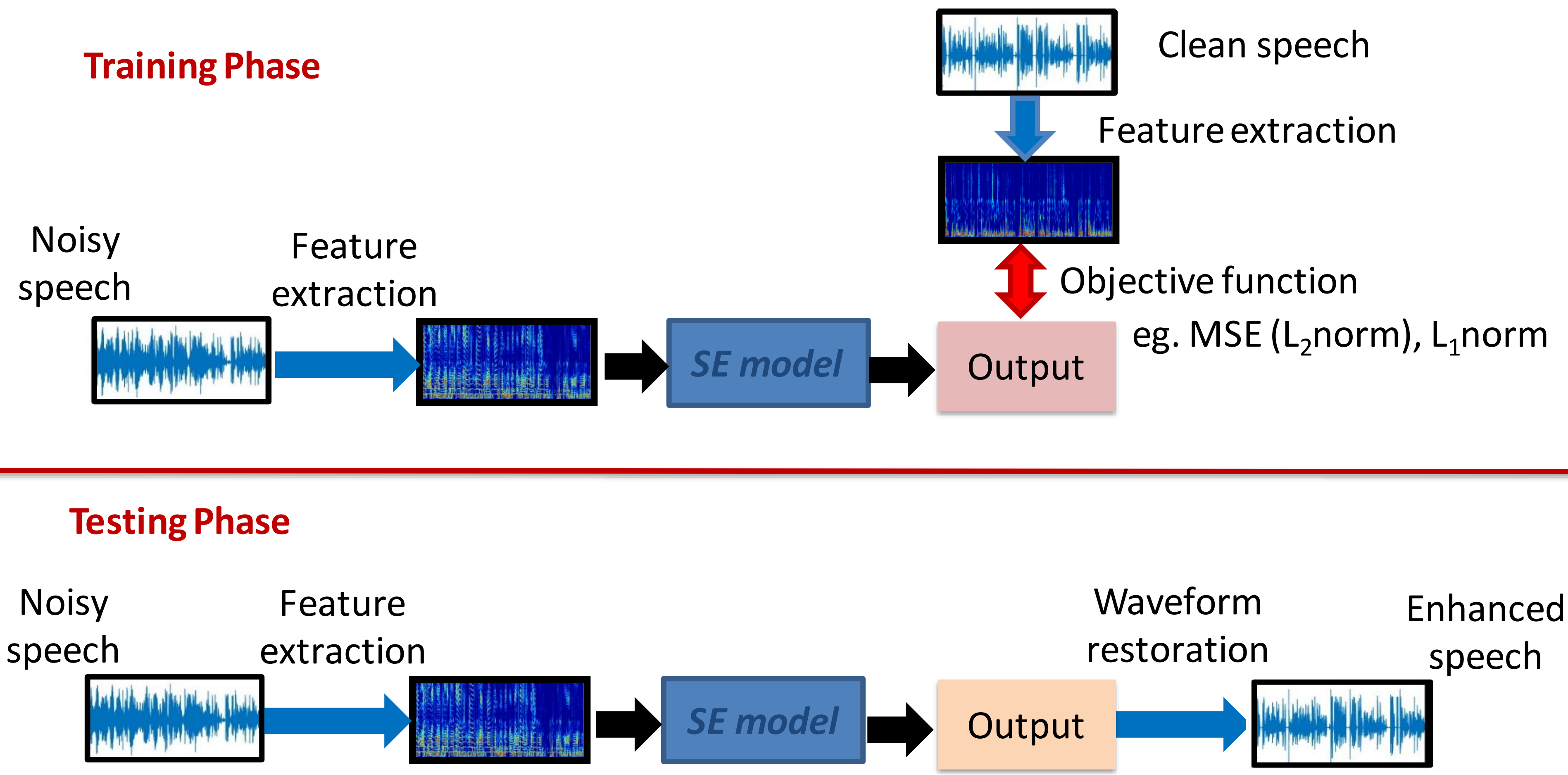
Outline

- Deep Learning based Speech Enhancement
  - System architecture
  - Six factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- Summary

Outline

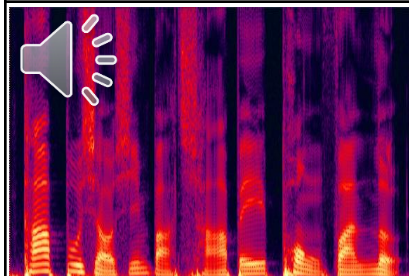
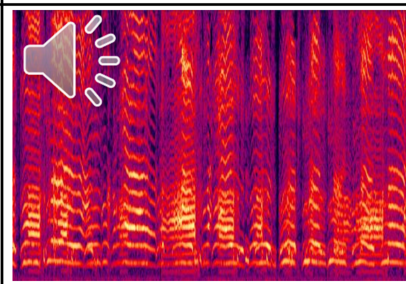
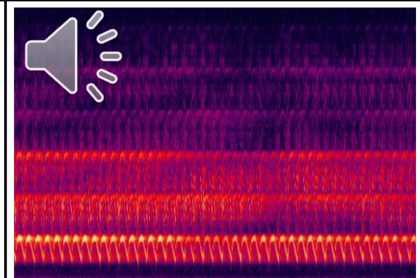
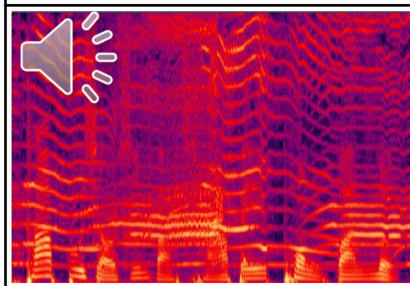
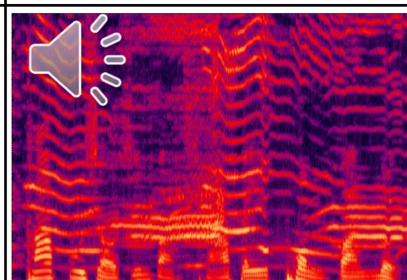
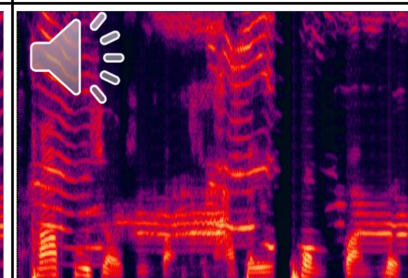
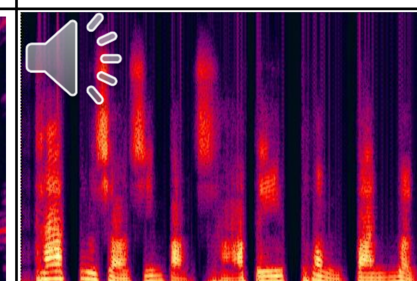
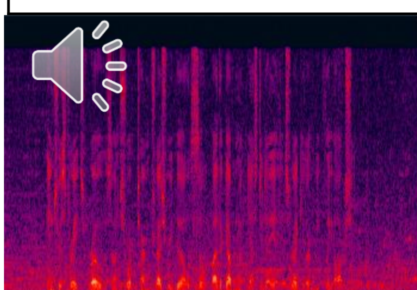
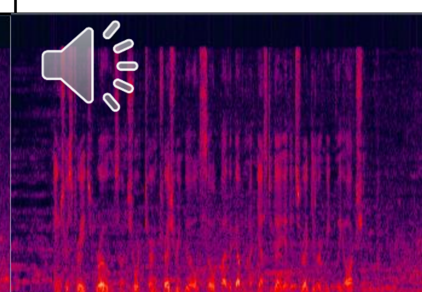
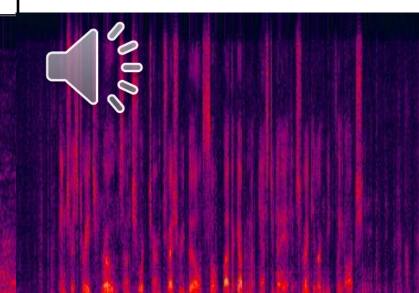
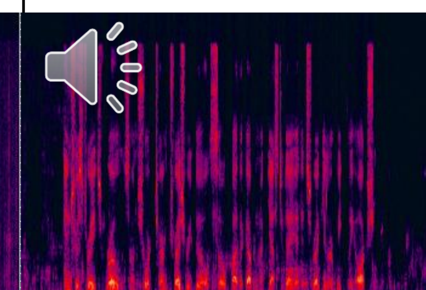
- Deep Learning based Speech Enhancement
  - **System architecture**
  - Six factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- Summary

Deep Learning Based SE System

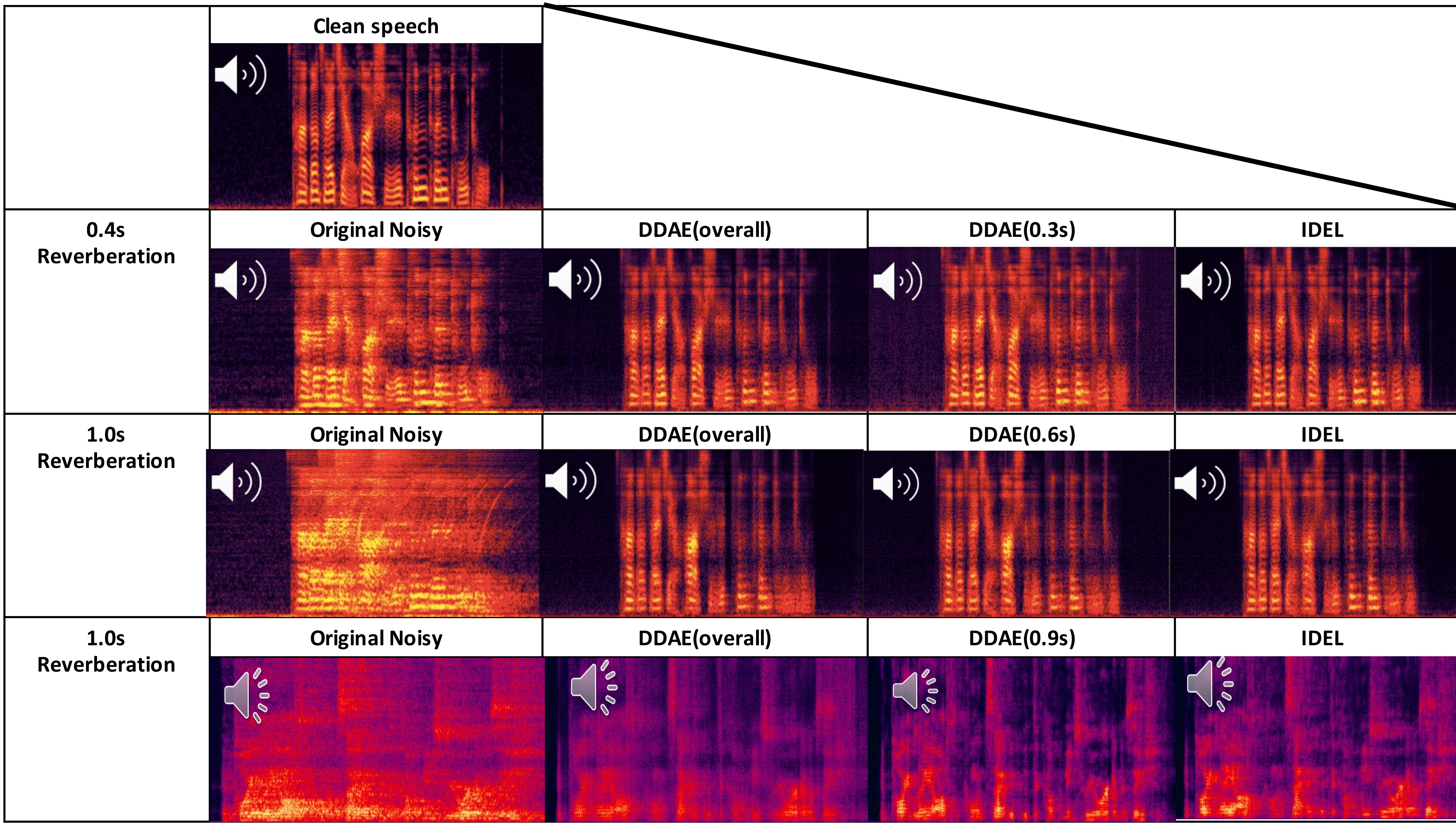


The first work of DL-based SE system: [Lu et al, Interpseech 2013].

DL for Noise Reduction (Denoising)

	Clean speech	Noise: baby Crying	Noise: Siren	
				
Baby Crying	Original Noisy	MMSE (Trandtional-1)	KLT (Trandtional-2)	DDAE
				
Siren	Original Noisy	MMSE (Trandtional-1)	KLT (Trandtional-2)	DDAE
				

DL for De-reverberation

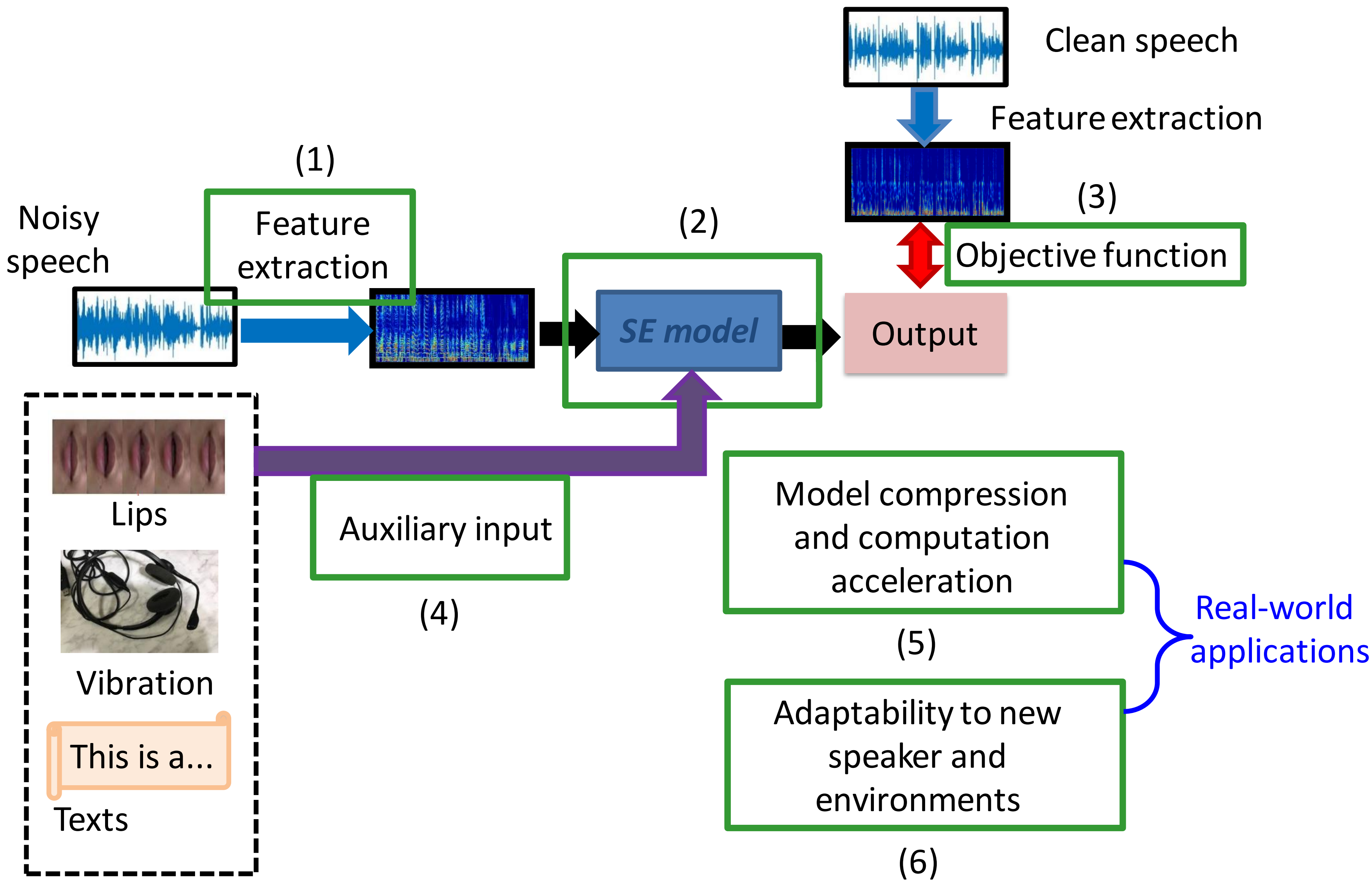


The examples were based on [Lee et. al., ICASSP 2018].

DL for Channel Compensation

The examples were based on [Liu et. al., Speech Comm. 2018].

Deep Learning Based SE System



Evaluation Metrics

- Perceptual Evaluation of Speech Quality **(PESQ)**: evaluating the quality of processed speech, with the score ranging from -0.5 to 4.5.
- Short-Time Objective Intelligibility **(STOI):** evaluating the speech intelligibility, with the score ranging from 0 to 1.
- Segmental Signal-to-Noise Ratio **(SSNR)**: the ratio of processed and noisy speech computed in a segment level.
- Log-Spectral-Distortion **(LSD)**: the difference of log spectrums of processed speech and clean reference.

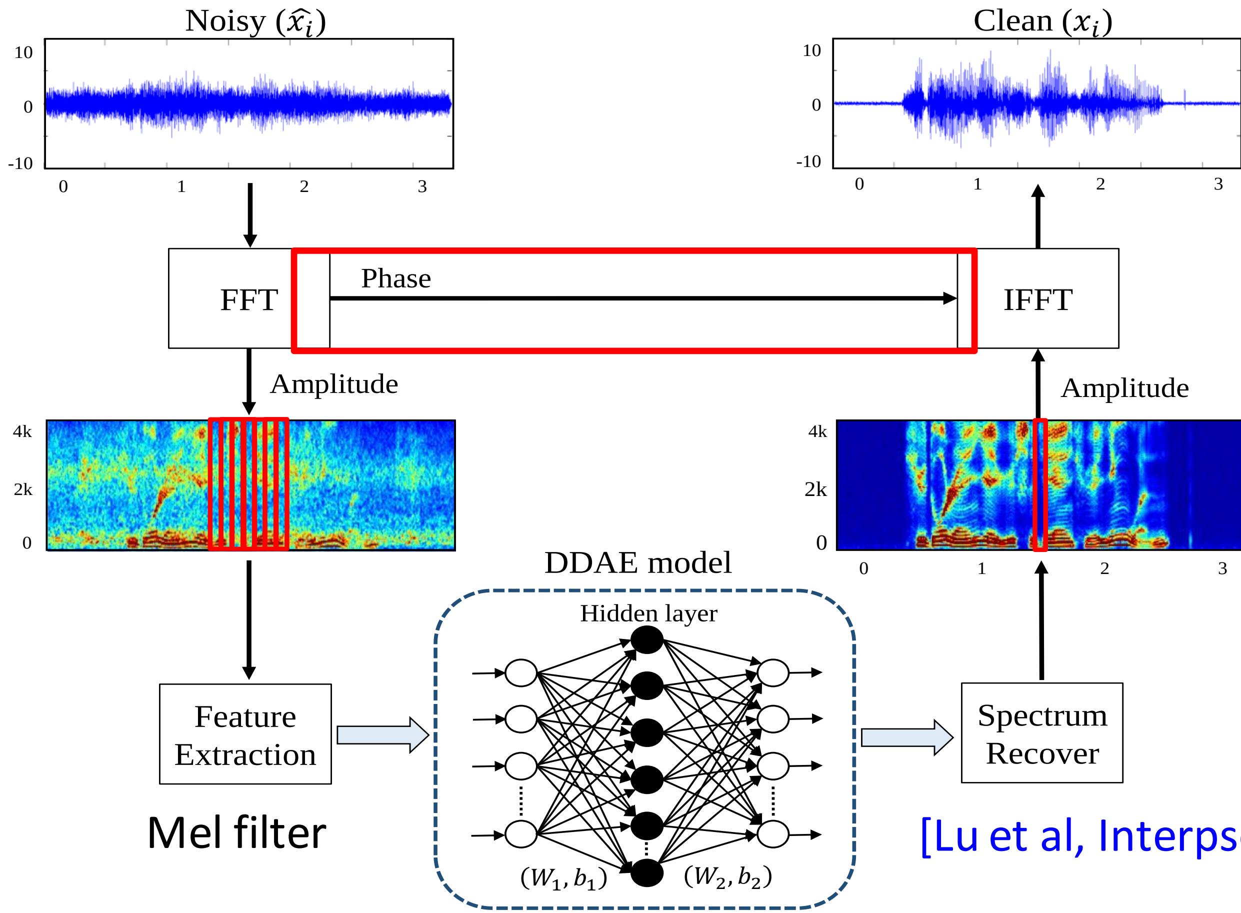
The goal of SE is to improve the speech intelligibility and quality.

