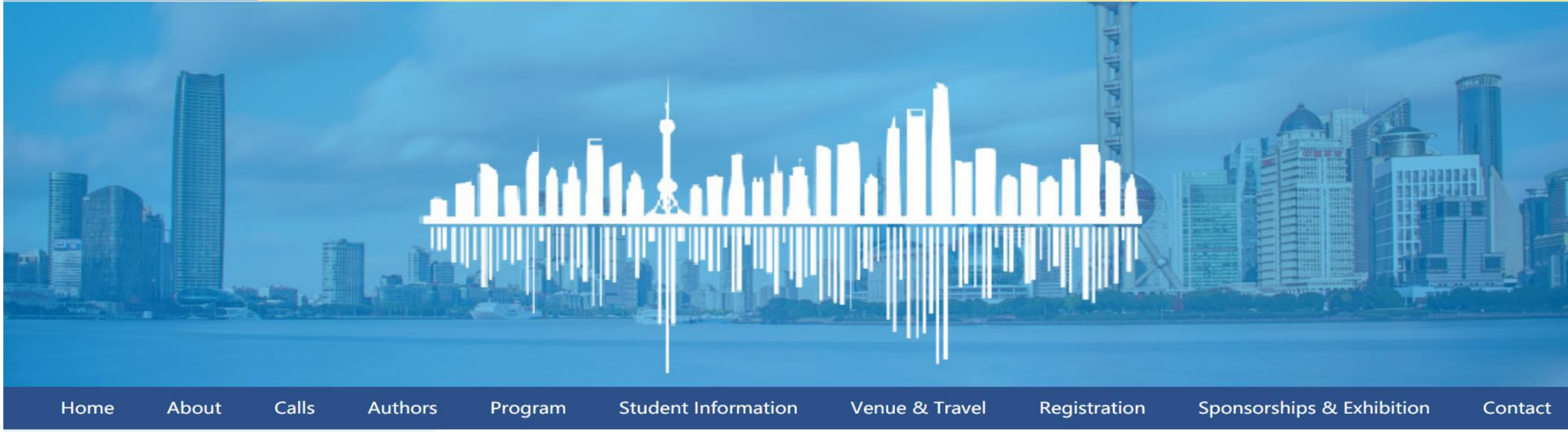


Intelligibility Evaluation and Speech Enhancement based on Deep Learning (Part II)

Yu Tsao

Research Center for Information Technology Innovation

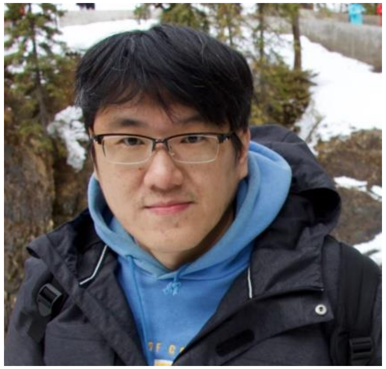
Academia Sinica

yu.tsao@citi.sinica.edu.tw

中央研究院

ACADEMIA SINICA

Dr. Yu Tsao (曹昱), *Research Fellow, Deputy Director*



– 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

- Research Fellow (Professor) and Deputy Director Research Center for Information Technology Innovation (2020/9-present)
- 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, and Speech Signal Processing