Outline

- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- Summary

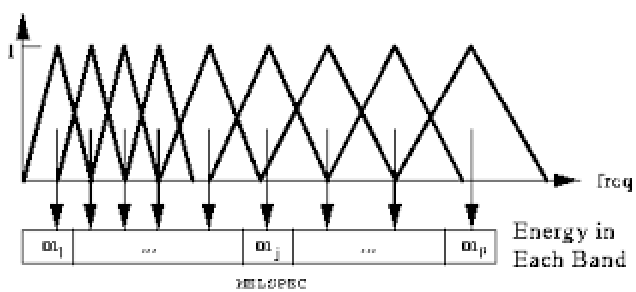
Outline

- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ **Feature types**

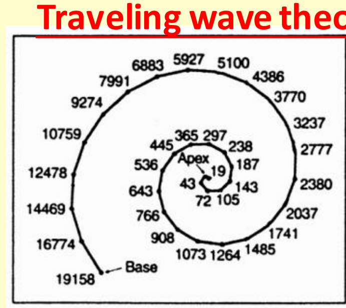


Mel filter

[Lu et al, Interpseech 2013]



Traveling wave theory

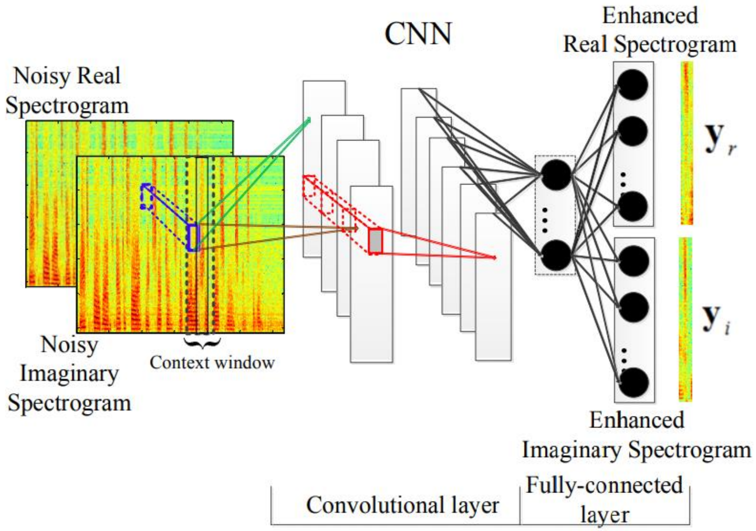


1961

Von Békésy, Georg (1960). Experiments in hearing. Ed. Ernest Glen Wever. Vol. 8. New York: McGraw-Hill.

Input Feature Types

- Complex spectrogram (CS) [Fu et. al., in MLSP, 2017]



RI spectrograms enhanced by CNN. Real and imaginary spectrograms are treated as different input channels.

(1) The motivation is to obtain more accurate phase information.
(2) The real and imaginary (RI) spectrograms can be considered as R, G, B in a color image and processed by a CNN model.

Input Feature Types (CS)

- LSD, SSNR, STOI, and PESQ scores:

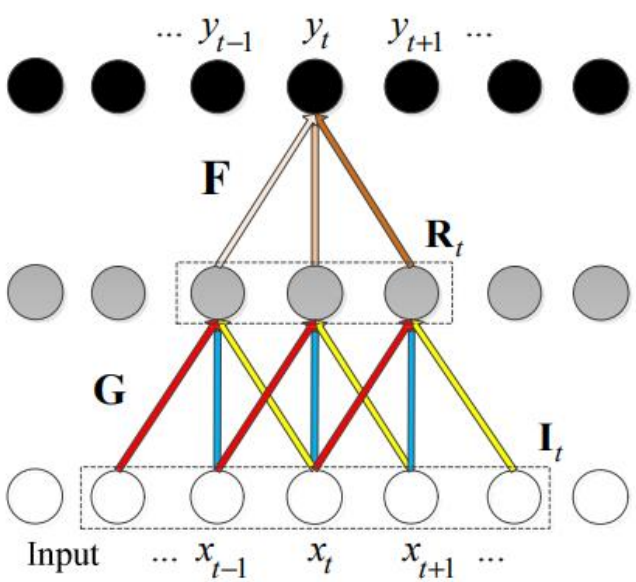
Performance comparisons of different models and input features in terms of LSD (log spectral distortion), SSNR, STOI, and PESQ.

	DNN-baseline				RI-DNN ($\alpha=1, \beta=0$)				RI-CNN ($\alpha=1, \beta=0$)			
SNR (dB)	**LSD**	**SSNR**	**STOI**	**PESQ**	**LSD**	**SSNR**	**STOI**	**PESQ**	**LSD**	**SSNR**	**STOI**	**PESQ**
12	**3.115**	-0.229	0.814	2.334	3.761	2.149	0.851	2.643	3.604	**3.042**	**0.886**	**2.741**
6	**3.404**	-1.243	0.778	2.140	3.936	1.113	0.817	2.404	3.844	**1.975**	**0.850**	**2.525**
0	**3.747**	-2.802	0.717	1.866	4.200	-0.454	0.750	2.088	4.150	**0.450**	**0.783**	**2.233**
-6	**4.114**	-4.974	0.626	1.609	4.521	-2.745	0.645	1.778	4.491	**-1.911**	**0.675**	**1.908**
-12	**4.426**	-7.070	0.521	1.447	4.838	-5.604	0.512	1.539	4.829	**-4.990**	**0.537**	**1.638**
Avg	**3.761**	-3.264	0.691	1.879	4.251	-1.108	0.715	2.090	4.183	**-0.286**	**0.746**	**2.209**

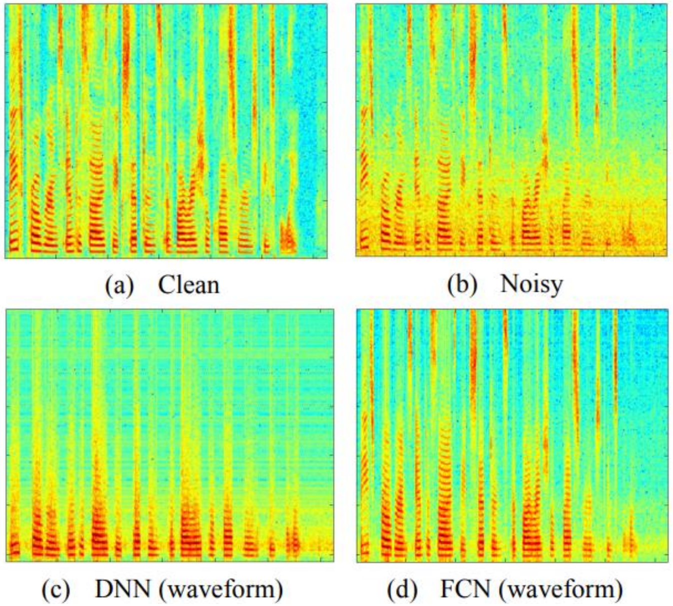
(1) Log-power-spectrum (LPS) with DNN gives lowest LSD.
(2) RI with DNN outperforms LPS with DNN in terms of PESQ and STOI.
(3) CNN outperforms DNN when using RI spectral features.

Input Feature Types

- Waveform as the input (Wav) [Fu et. Al., APSIPA, 2017]



Local connection in FCN



Spectrograms of a TIMIT utterance: (a) clean speech, (b) noisy speech, (c) DNN, and (d) FCN

(1) Using waveform can address the issue of phase estimation.
(2) We observe that fully convolutional network (FCN) architecture is more suitable than fully connected neural networks.

Input Feature Types (Wave)

- Waveform versus LPS:

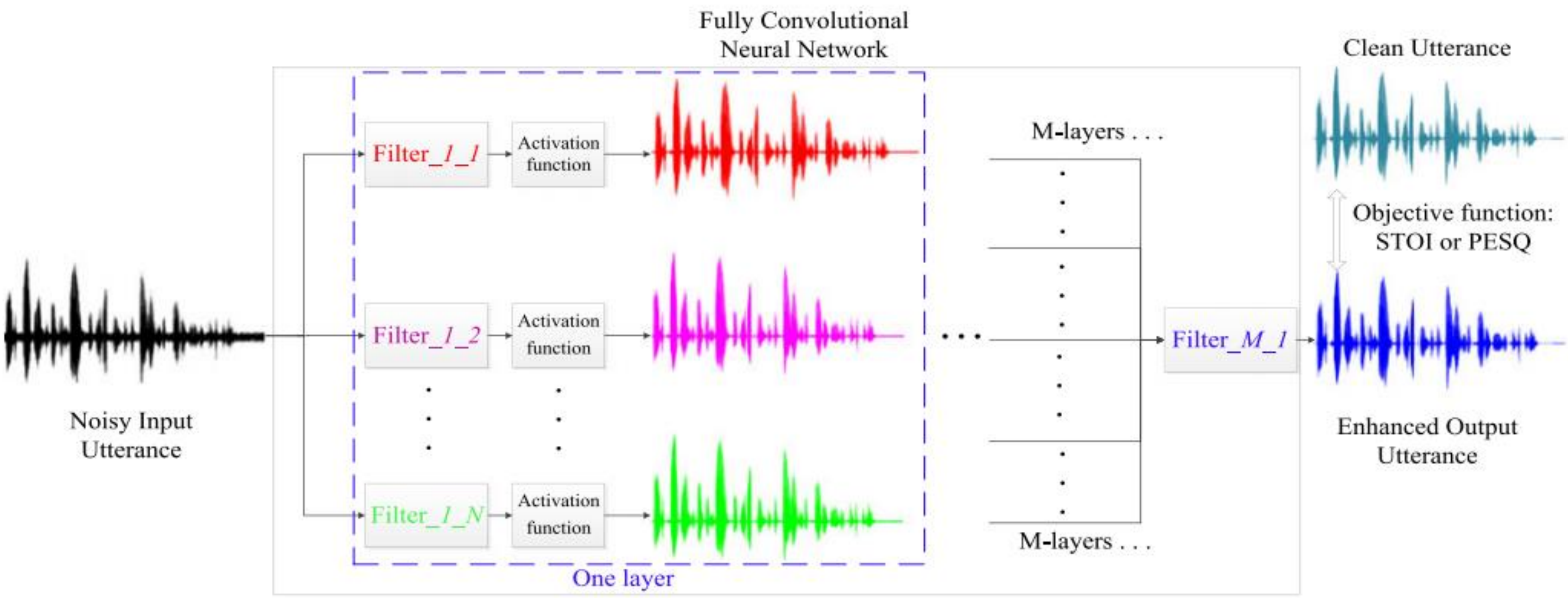
Comparison of different models and input features in terms of STOI, and PESQ.

	DNN-baseline (LPS)		DNN (waveform)		CNN (waveform)		FCN (waveform)	
SNR (dB)	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**
12	0.814	2.334	0.737	2.548	0.788	2.470	0.874	2.718
6	0.778	2.140	0.715	2.396	0.753	2.302	0.833	2.346
0	0.717	1.866	0.655	2.118	0.673	2.011	0.758	1.995
-6	0.626	1.609	0.549	1.816	0.561	1.707	0.639	1.719
-12	0.521	1.447	0.429	1.573	0.441	1.453	0.506	1.535
Avg.	0.691	1.879	0.617	**2.090**	0.643	1.989	**0.722**	2.063

(1) Waveform with FCN achieves the highest STOI score.
(2) Waveform with DNN achieves the highest PESQ score.

Input Feature Types

- Utterance waveform (UWave) [Fu et. al., TASLP, 2018]



Utterance enhancement by fully convolutional networks (FCN).
The FCN model has multiple layers, each layer consisting of multiple filters.
The model can take inputs with arbitrary lengths.

Input Feature Types (UWave)

- A comparison of utterance-based and frame-based waveform as the inputs

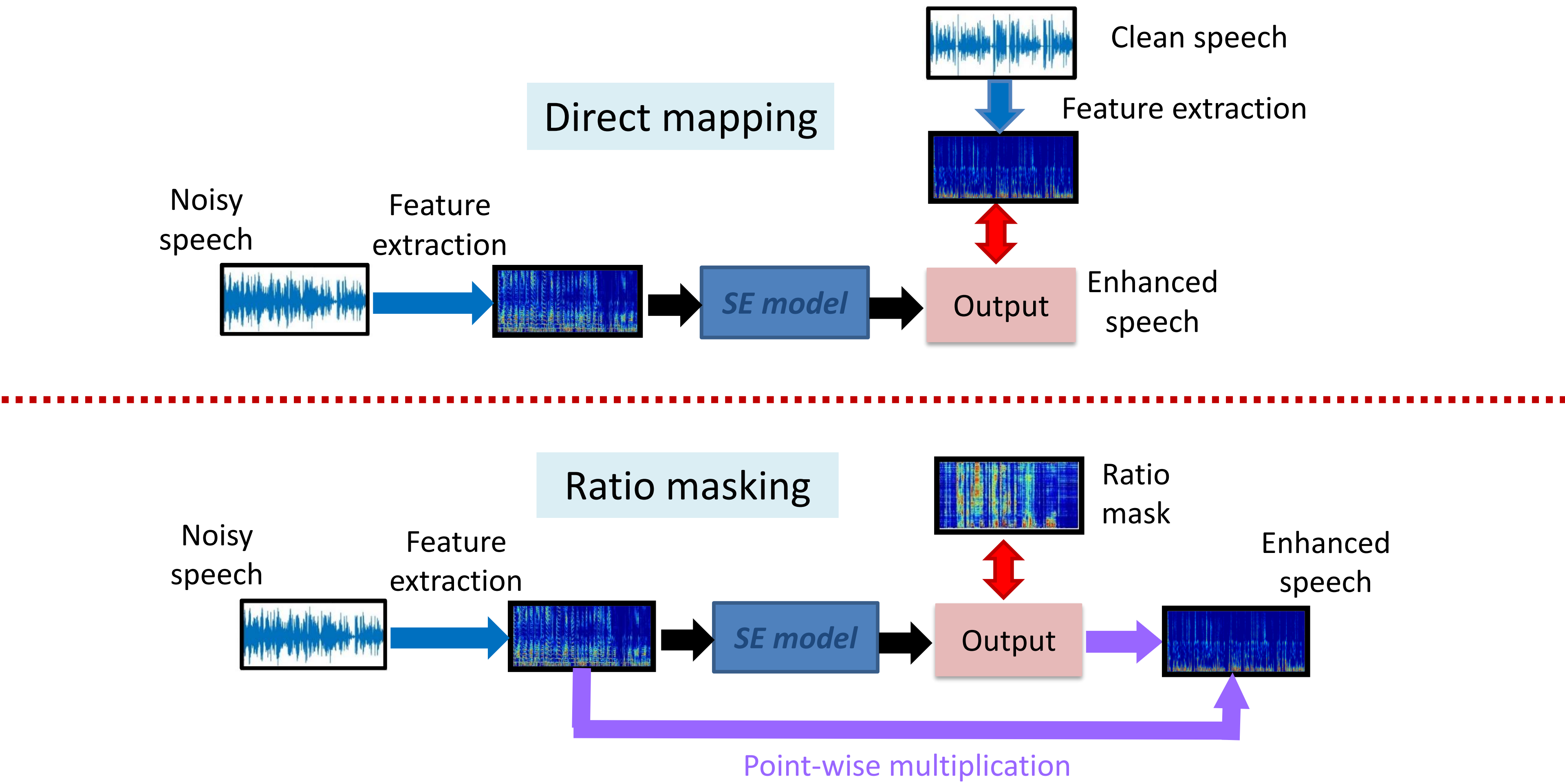
Comparison of different models and input features in terms of STOI and PESQ.

	Frame-based		Utterance-based			
	FCN (obj=MSE)		**FCN** (obj=MSE)		**FCN** (obj= STOI)	
SNR (dB)	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**
12	0.874	2.718	0.909	**2.909**	**0.931**	2.587
6	0.833	2.346	0.864	**2.481**	**0.888**	2.205
0	0.758	1.995	0.780	**2.078**	**0.814**	1.877
-6	0.639	1.719	0.647	**1.754**	**0.699**	1.608
-12	0.506	1.535	0.496	**1.536**	**0.562**	1.434
Avg.	0.722	2.063	0.739	**2.152**	**0.779**	1.942

(1) Utterance-based waveform outperforms frame-based counterpart.
(2) Utterance-based waveform combines better with STOI (correlation).

Output Feature Types

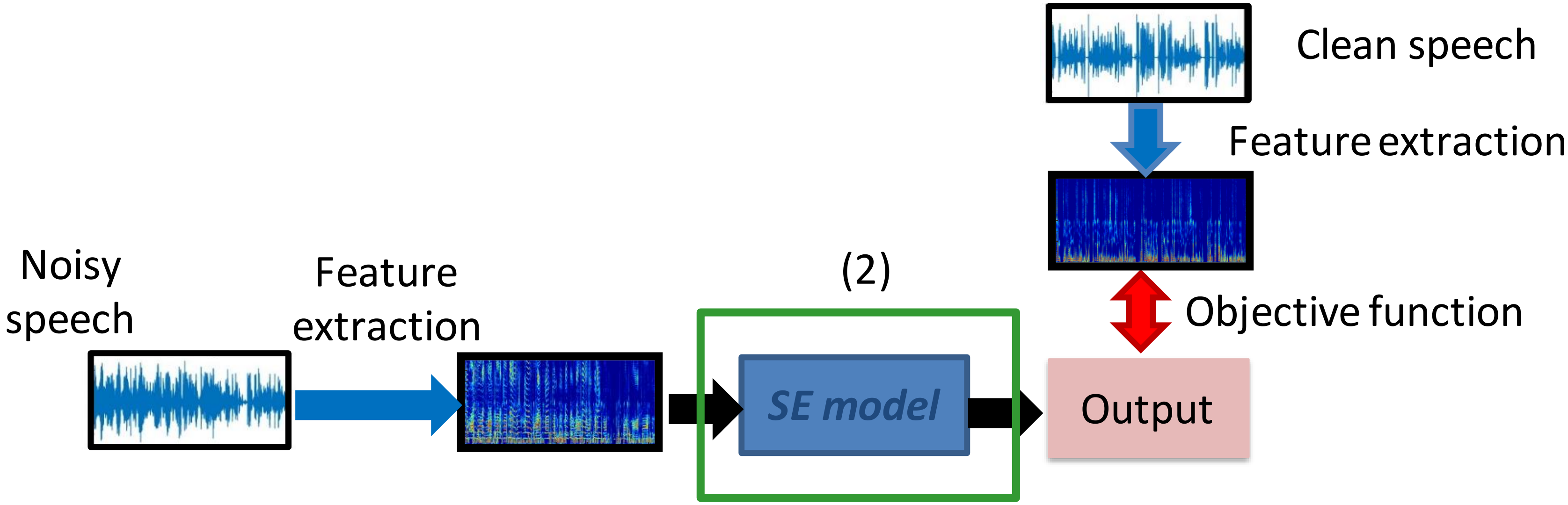
- Mapping vs. masking based SE : [Wang and Chen, TASLP 2018]



Outline

- Deep Learning (DL) based Speech Enhancement (SE)
  - ➢ System architecture
  - ➢ **Six factors need to consider**
    - ✓ Feature types
    - ✓ **Model types**

Model Types



Model types:

DNN [Wang et al. NIPS 2012; Xu et al., SPL 2014], DDAE [Lu et al., Interspeech 2013], RNN (LSTM) [Chen et al., Interspeech 2015; Weninger et al., LVA/ICA 2015], CNN [Fu et al., Interspeech 2016], CRNN [Zhao et al., ICASSP 2018], FCN [Fu et al, TASLP 2018], HELM [Hussain et al., IEEE Access 2017].

Advanced architecture:

Skip connection [Tu and Zhang ICASSP 2017], Highway [Santos and Falk, NIPS workshop 2018], Densely connection [Zhen et al., ICASSP 2019], Attention mechanism [Hao et al., ICASSP 2019], U-Net architecture [Pascual et al., Interspeech 2017], and Complex parameters [Y.-S. Lee et al., ICASSP 2017].

Outline

- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ Feature types
    - ✓ Model types
    - ✓ **Objective function**

聞

聽

大學曰：心不在焉，聽而不聞

Hear but pay no attention; listen but not hear

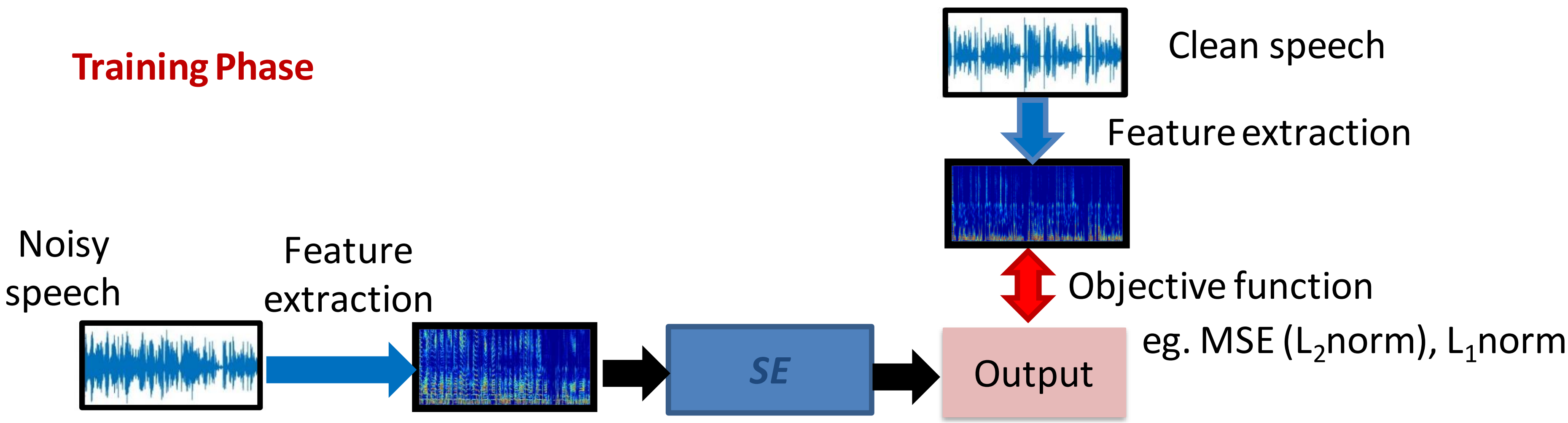
Intelligibility and Quality are different

Objective Function

Speech perception
- Quality
- Intelligibility

Automatic speech recognition
- Recognition accuracy

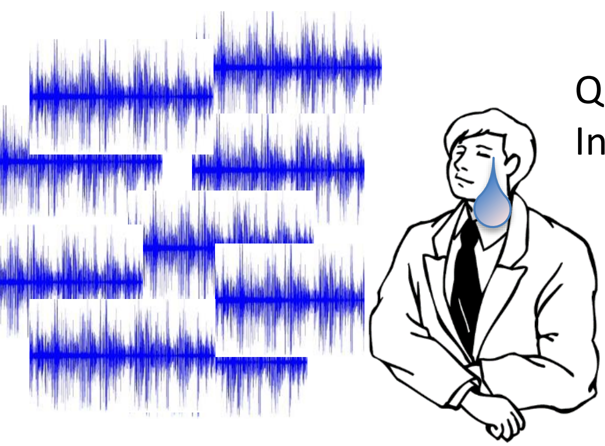
Training Phase



Mean squared error (MSE) and L1 losses aim to minimize the differences of enhanced and target and do not directly consider human perception and ASR performance.

Objective Function

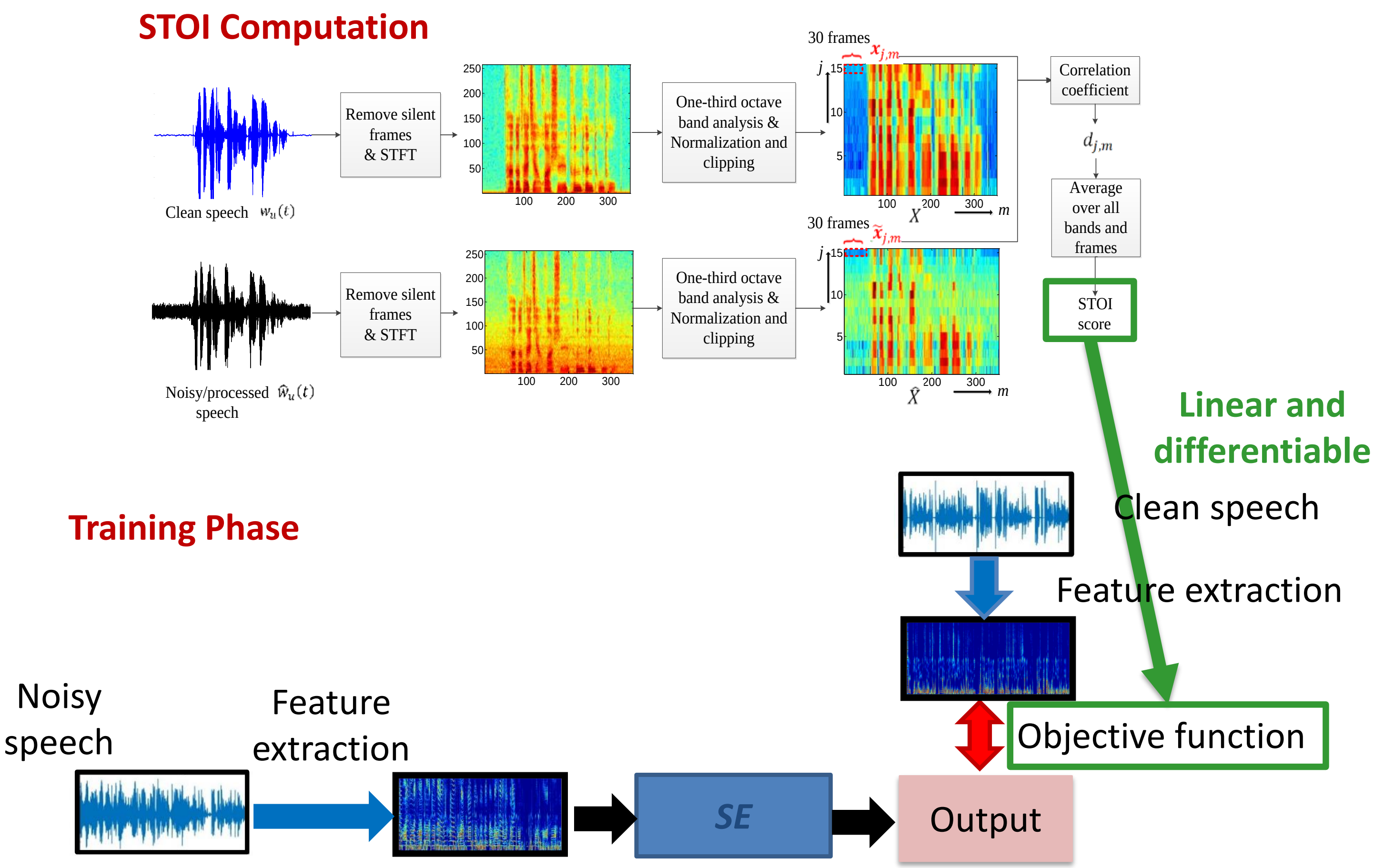
Speech perception
- Quality
- Intelligibility



(1) **STOI** was designed to compute the speech **intelligibility**, and the STOI score ranges from 0 to 1.
(2) **PESQ** was designed to evaluate the **quality** of processed speech, and the PESQ score ranges from -0.5 to 4.5.

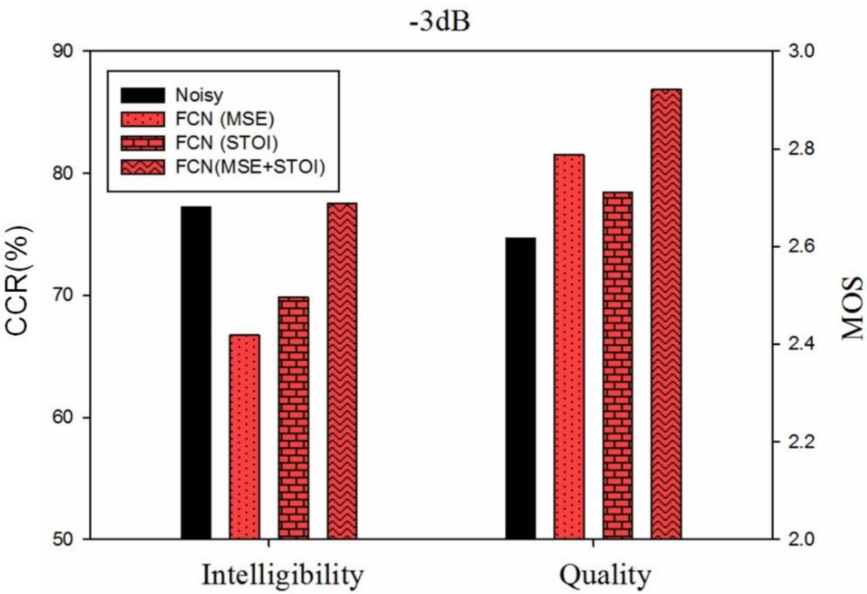
Objective Function

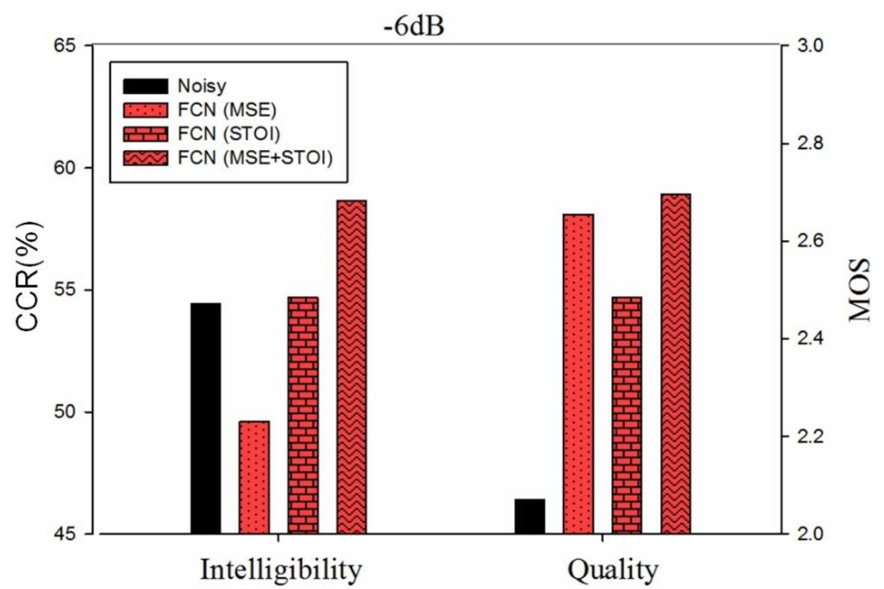
- STOI-based Objective Function [Fu et al, TASLP 2018]



Objective Function (STOI)

- Experimental Results (Human Listening Test)





Average character error rate (CCR) and quality scores (MOS) of human subjects for (a) −3 dB and (b) −6 dB SNR.

(1) Intelligibility: FCN (MSE+STOI)> FCN (STOI)>FCN (MSE);
(2) Quality: FCN (MSE+STOI) performs the best.

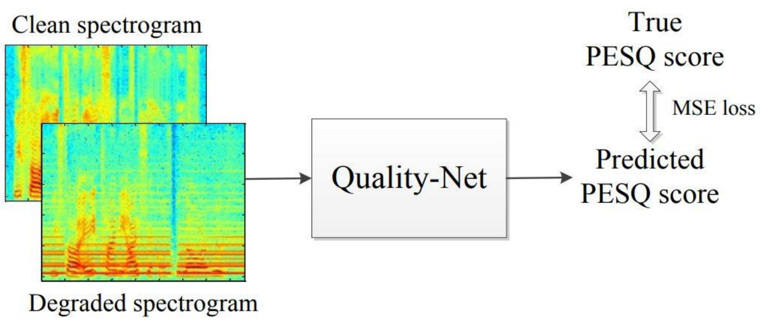
Objective Function

- PESQ-based Objective Function [Fu et al, IEEE SPL 2019]
  - However, when evaluation metrics are complicated and non-linear, such as PESQ (with more than 2700 lines in Matlab codes), it is difficult to directly derive an objective function using PESQ.
  - We can apply reinforcement learning (RL), where the PESQ score is used to form the reward function, to optimize the SE model [Koizumi et al, ICASSP 2017; Koizumi et al, TASLP 2018].
  - We can use direction sampling [Zhang et al., ICASSP 2018].
  - We can approximate the PESQ function and make it differentiable to update the SE model [Martin-Donas et al, IEEE SPL 2018].
  - Recently, we proposed a two-step strategy: (1) learn a deep learning model, Quality-Net, that can predict PESQ scores; (2) train the SE model based on the learned Quality-Net [Fu et al, IEEE SPL 2019].

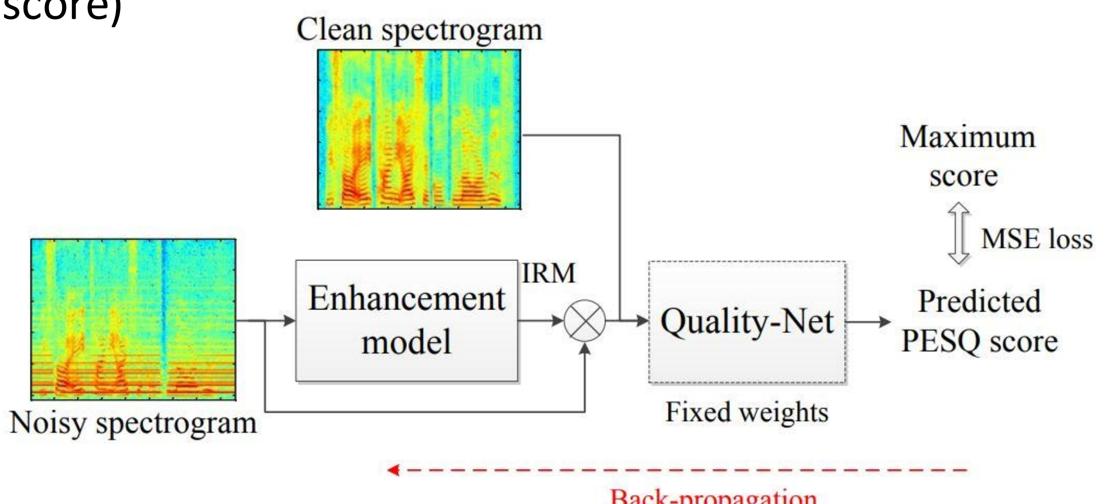
Objective Function

- PESQ-based Objective Function [Fu et al, IEEE SPL Accepted]

Stage 1: train a Quality-Net (input: paired clean and noisy speech; output: PESQ score)

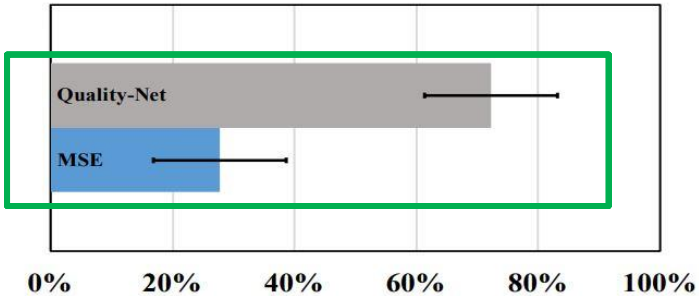


Stage 2: train the SE model based on the Quality-Net (input: paired clean and noisy speech; output: PESQ score)



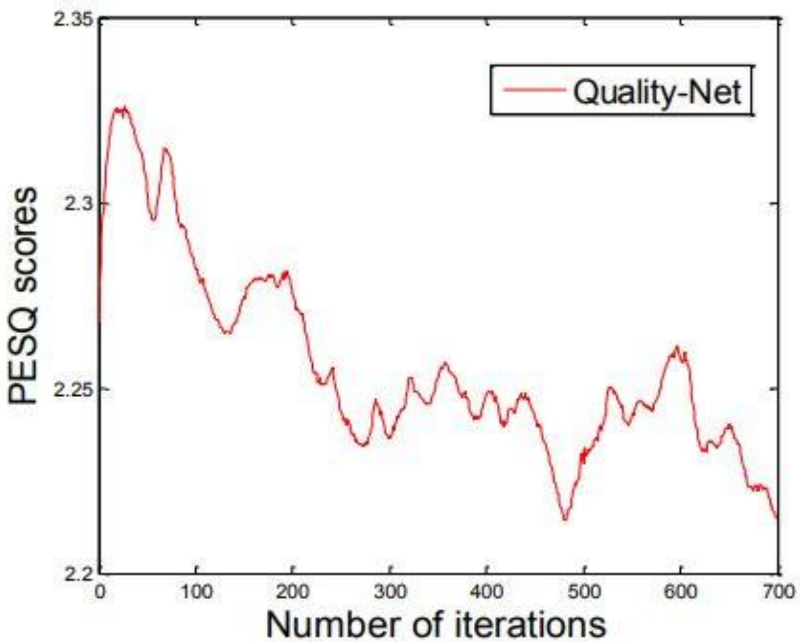
Objective Function (Quality-Net)

- Subjective and objective tests



	PESQ	STOI
Noisy	1.970	0.820
Wiener filter	2.223	0.914
BLSTM (MSE)	2.529	**0.935**
BLSTM (Quality-Net)	**2.713**	0.932

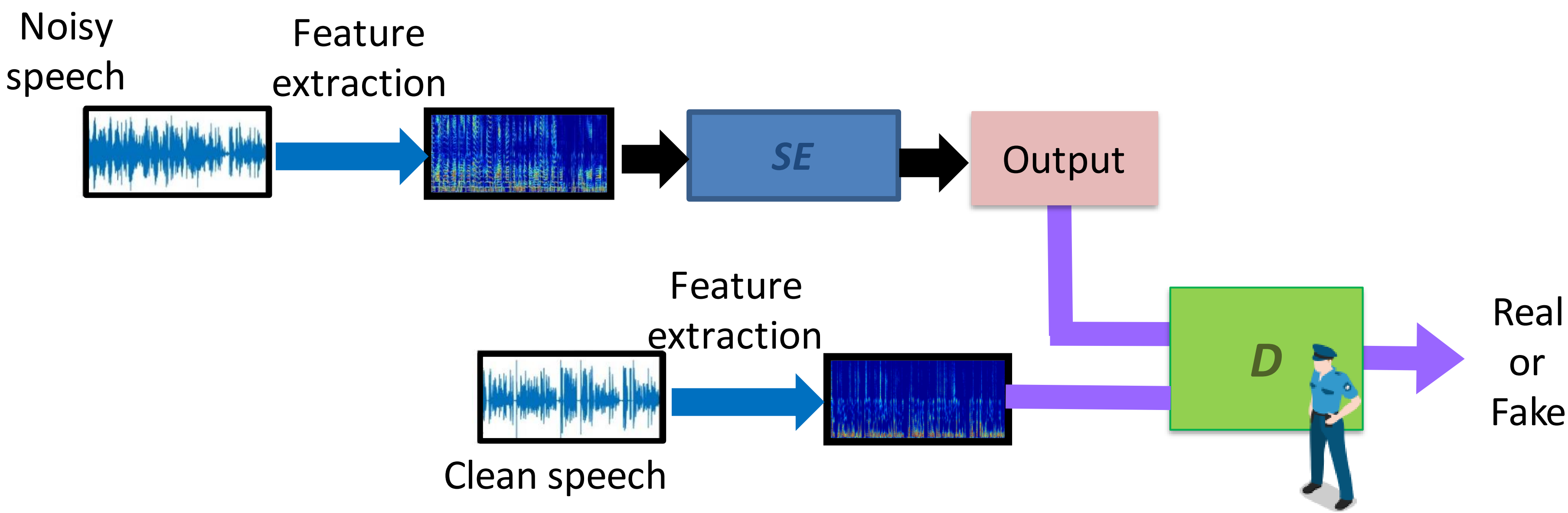
Listening tests and objective evaluation results



(1) The proposed method achieves the best MOS and PESQ scores.
(2) When we train the system more iterations, the true PESQ scores start to **decrease**.
(3) The Quality-Net **has not seen** the speech generated by the **updated enhancement model**.
(4) We further propose a **GAN-based** approach to handle this issue.

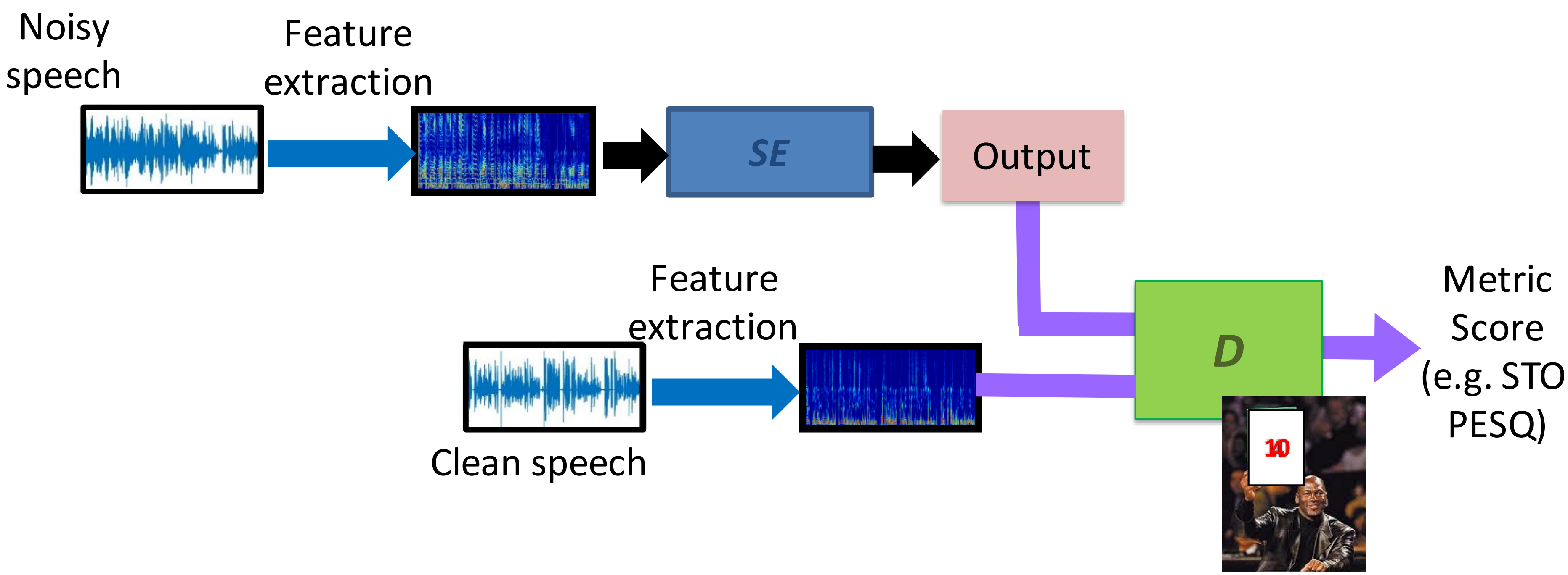
Objective Function

- Generative Adversarial Networks (GAN) based Methods: SEGAN [Pascual et al., Interspeech 2017]; Pix2Pix [Michelsanti et al., Interpsech 2017]; Mask estimation [Pandey and Wang, ICASSP 2018; Neil et al., APSIPA 2018]



Objective Function

- MetricGAN [Fu et al., ICML 2019]



Objective Function

- Conditional GAN (CGAN) versus MetricGAN [Fu et al., ICML 2019]

Discriminator in CGAN (LSGAN):

$$L_D(CGAN) = \mathrm{E}_{x,y}[(D(y,x)-1)^2 + (D(G(x),x)-0)^2]$$

where *x* and *y* are noisy and clean speech, respectively.

Discriminator in MetricGAN:

$$L_D(MetricGAN) = \mathrm{E}_{x,y}[(D(y,y)-1)^2 + \big(D(G(x),y) - Q'(G(x),y)\big)^2]$$

$0 \leq Q'(G(x),y) < 1$ is the normalized evaluation metric (1 represents the highest evaluation score).

(1) For CGAN, *D* tries to distinguish real and enhanced samples.
(2) For MetricGAN, *D* tries to learn the PESQ\STOI function.

Objective Function

- Conditional GAN (CGAN) versus MetricGAN [Fu et al., ICML 2019]

Generator in CGAN (LSGAN):

$$L_G(CGAN) = \mathrm{E}_x[\lambda(D(G(x),x) - 1)^2] + \|G(x) - y\|_1$$

where *x* and *y* are noisy and clean speech, respectively.

Generator in MetricGAN:

$$L_G(MetricGAN) = \mathrm{E}_x[(D(G(x),y) - \boxed{s})^2]$$

where *s* is the desired assigned score.

(1) We can specify any particular score *s*.
(2) With a large number *s* (e.g.,1), we get a speech **enhancement** model.
(3) With a small number *s* (e.g., 0), we get a speech **degradation** model.

Objective Function (MetricGAN)

- MetricGAN (P) and MetricGAN (S) with related works

Performance comparisons on TIMIT of different methods in terms of PESQ & STOI

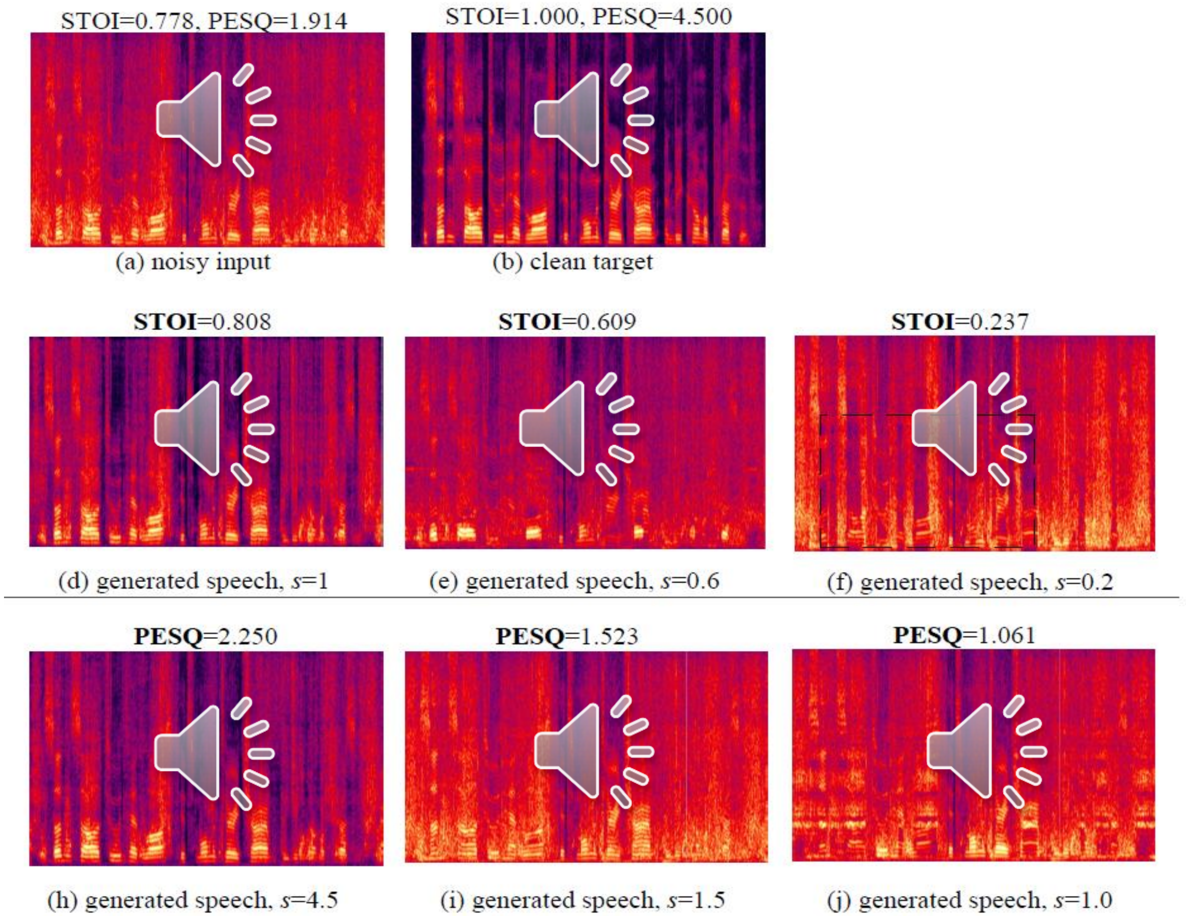
	Noisy		IRM (L1)		IRM (CGAN)		PE policy grad*(P)		MetricGAN (P)		MetricGAN (S)	
SNR (dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
12	2.375	0.919	2.913	0.935	2.879	0.936	**2.995**	0.927	2.967	0.936	2.864	**0.939**
6	1.963	0.831	2.52	0.878	2.479	0.876	2.595	0.869	**2.616**	0.881	2.486	**0.885**
0	1.589	0.709	2.086	0.787	2.053	0.786	2.144	0.776	**2.200**	0.796	2.086	**0.802**
-6	1.242	0.576	1.583	0.655	1.551	0.653	1.634	0.644	**1.711**	0.668	1.599	**0.679**
-12	0.971	0.473	1.061	0.508	1.046	0.507	1.124	0.500	**1.169**	0.521	1.090	**0.533**
Avg.	1.628	0.702	2.033	0.753	2.002	0.751	2.098	0.743	**2.133**	0.760	2.025	**0.768**

(P: PESQ) (S: STOI)

(1) GAN is not helpful for this task (TIMIT).
(2) MetricGAN (P) achieves the best PESQ (quality) scores.
(3) MetricGAN (S) achieves the best STOI (intelligibility) scores.

Objective Function (MetricGAN)

- Arbitrary target scores



Results of assigning different scores (*s*) for the generator training.

We can specify a metric score to increase or decrease the speech quality or ineligibility.

Objective Function (MetricGAN)

- Objective evaluation on the Voice Bank corpora

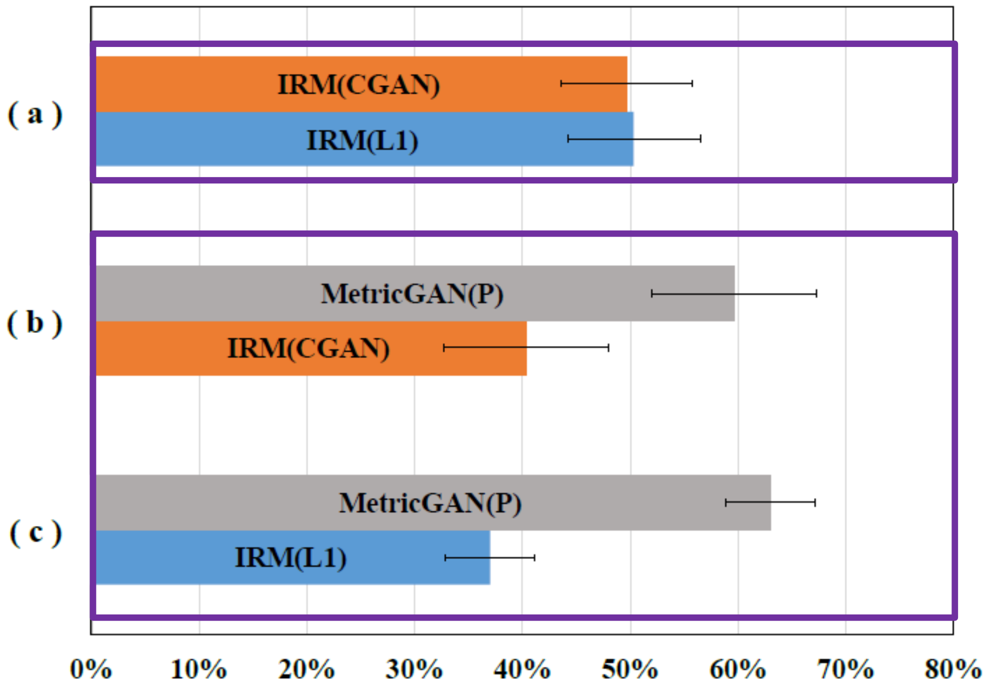
Comparisons of MetricGAN with other state-of-the-art methods. The highest score per metric is highlighted with bold text.

	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
SEGAN	2.16	3.48	2.94	2.80
MMSE-GAN	2.53	3.80	3.12	3.14
WGAN-GP	2.54	-	-	-
Deep Feature Loss	-	3.86	**3.33**	3.22
SERGAN	2.62	-	-	-
MetricGAN (P)	**2.86**	**3.99**	3.18	**3.42**

(1) MetricGAN outperforms CGAN (SEGAN) for all the evaluation metrics.
(2) MetricGAN achieves the state-of-the-art results for PESQ and other metrics.

Objective Function (MetricGAN)

- Subjective listening tests obtained from 15 listeners

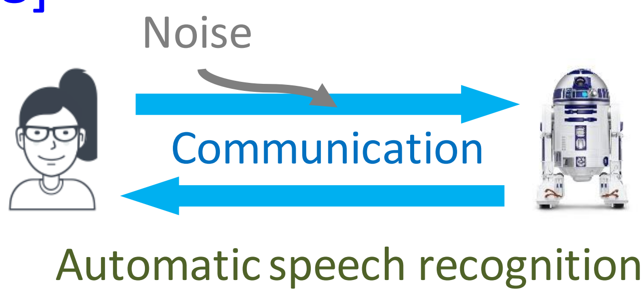


Results of AB preference test (with 95% confidence intervals) on speech quality compared between proposed MetricGAN(P) and the two baseline models.

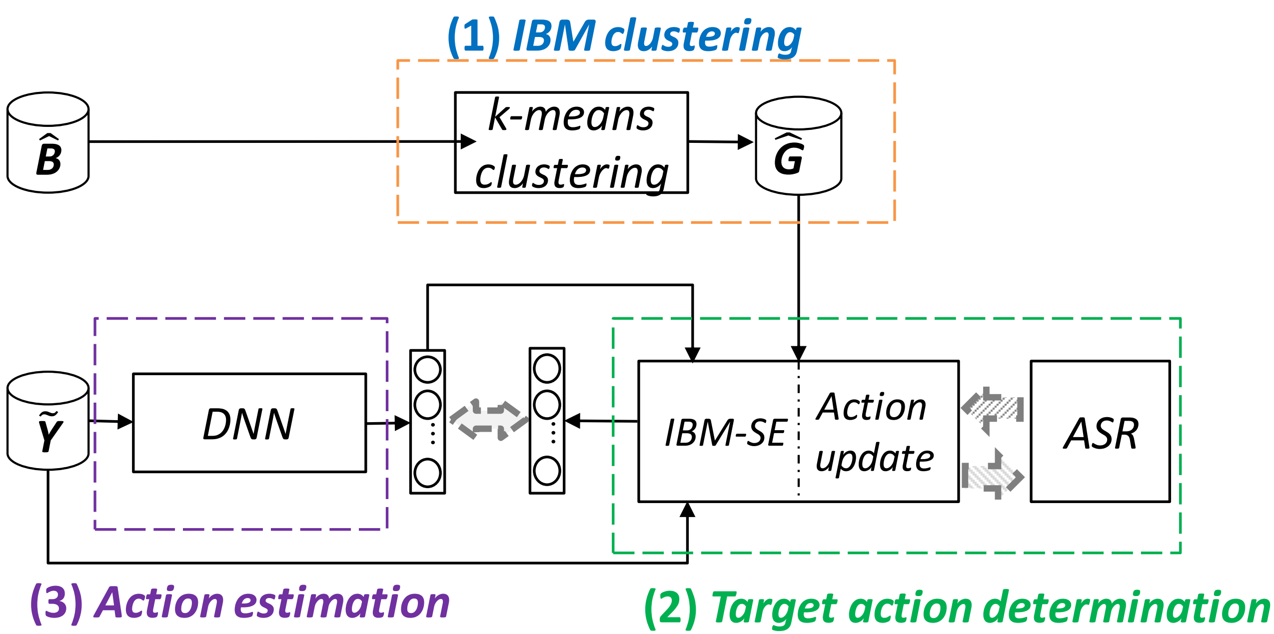
(1) CGAN does not give clearly better listening test results than L1 norm.
(2) MetricGAN(P) outperforms SE models trained by CGAN and L1 norms.

Objective Function

- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]

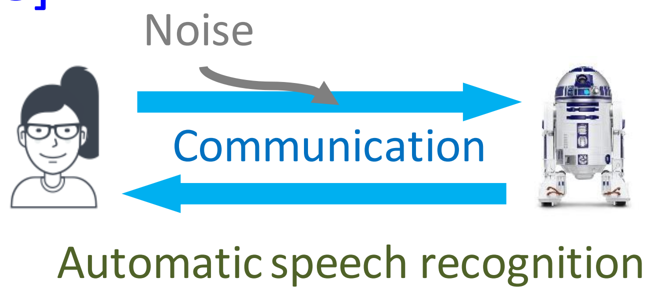


- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system

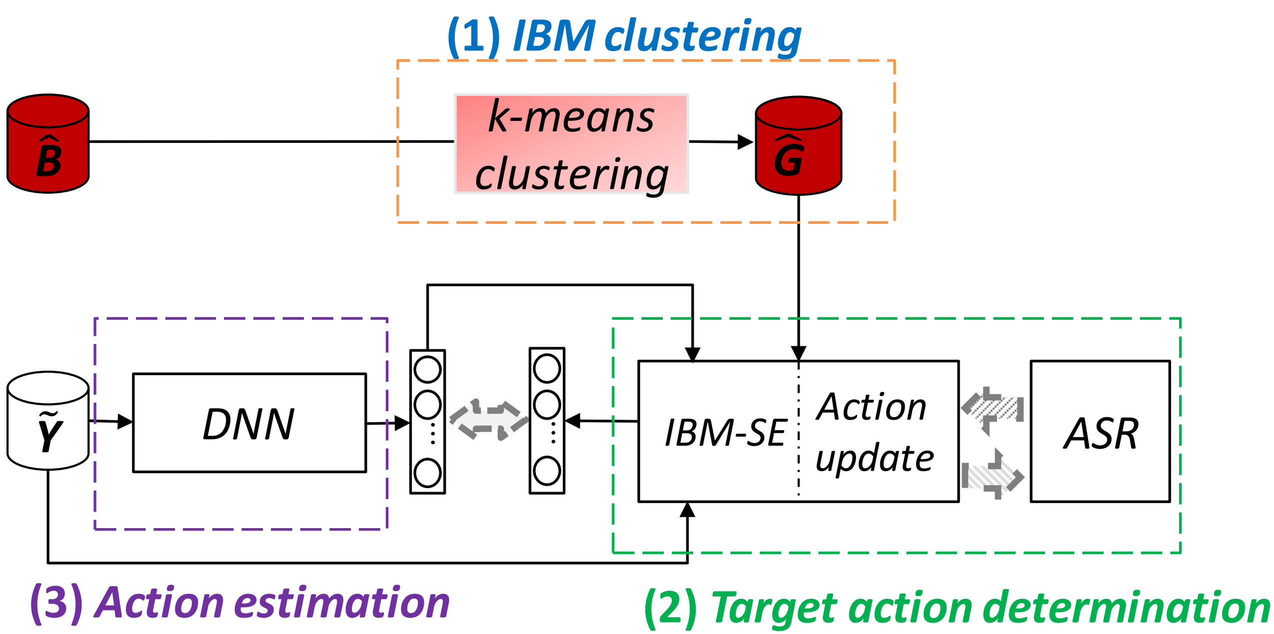


Objective Function

- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]



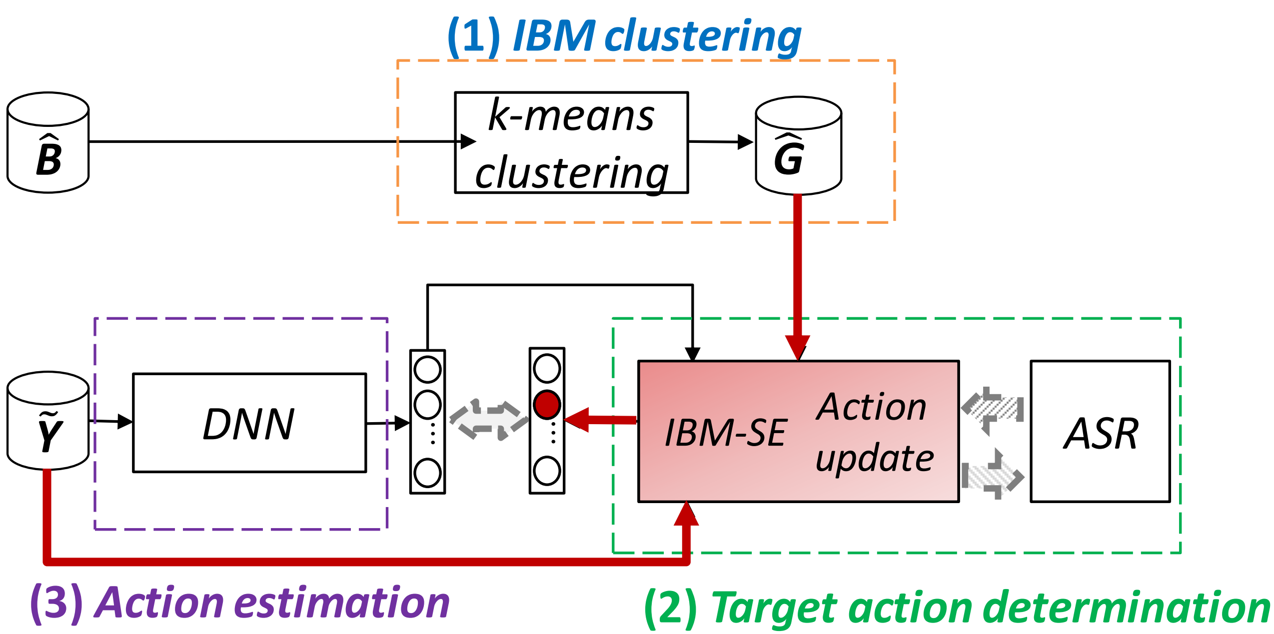
- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system



Objective Function

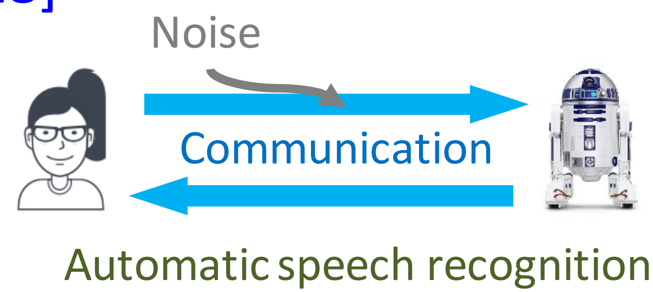
- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]

- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system

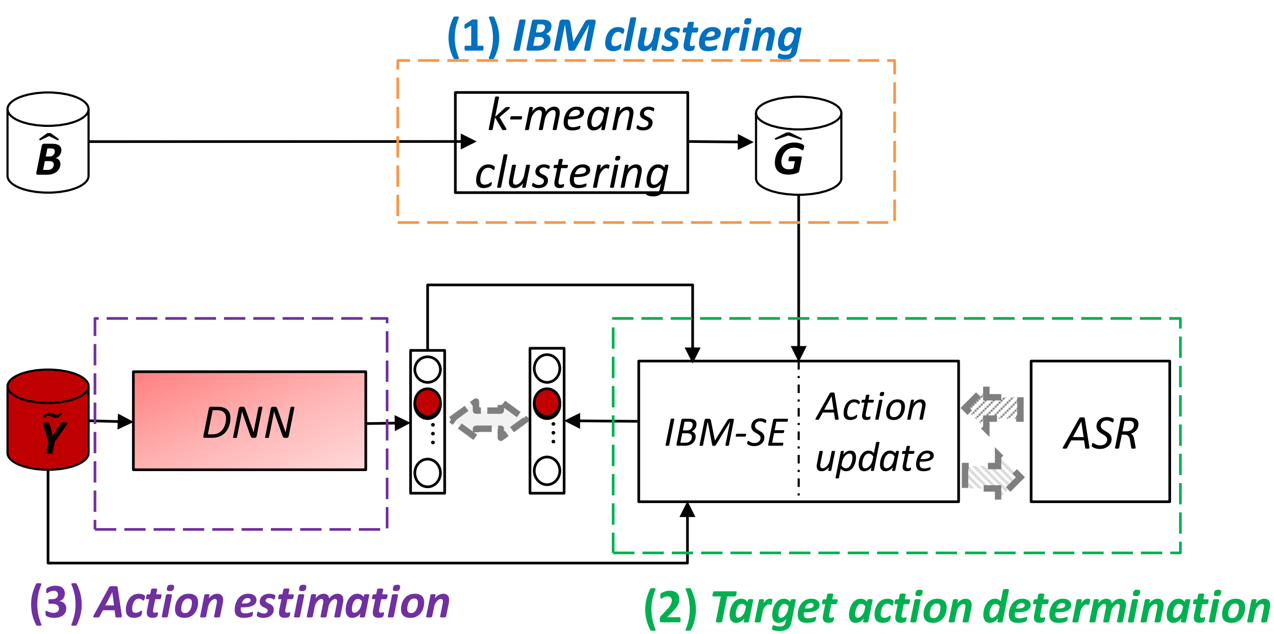


Objective Function

- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]



- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system



Objective Function (RLSE)

- Results on ASR and STOI and PESQ

The average CERs of Noisy (the baseline), *1nnSE*, *$RLSE_1$*, and *$RLSE_2$* at 0 and 5 dB SNR conditions.

SNR	*Noisy*	$1nnSE$	$RLSE_1$	$RLSE_2$
5 dB	56.14	73.09	55.60	**49.18**
0 dB	81.40	85.79	77.20	**65.75**

The average STOI and PESQ of Noisy (the baseline), *$RLSE_1$*, and *$RLSE_2$* at 0 and 5 dB SNR conditions.

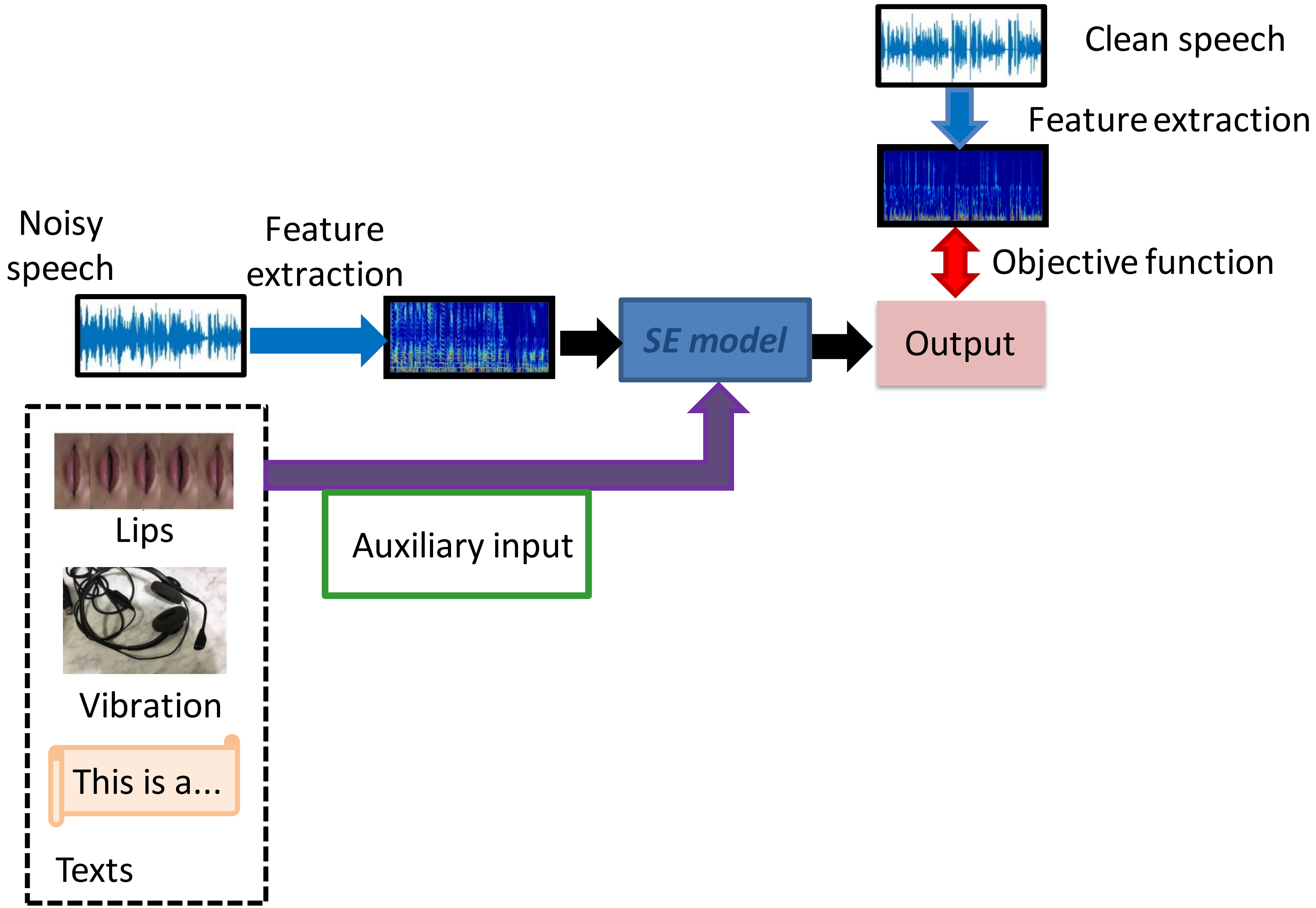
SNR	STOI			PESQ		
	Noisy	$RLSE_1$	$RLSE_2$	*Noisy*	$RLSE_1$	$RLSE_2$
5 dB	0.82	0.82	**0.86**	1.85	1.67	**1.96**
0 dB	0.74	0.77	**0.81**	1.45	1.42	**1.59**

(1) The training can be done without noisy-clean paired speech.
(2) Speech recognition accuracy-based objective function improves ASR performance and objective measures (human listening).

Outline

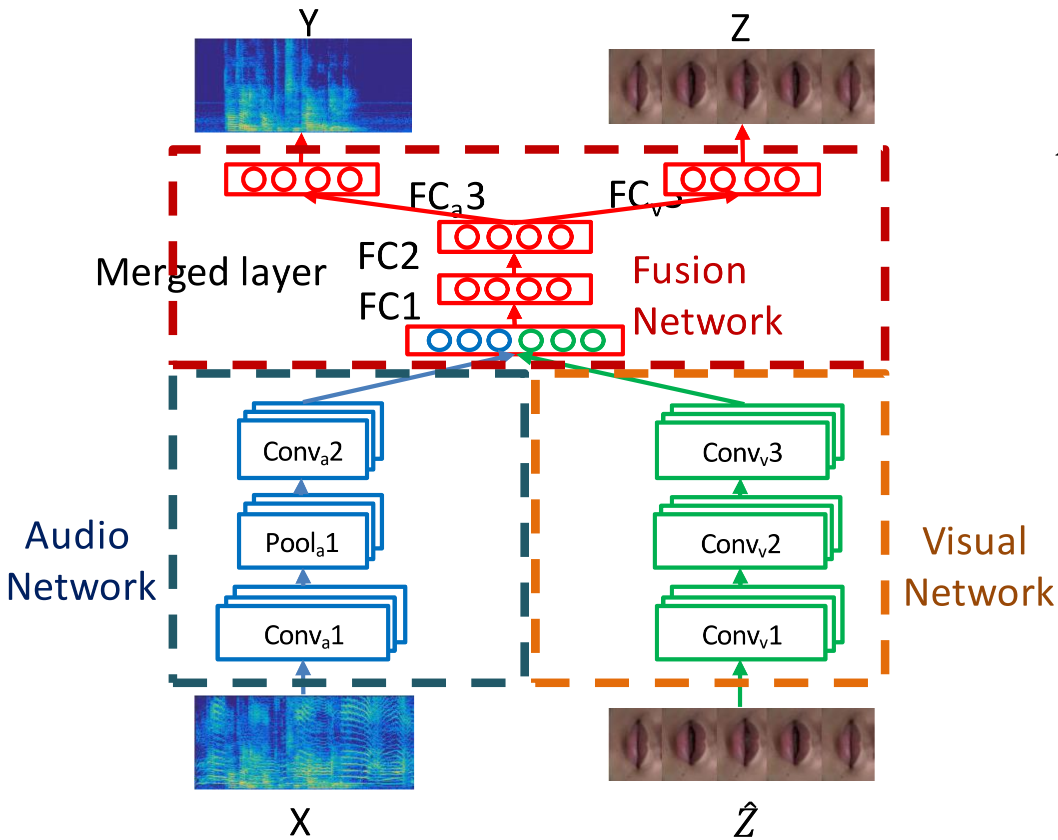
- Deep Learning based Speech Enhancement
  - ➢ System architecture
  - ➢ **Six factors need to consider**
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ **Auxiliary input**

Auxiliary Input



Auxiliary Input

- Audio-visual SE [Hou et al., TETCI 2018]



$$A_i = Conv_a2\left(Pool_a1\left(Conv_a1(X_i)\right)\right),$$

$$V_i = Conv_v3\left(Conv_v2\left(Conv_v1(\hat{Z}_i)\right)\right),$$

$$F_i = [A_i{}' \;\; V_i{}']',$$

$$Y_i = FC_a3\left(FC2\left(FC1(F_i)\right)\right),$$

$$Z_i = FC_v3\left(FC2\left(FC1(F_i)\right)\right),$$

where $i = 1 \ldots K$,

K is the number of training samples.

Objective function:

$$\min\left(\frac{1}{K}\sum_{i=1}^{K}\left\|Y_i - \hat{Y}_i\right\|_2^2 + \mu\left\|Z_i - \hat{Z}_i\right\|_2^2\right)$$, where μ is a mixing weight.

Auxiliary Input (AVSE)

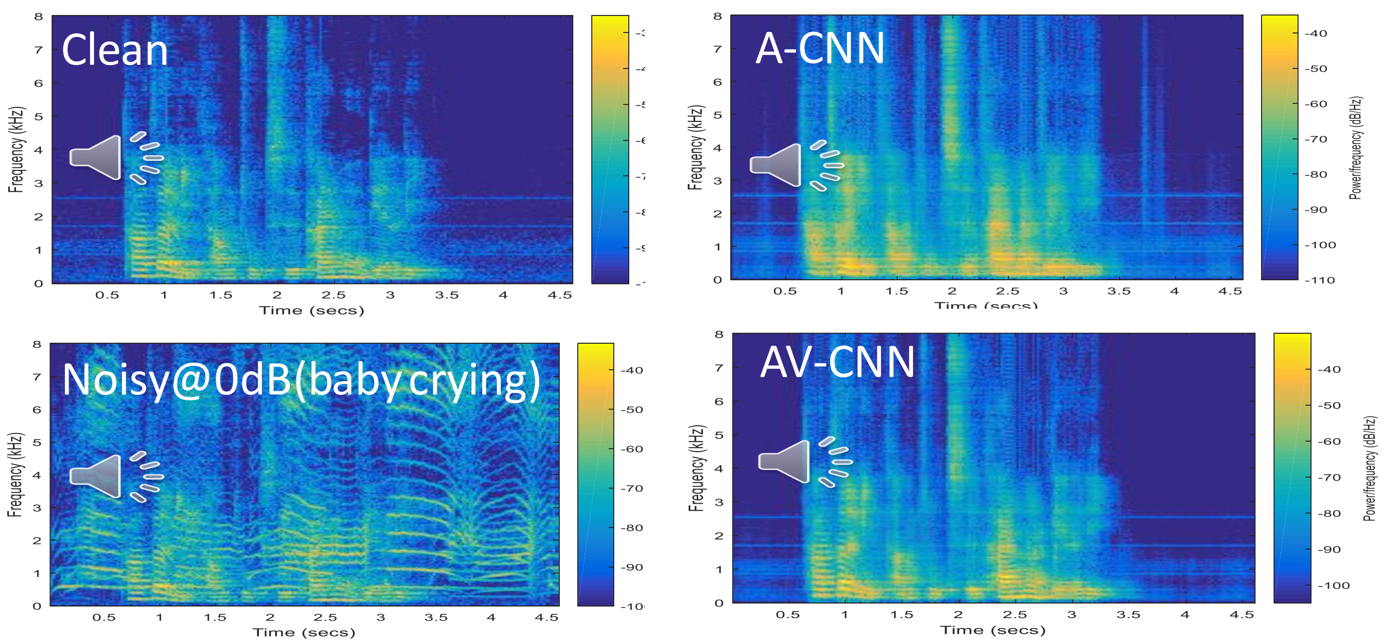
- Audio-visual versus audio only [Hou et al., TETCI 2018]

Experimental setup

- Single speaker (speaker-dependent case), 320 clean utterances (training:280, testing:40)
- Training: 91 types of interference noise at 10dB, 6dB, 2dB, -2dB, -6dB SIRs
  + ambient noise (car engine noise) at 10dB, 6dB, 2dB, -2dB, -6dB SARs
- Testing: 10 types of interference noise at 5dB, 0dB, -5dB SIRs
  + ambient noise (car engine noise) at 5dB, 0dB, -5dB SARs

*SIRs: signal-to-interference noise ratios, SARs: signal-to-ambient noise ratios.

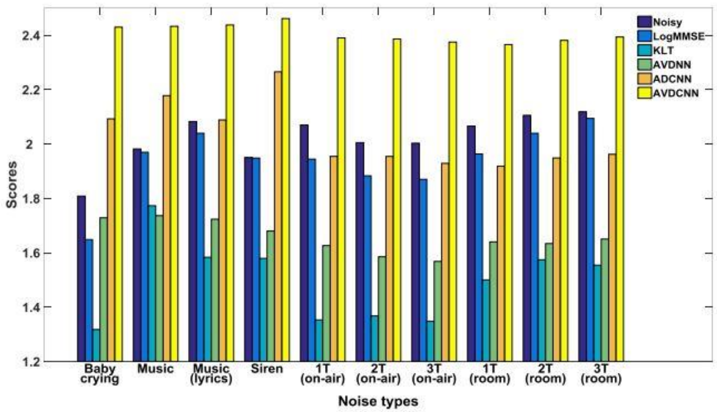
Spectrogram comparison



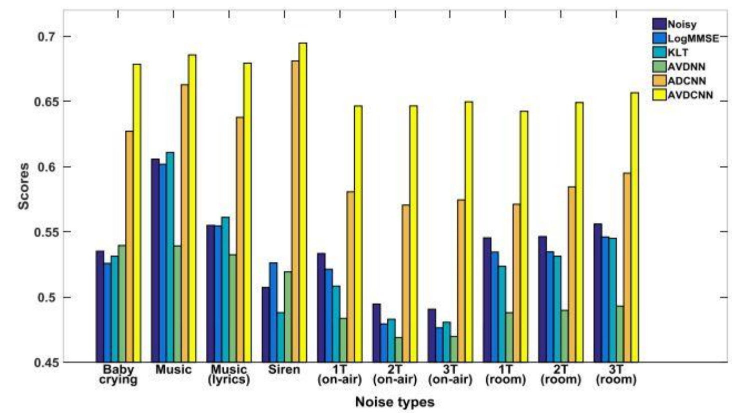
Auxiliary Input (AVSE)

- Audio-visual versus audio only [Hou et al., TETCI 2018]

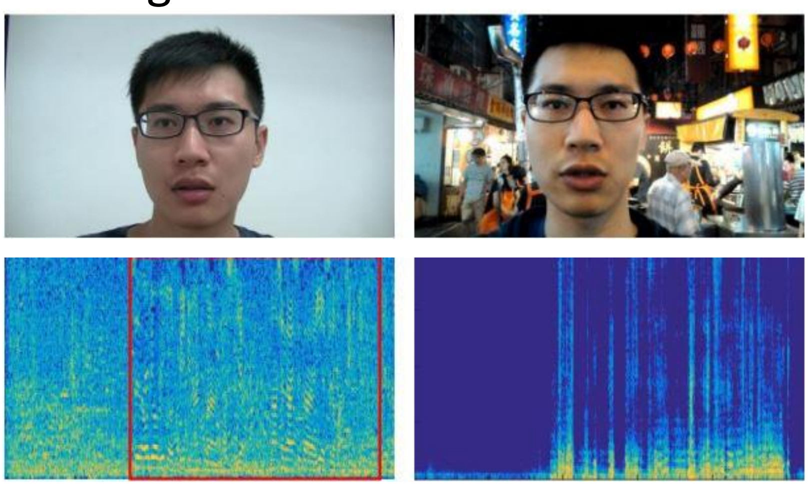
The PESQ scores



The STOI scores



Testing in the real-world conditions



(1) Visual information improves the SE performance.
(2) The performance is robust against recording conditions as long as lips can be recorded well.

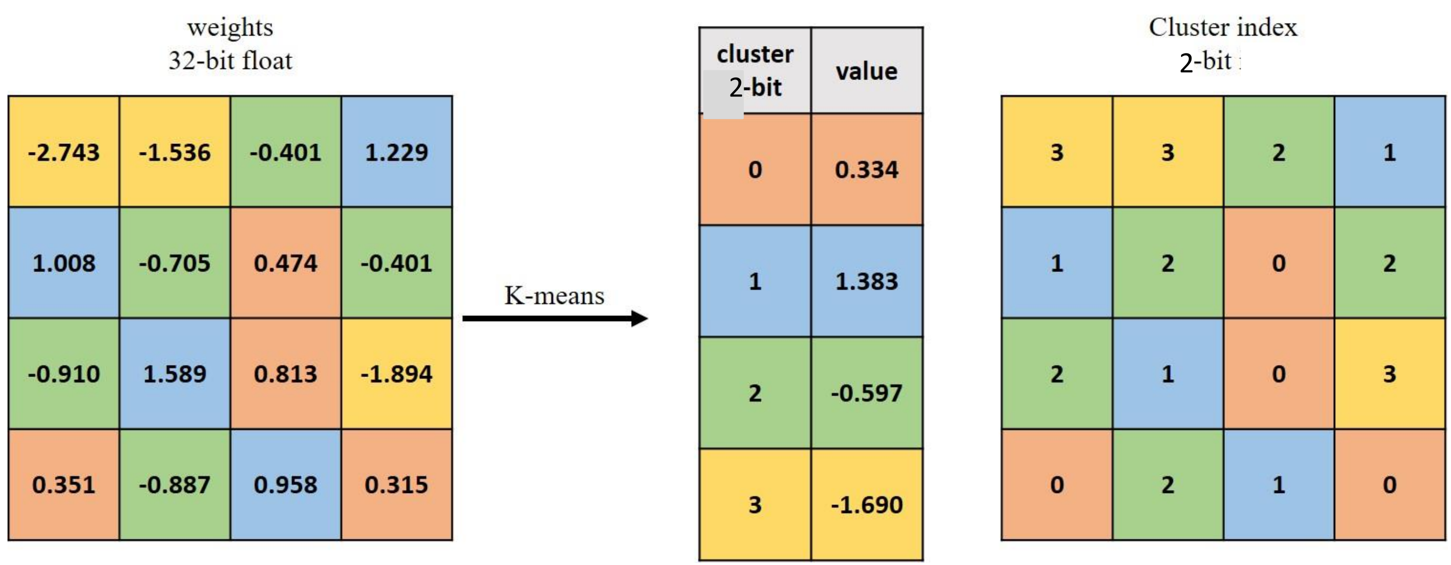
Outline

- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ **Model compression**

https://www.vology.com/resource/benefits-of-edge-computing/

Model Compression

- Weight sharing (WS) based on K-means
  - Clustering weights into ***c*** **clusters** with K-means algorithm.
  - Replacing 32-bit weights with $(\log_2 c)$-bit cluster index; each index represent a specific **cluster centroid**; the same cluster share the same centroid.



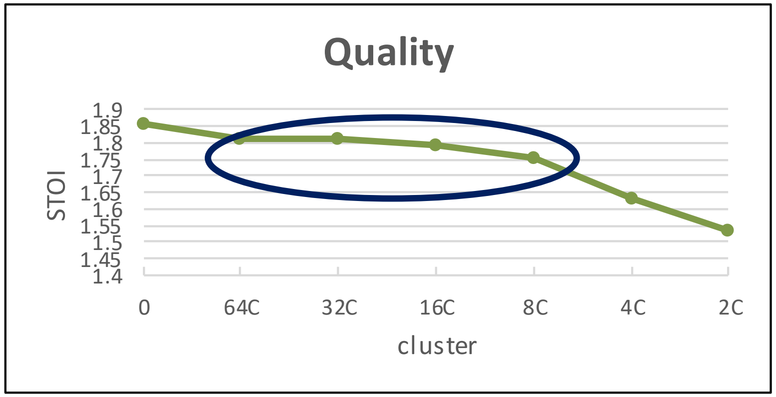
16×32-bit =(512) bits   4×32-bit + 16×2-bit =(160) bits

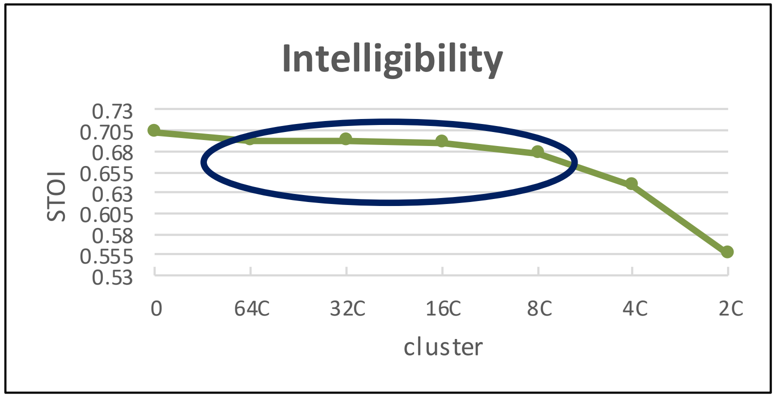
Model Compression (WS-SE)

- WS for SE model [Wu et al., IEEE SPL Accepted]

Cluster: 64, 32, 16, 8, 4, 2;
cluster = 0 is original model.

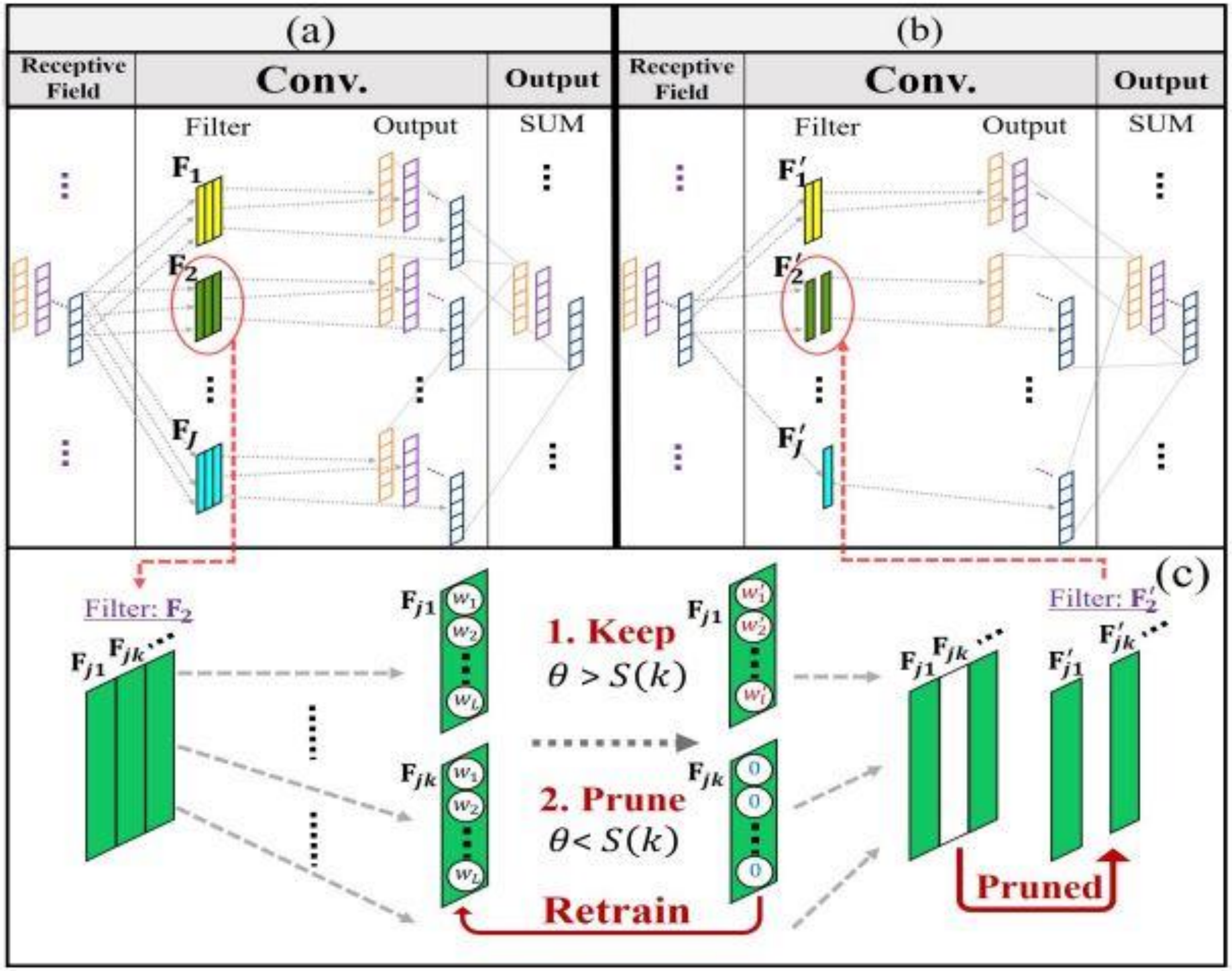
cluster	PESQ	STOI
0	1.85385	0.70231
64C	1.8063	0.6941
32C	1.7967	0.6927
16C	1.8088	0.6896
8C	1.7606	0.6786
4C	1.5852	0.6269
2C	1.4558	0.5568
Noisy	1.63713	0.66977





(1) Performance does not change much when the cluster number increases from 0 to 16.
(2) However, the performance drops significantly when K> 16.

Model Compression



Model Compression

- CPO performs filter pruning to reduce model size and online computational costs [Wu et al., IEEE SPL 2019].
- Three steps in CPO:

  (1) For a specified channel *c* in a conv layer, the **mean value** of all **absolute filter weights** at that channel is computed:

$$M_c = \frac{\sum_{n,w} |k_{nw}|}{N \times W} \qquad \begin{array}{l} N: filter\ number \\ W: weight\ number \end{array}$$

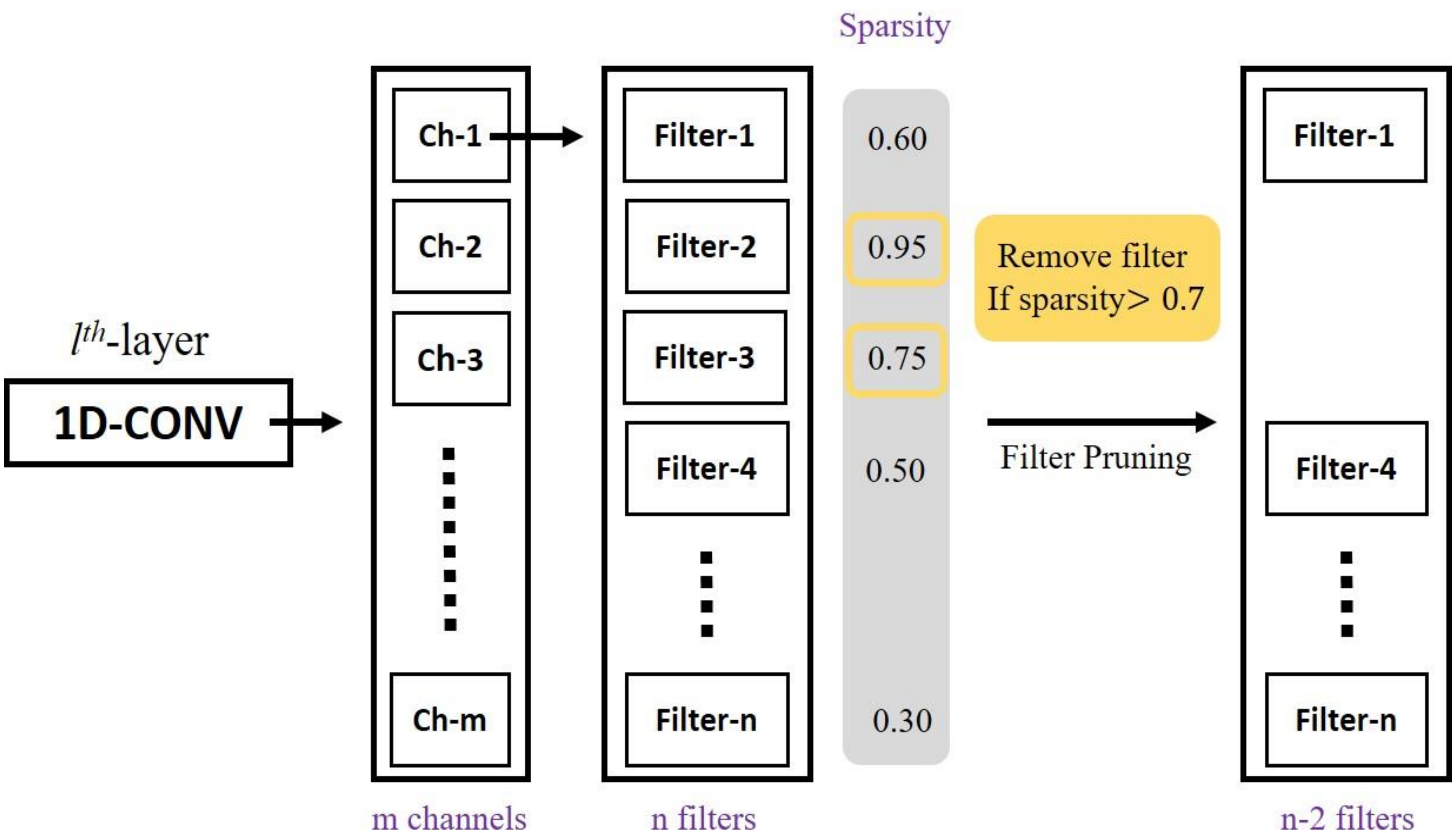
  (2) Compute the $\boldsymbol{sparsity}$ of the *n*-th filter:

$$S(n) = \frac{\sum_{w} \sigma(k_w)}{W}, \qquad \sigma(x) = \begin{cases} 1, if\ x < M_c \\ 0, otherwise \end{cases}$$

  (3) **A Threshold Θ** is specified. If $sparsity > \Theta$, the filter will be removed.

Model Compression

- CPO for SE model [Wu et al., IEEE SPL 2019]



$N: filter\ number$

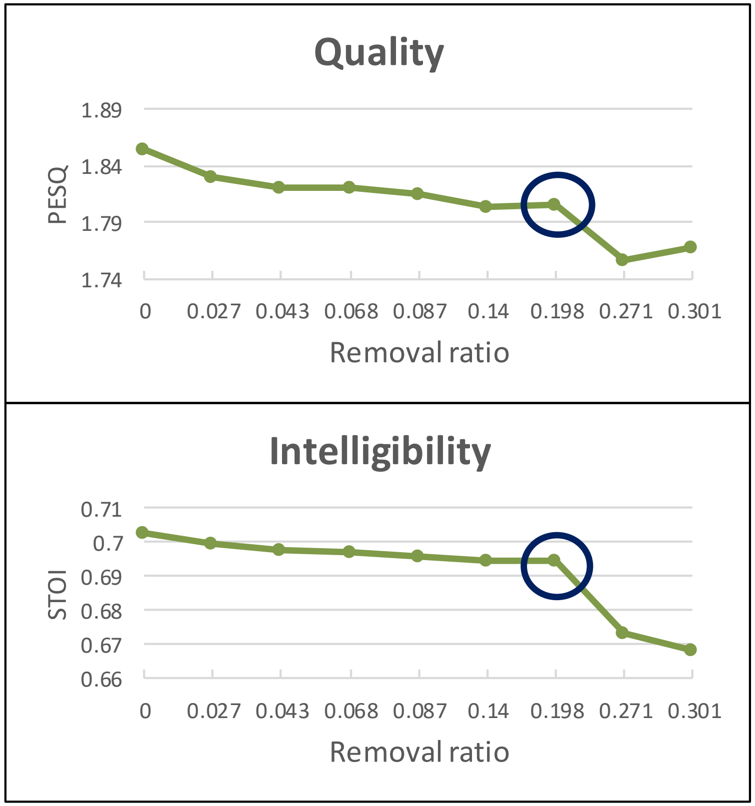
$W: weight\ number$

Model Compression (CPO-SE)

- The results of CPO

A Threshold Θ is specified
If $sparsity > \Theta$, the filter will be removed

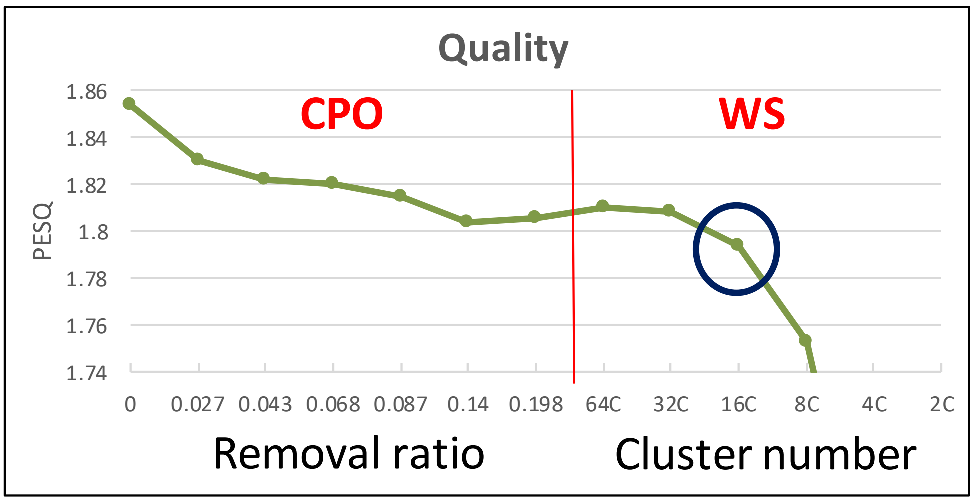
Threshold	Removal ratio	PESQ	STOI
1.0	0	1.85385	0.70231
0.95	0.027	1.83	0.6995
0.9	0.043	1.8215	0.6975
0.85	0.068	1.8197	0.697
0.8	0.087	1.8147	0.6957
0.75	0.14	1.8034	0.6941
0.7	0.198	1.805	0.6943
0.65	0.271	1.7558	0.673
0.6	0.301	1.7687	0.6683
Noisy		1.63713	0.66977

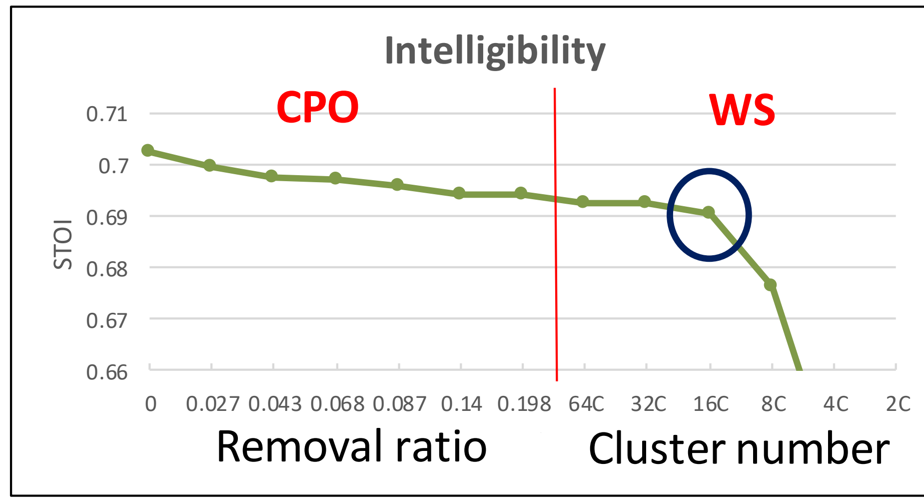


A notable performance drop when Threshold <0.65.

Model Compression (CPO+WS SE)

- The results of CPO+WS
  - We first define the expected performance loss ratio (=0.95)
  - Gradually reducing the Threshold (removal ratio = 20%)
  - Gradually decreasing the number of clusters (C = 16)

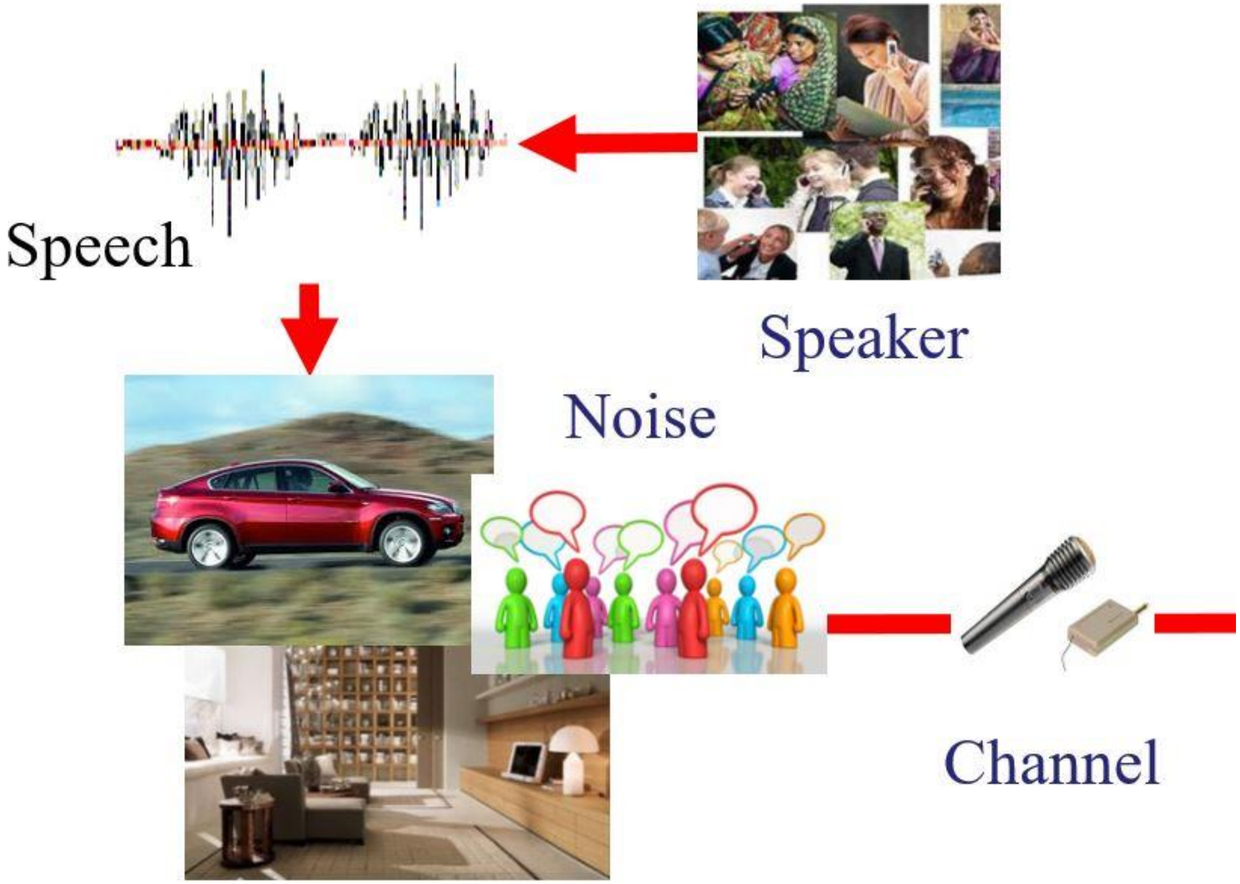




(1) The **model size** of the compressed model is only **9.76%** as compared to the original model.
(2) The **computation cost** is reduced by **20%**.

Outline

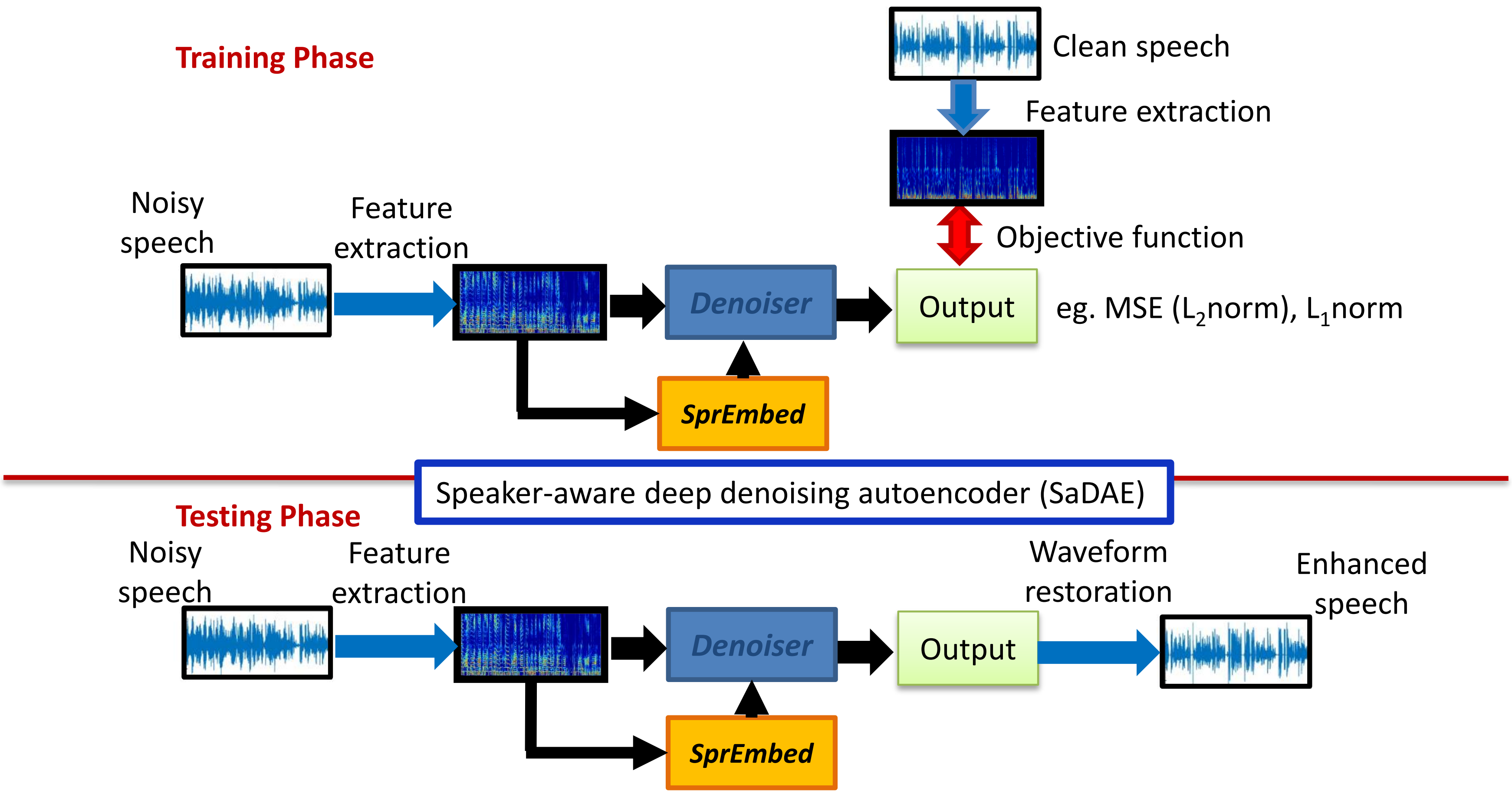
- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ **Increasing adaptability**



Speech
Speaker
Noise
Channel

Speaker Adaptability

- Speaker-aware Deep Autoencoder (SaDAE)
  [Chuang et al., Interspeech 2019]



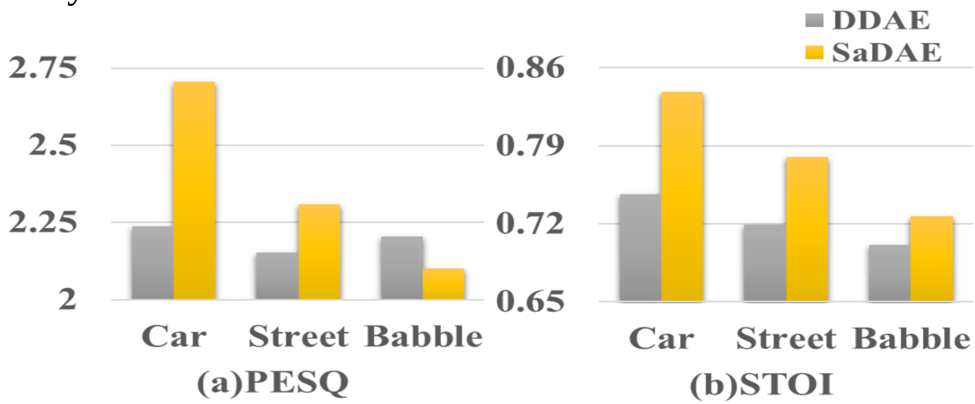
Speaker Adaptability (SaDAE)

- The results of SaDAE

The averaged PESQ, STOI and SDI results over all noisy utterances in the test set.

Testing	PESQ	STOI	SDI.
Noisy	2.0280	0.7493	1.1450
DDAE	2.1987	0.7225	0.7501
SaDAE	**2.3715**	**0.7815**	**0.3228**

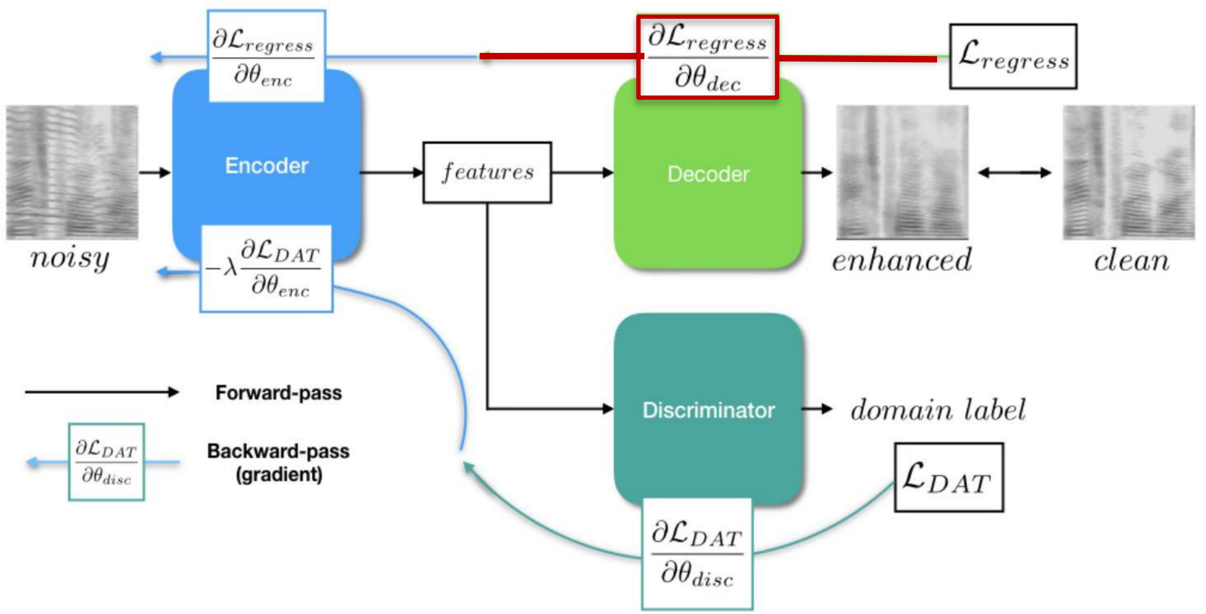
The averaged PESQ and STOI results over noisy utterances with respect to three noisy environments.



SaDAE outperforms conventional DDAE for both PESQ and STOI.

Environment Adaptability

- Noise-adaptive DAT (NADAT) [Liao et al., Interspeech 2019]



$$\theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G}$$ Min reconstruction error

$$\theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D}$$ Max domain accuracy

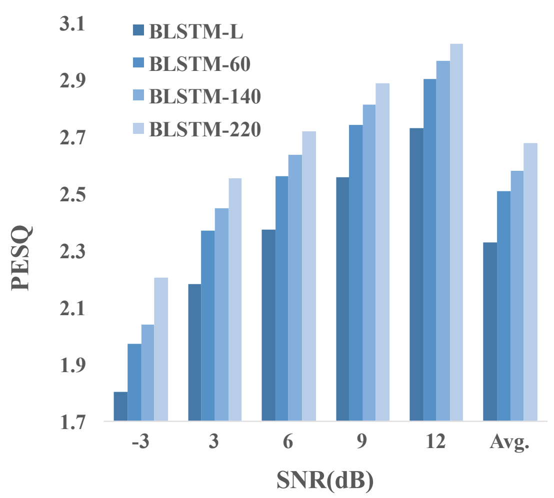
$$\theta_E \leftarrow \theta_E - \epsilon \left(\frac{\partial V_y}{\partial \theta_E} \right) + \alpha \frac{\partial V_z}{\partial \theta_E}$$

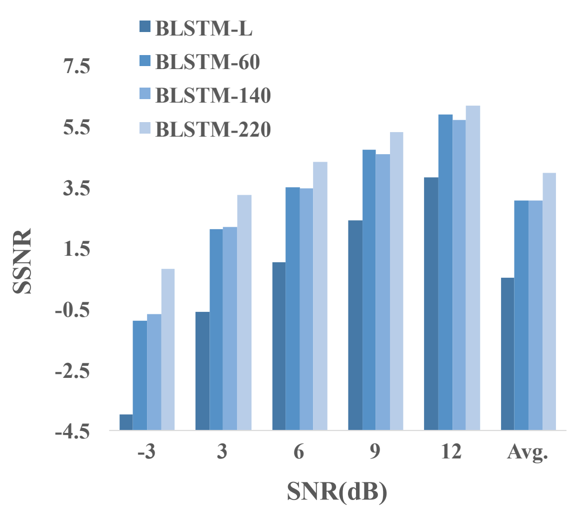
Min reconstruction error and Min domain accuracy

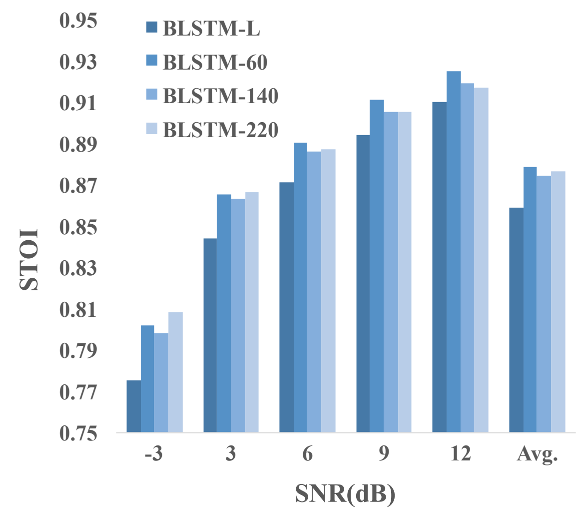
Environment Adaptability (NADAT)

- Adapting to new noise type (Baby cry)

Different amount of target domain data for adaptation





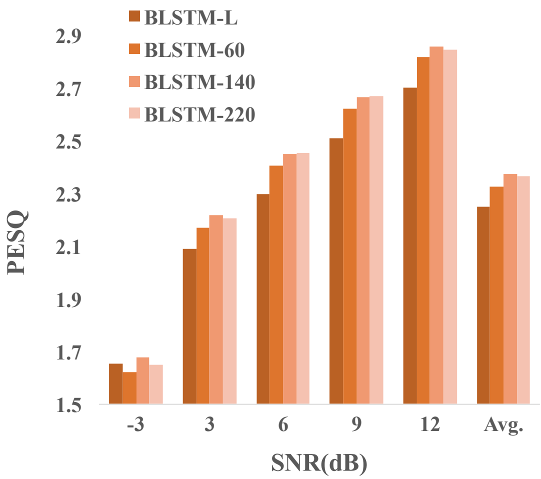


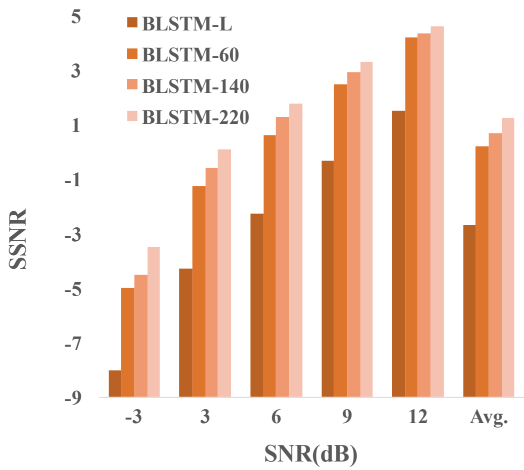
(1) DAT achieves good unsupervised adaptation performance (without paired noisy-clean adaptation data).
(2) More adaptation data gives higher scores.

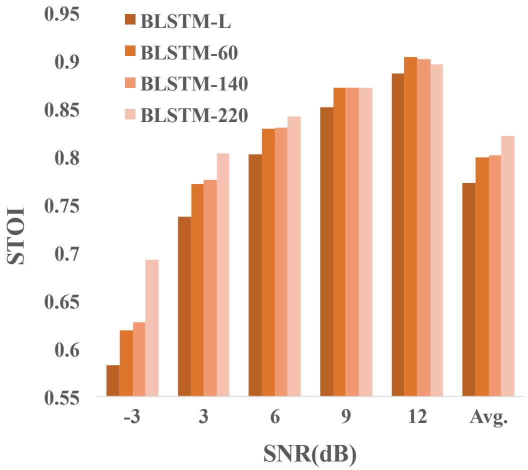
Environment Adaptability (NADAT)

- Adapting to new noise type (Cafeteria)

Different amount of target domain data for adaptation







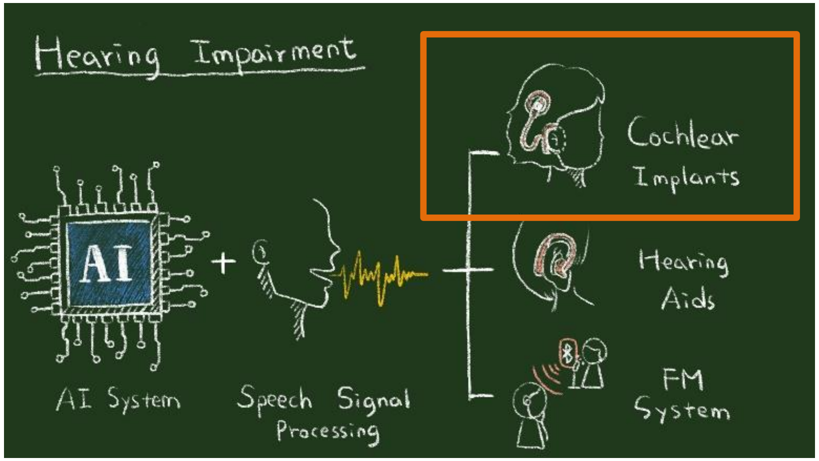
(1) DAT achieves good unsupervised adaptation performance (without paired noisy-clean adaptation data).
(2) More adaptation data gives higher scores.

Outline

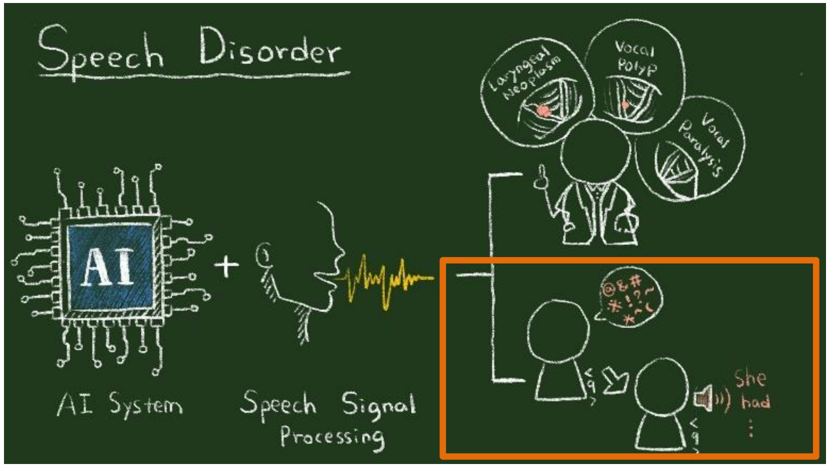
- Deep Learning based Speech Enhancement
  - System architecture
  - Five factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- **Assistive Voice Communication Technologies**

Assistive Voice Communication

- Assistive listening



- Assistive speaking

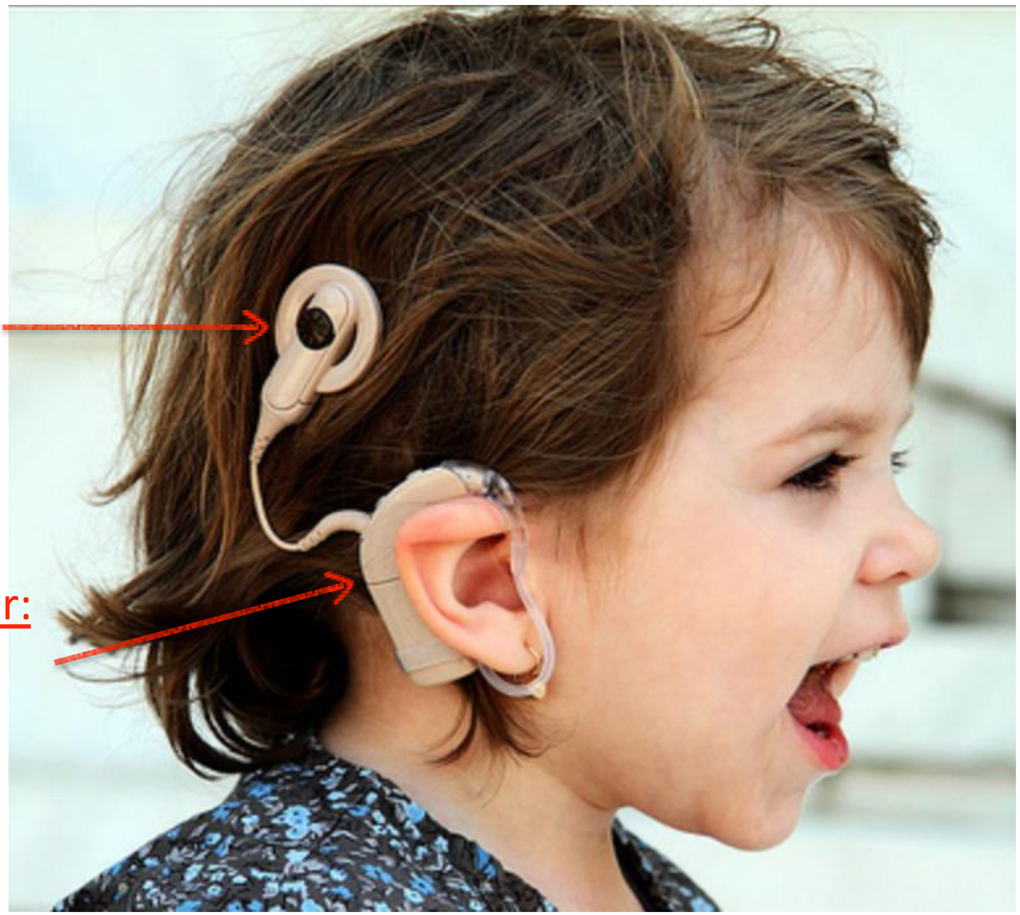


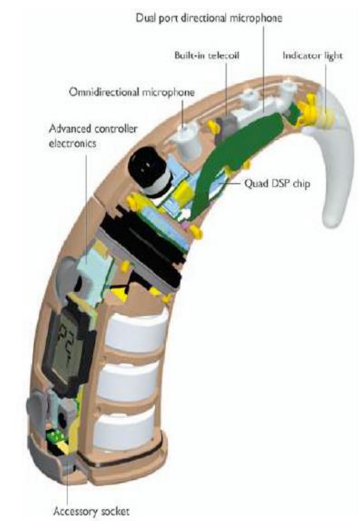
Cochlear Implant

Transmitter coil

<u>Speech processor:</u>

1. Microphone.
2. DSP chip.
3. Battery
4. Others…



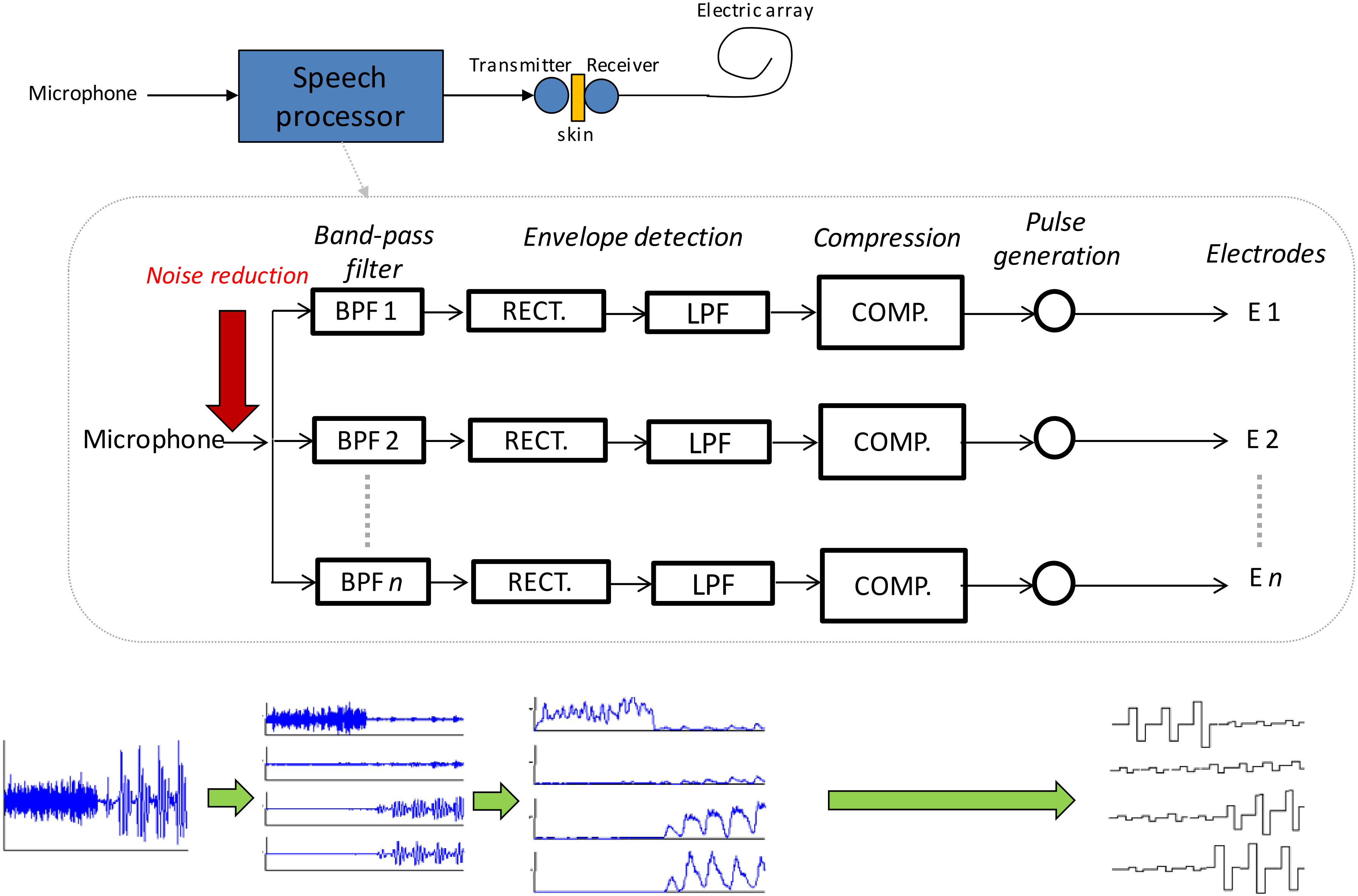


SE for Cochlear Implant

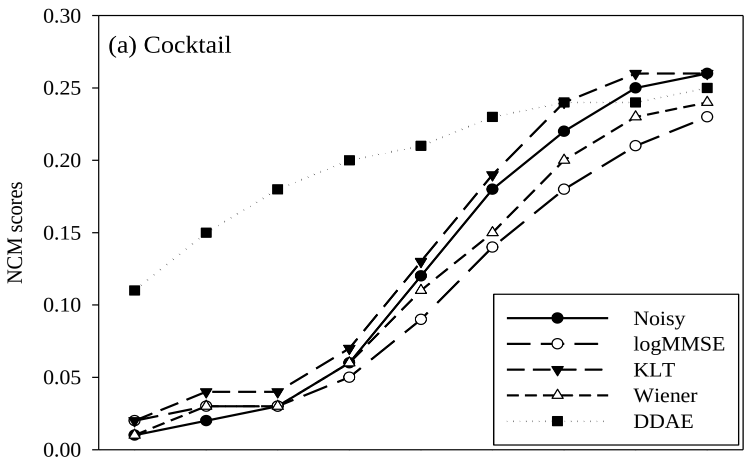
- The tremendous progress of CI technologies in the past three decades has enabled many CI users to enjoy **high level** of speech understanding **in quiet**.
- For most CI users, however, the performance of speech understanding **in noise** still remains **challenging**.
  - ➢ F. Chen, Y. Hu, and M. Yuan, "Evaluation of Noise Reduction Methods for Sentence Recognition by Mandarin-Speaking Cochlear Implant Listeners," Ear and hearing, vol. 36, no. 1, pp. 61-71, 2015.
- **Deep learning** based speech enhancement (SE) for CI.



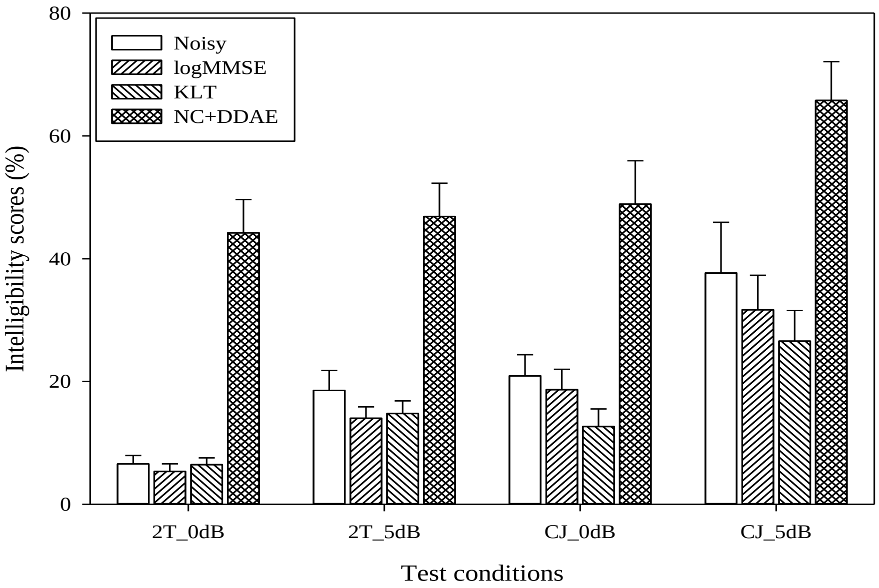
SE for Cochlear Implant



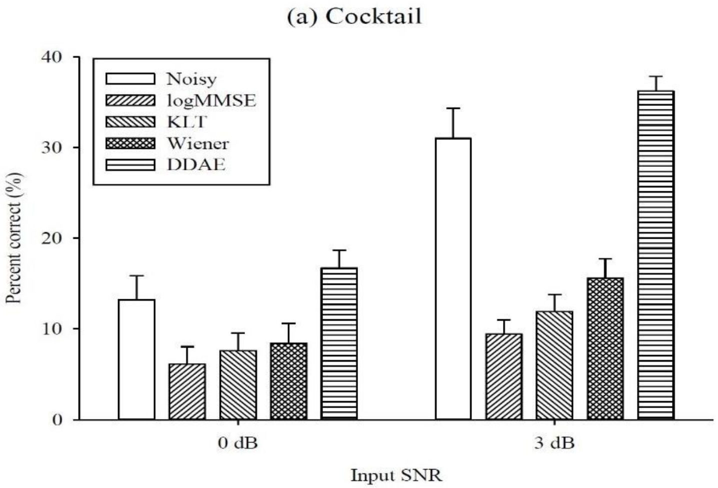
Testing Results



Objective evaluation (NCM)



Clinical trial: 9 CI subjects.



Vocoder results: 10 normal hearing subjects.

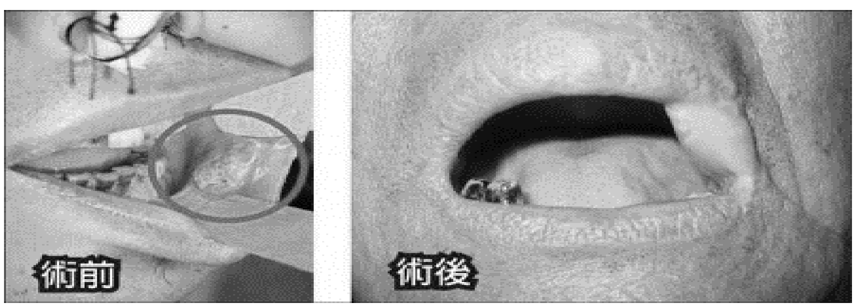
- Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation," IEEE Transactions on Biomedical Engineering.
- Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, K.-C. Chen, Y.-H. Chen, L.-C. Chen, P.-H. Li, and C.-H. Lee, "Deep Learning based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients," Ear and Hearing.
- **National Innovation Award 2018 and 2019.**

SE for Speaking Disorder

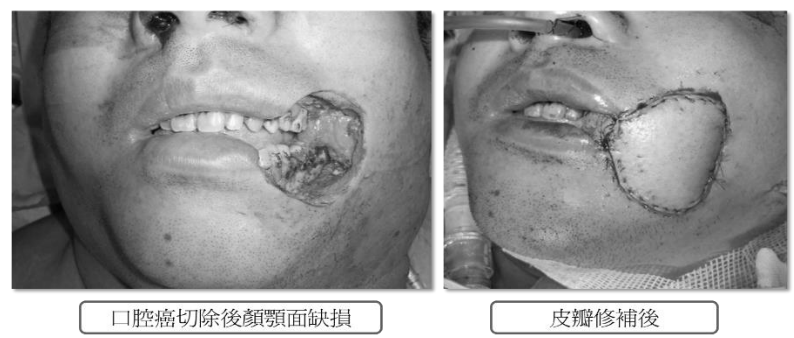
Previous applications: hearing aids; murmur to normal speech; bone-conductive microphone to air-conductive microphone.

Proposed: improving the speech intelligibility of surgical patients.

Target: oral cancer (top five cancer for male in Taiwan).



摘自自由時報



臺北榮民總醫院(口腔醫學部)

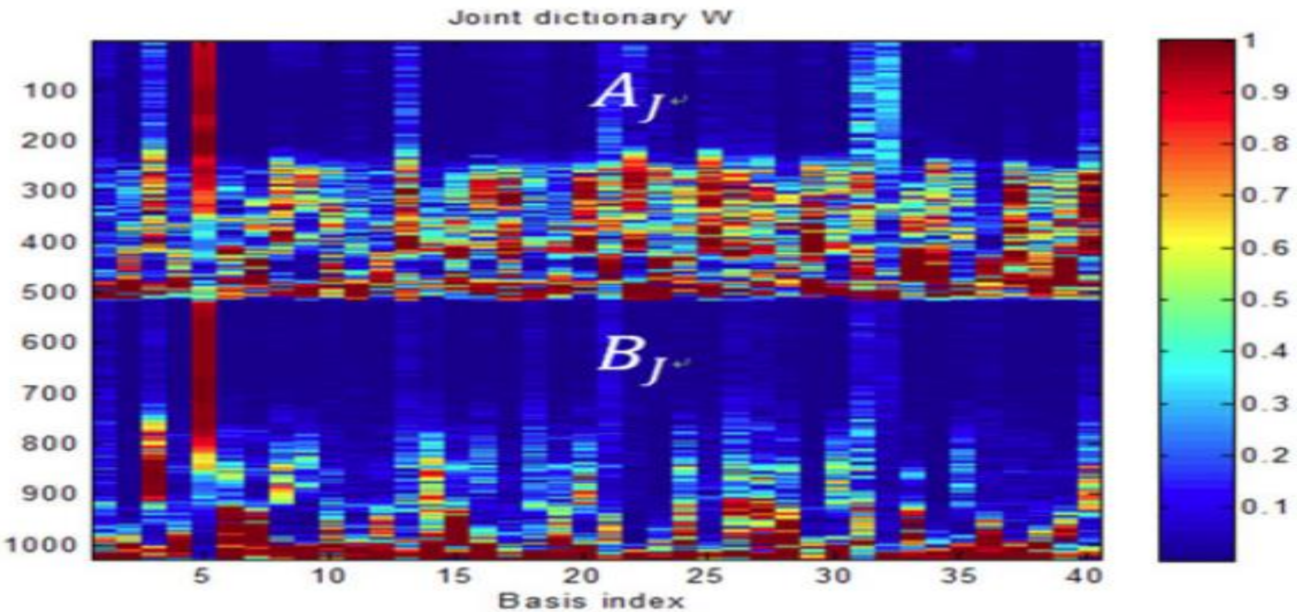
SE for Speaking Disorder

Related works:

Exemplar-based approach ---------------------------- (high computational cost)
Traditional NMF approach ---------------------------- (not well-aligned)

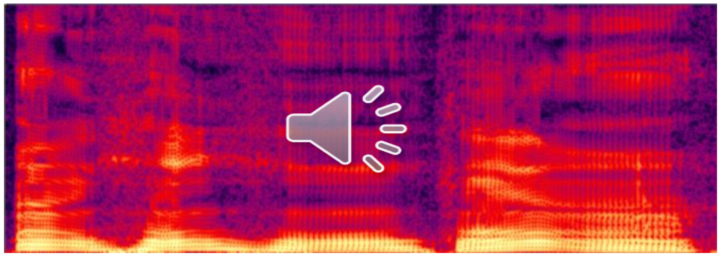
Proposed: joint training of source and target dictionaries:

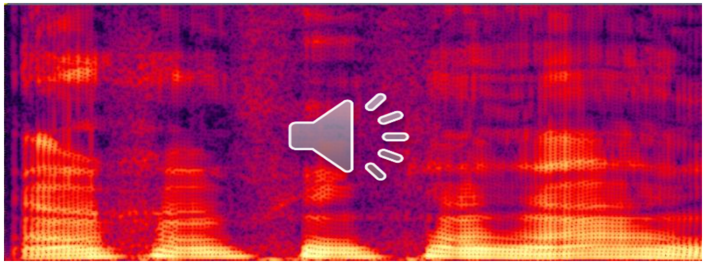
$$\mathbf{A}_j, \mathbf{B}_j = \arg\min_{\mathbf{A}_j,\mathbf{B}_j} d(\mathbf{X}, \mathbf{A}_j\mathbf{H}) + d(\mathbf{X}, \mathbf{B}_j\mathbf{H}) + \lambda\|\mathbf{H}\|_1$$



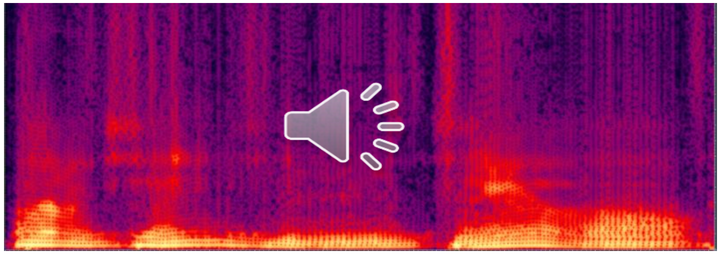
Testing Results

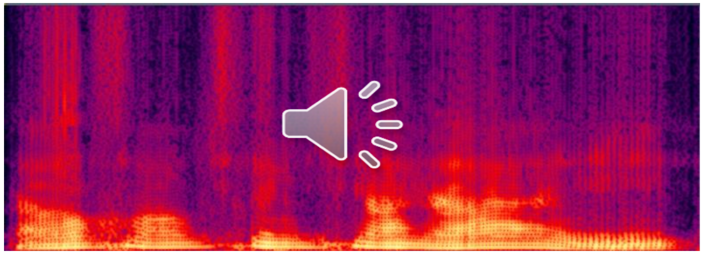
Original:





After Conversion:





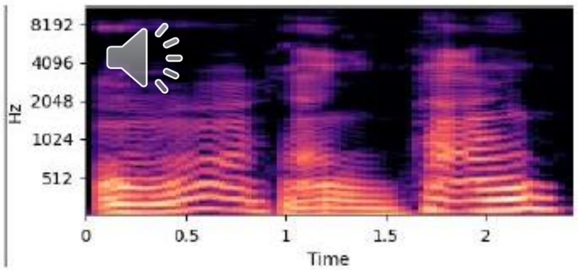
衛生紙給我

遙控器在哪裡

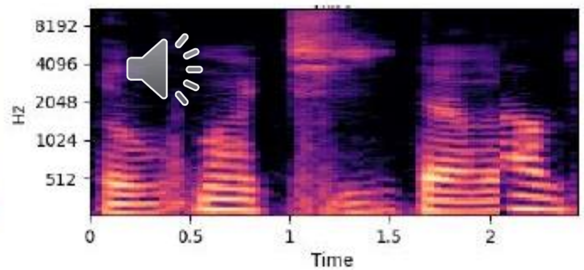
Speech samples were from [Fu et. al., TBME 2017]

SE(CGAN) for Speaking Disorder

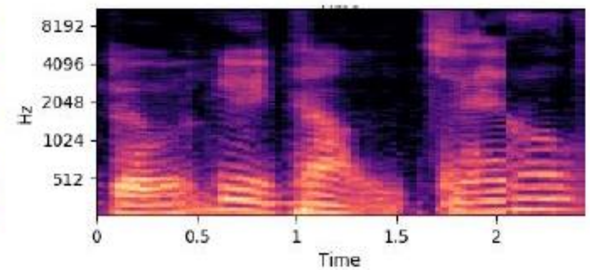
- Spectrogram analysis



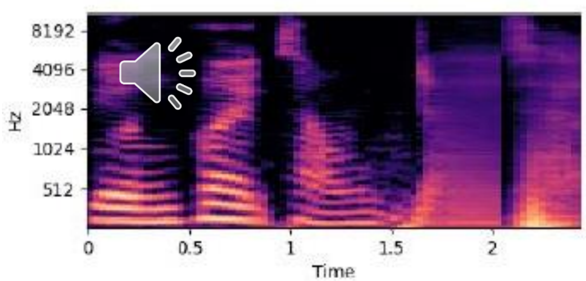
(a) Impaired speech



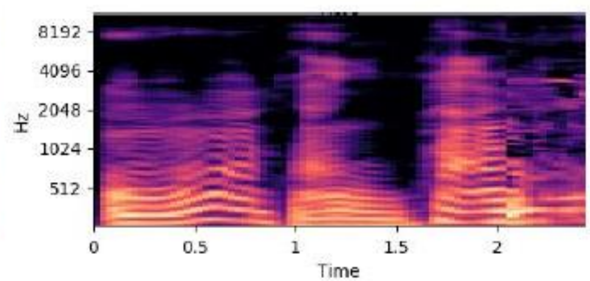
(b) Proposed model



(c) NoSD model



(d) CycleGAN transformed



(e) CycleGAN reconstructed

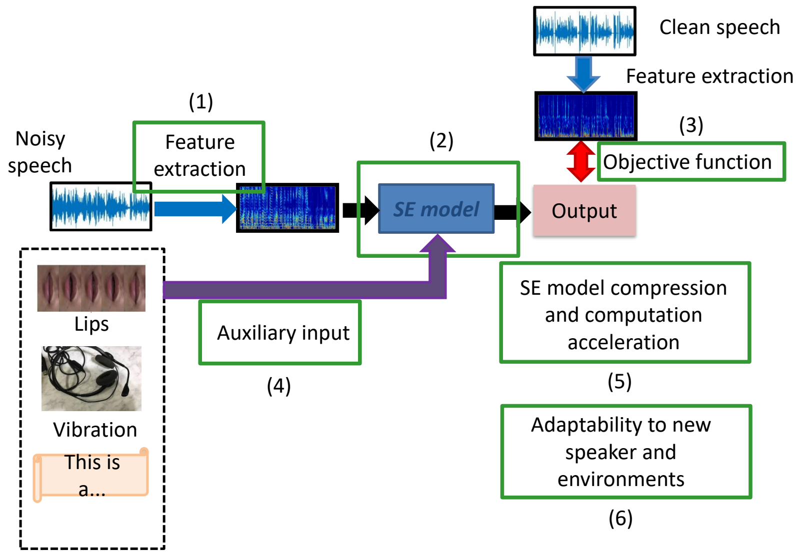
Speech samples were from [Chen et. al., Interspeech 2019]

Outline

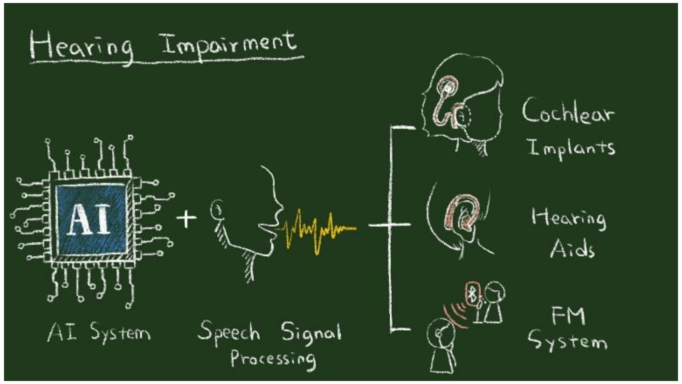
- Deep Learning based Speech Enhancement
  - ➢ System architecture
  - ➢ Six factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- **Summary**

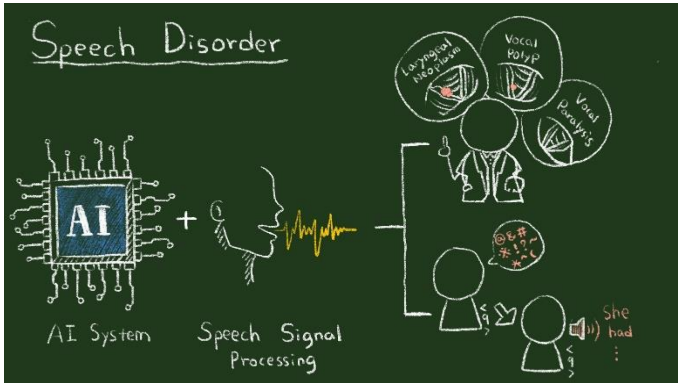
Summary

- Six Factors

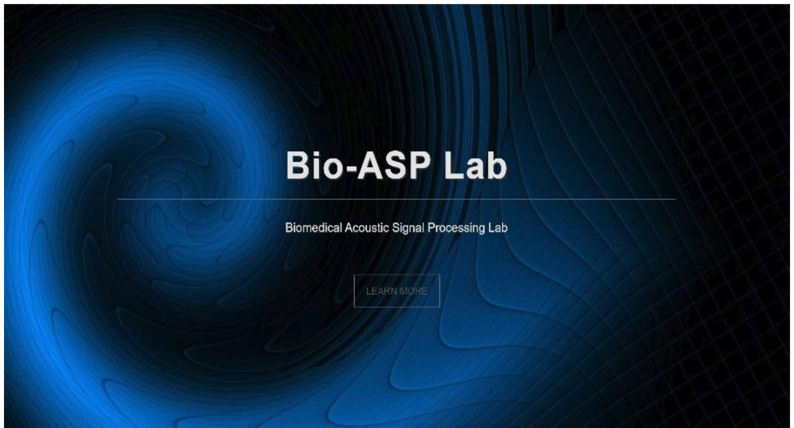


- Assistive Voice Communication Technologies





Bio-ASP Lab in CITI, Academia Sinica (中央研究院資訊科技創新研究中心)



Contact: yu.tsao@citi.sinica.edu.tw
More Information: http://bio-asplab.citi.sinica.edu.tw/
Publications: https://www.citi.sinica.edu.tw/pages/yu.tsao/publications_en.html

References

- X. Lu, Y. Tsao, S. Matsuda, H. Chiroi, Speech enhancement based on deep denoising autoencoder, Interspeech 2012.
- W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, Speech dereverberation based on integrated deep and ensemble learning algorithm, ICASSP, 2018.
- H.-P. Liu, Y. Tsao, and C.-S. Fuh, Bone conducted speech enhancement using deep denoising autoencoder, Speech Communication 2018.
- S.-W. Fu, T.-y. Hu, Y. Tsao, X. Lu, Complex spectrogram enhancement by convolutional neural network with multi-metrics learning, MLSP 2017.
- S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, Raw waveform-based speech enhancement by fully convolutional networks, APSIPA 2017.
- S.-W. Fu, Y. Tsao, X.-G. Lu, and Hisashi Kawai, End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks, IEEE/ACM TASLP, 2018.
- D. Wang and J. Chen, Supervised speech separation based on deep learning: An overview," IEEE/ACM TASLP 2018.
- Y.-X. Wang and D.-L. Wang, Cocktail party processing via structured prediction, NIPS 2012.
- Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, An experimental study on speech enhancement based on deep neural networks, IEEE SPL, 2014.
- Y. Xu, J. Du, L.-R. Dai, and Chin-Hui Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM TASLP, 2015.
- Z. Chen, S. Watanabe, H. Erdogan, J. R. Hershey, Integration of speech enhancement and recognition using long-short term memory recurrent neural network, Interspeech 2015.
- F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, Speech enhancement with LSTM recurrent neural networks and Its application to noise-robust ASR, LVA/ICA, 2015.
- S.-W. Fu, Y. Tsao, and X.-G. Lu, SNR-aware convolutional neural network modeling for speech enhancement, Interspeech, 2016.
- T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao and W.-H. Liao, Experimental study on extreme learning machine applications for speech enhancement, IEEE Access 2017.
- M. Tu and X. Zhang, Speech enhancement based on deep neural networks with skip connections, ICASSP 2017.
- J. F. Santos and T. H. Falk, Speech dereverberation with contextaware recurrent neural networks, IEEE/ACM TASLP 2018.
- T. Gao, J. Du, L. R. Dai, C.-H. Lee, Densely connected progressive learning for LSTM-based speech enhancement, ICASSP 2018.
- Xiang Hao, Changhao Shan, Yong Xu, Sining Sun, and Lei Xie. An attention-based neural network approach for single channel speech enhancement." ICASSP 2019.
- P. Santiago, B. Antonio, and S. Joan, SEGAN: Speech enhancement generative adversarial network, Interspeech, 2017.

References

- Y.-S. Lee, C.-Y. Wang, S.-F. Wang, J.-C. Wang, and C.-H. Wu, Fully complex deep neural network for phase-incorporating monaural source separation, ICASSP 2017.
- Y. Koizumi, K. Niwa, Y. Hioka, K. Koabayashi, and Y. Haneda, DNN-based source enhancement to increase objective sound quality assessment score, IEEE/ACM TASLP 2018.
- Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements, ICASSP 2017.
- H. Zhang, X. Zhang, and G. Gao, Training supervised speech separation system to improve STOI and PESQ directly, ICASSP 2018.
- J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, A deep learning loss function based on the perceptual evaluation of the speech quality, IEEE SPL 2018.
- S.-W. Fu, C.-F. Liao, Y. Tsao, Learning with learned loss function: speech enhancement with Quality-Net to improve perceptual evaluation of speech quality," to appear in IEEE SPL.
- D. Michelsanti, and Z.-H. Tan, Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification, Interspeech, 2017.
- C. Donahue, B. Li, and P. Rohit, Exploring speech enhancement with generative adversarial networks for robust speech recognition, ICASSP, 2018.
- A. Pandey and D. Wang, On adversarial training and loss functions for speech enhancement, ICASSP 2018.
- M. H. Soni, Neil Shah, and H. A. Patil, Time-frequency masking-based speech enhancement using generative adversarial network, ICASSP 2018.
- S.-W. Fu, C.-F. Liao, Y. Tsao, S.-D. Lin, MetricGAN: generative adversarial networks based black-box metric scores optimization for speech enhancement," ICML 2018.
- Y.-L. Shen, C.-Y. Huang, S.-S. Wang, Y. Tsao, H.-M. Wang, and T.-S. Chi, Reinforcement learning based speech enhancement for robust speech recognition, ICASSP 2019.
- J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, Audio-visual speech enhancement using multimodal deep convolutional neural networks, IEEE TETCI 2018.
- J.-Y. Wu, C. Yu, S.-W. Fu, C.-T. Liu, S.-Y. Chien, Y. Tsao, Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques, to appear in IEEE SPL.

References

- F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement, Interspeech 2019.
- C.-F. Liao, Y. Tsao, H.-y. Lee and H.-M. Wang, Noise adaptive speech enhancement using domain adversarial training, Interspeech 2019.
- Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation, IEEE TBME 2017.
- Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, K.-C. Chen, Y.-H. Chen, L.-C. Chen, P.-H. Li, and C.-H. Lee, Deep learning based noise reduction approach to improve speech intelligibility for cochlear implant recipients, Ear and Hearing 2018.
- S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery, IEEE TBME 2017.
- L.-W. Chen, H.-Y. Lee, and Y. Tsao, Generative adversarial networks for unpaired voice transformation on impaired speech, Interspeech 2019.