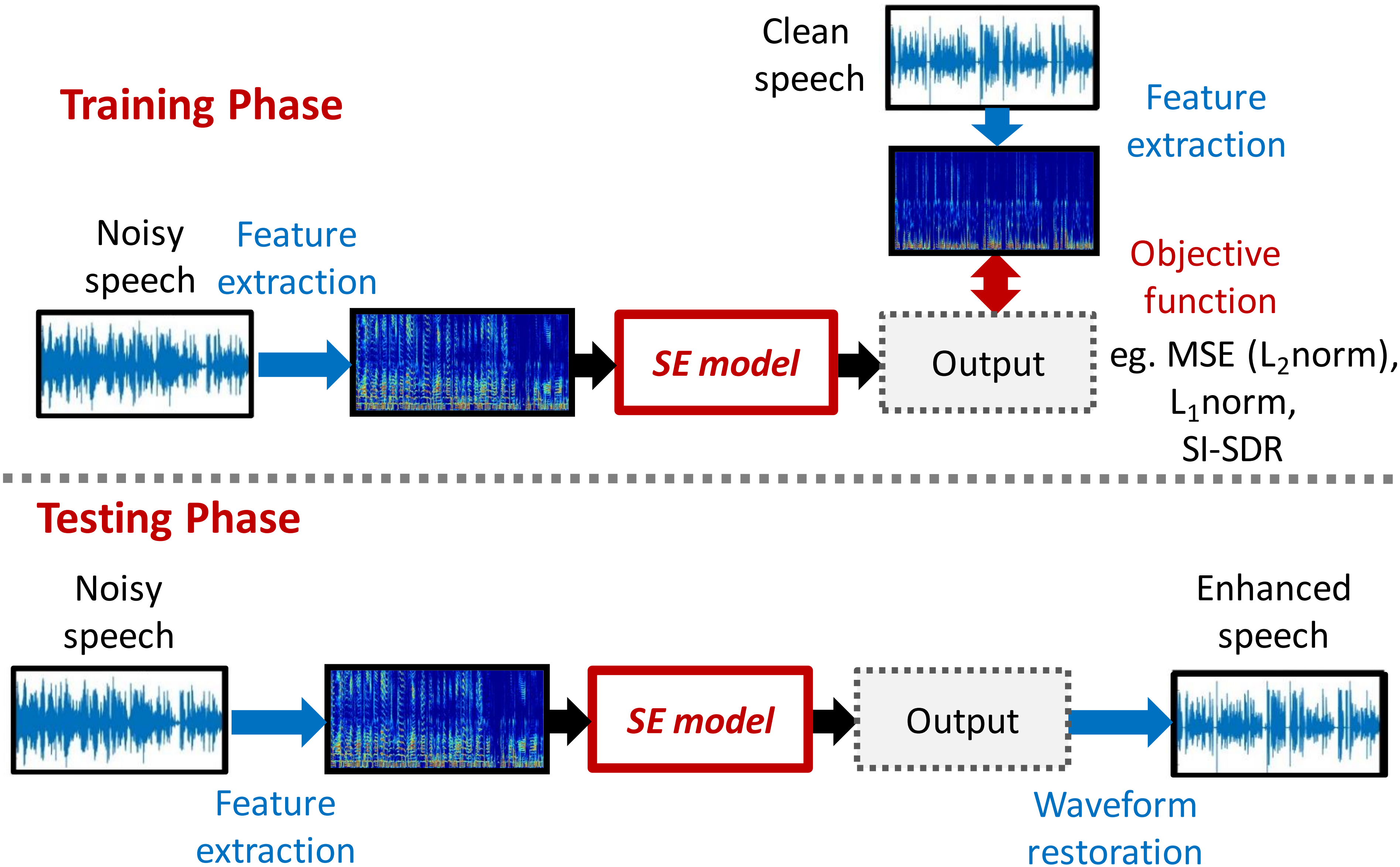
Outline

- 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

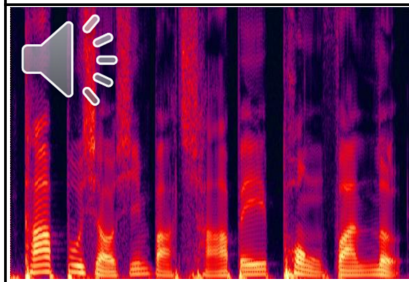
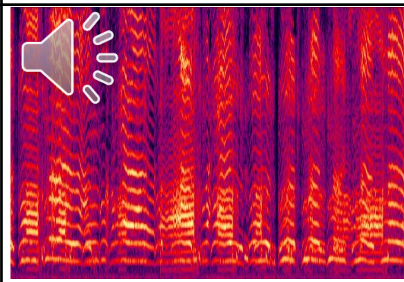
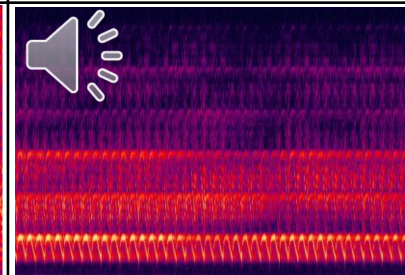
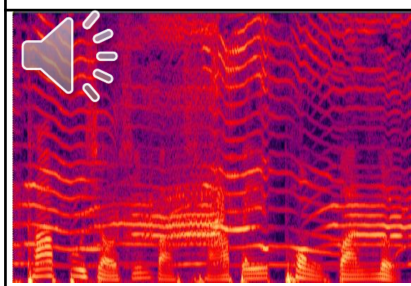
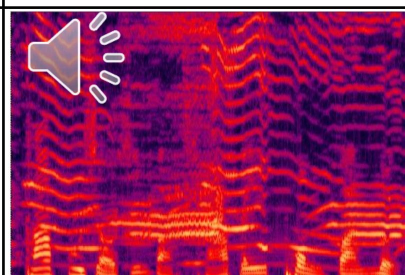
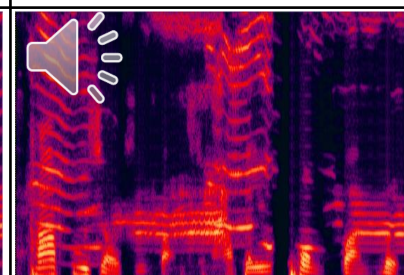
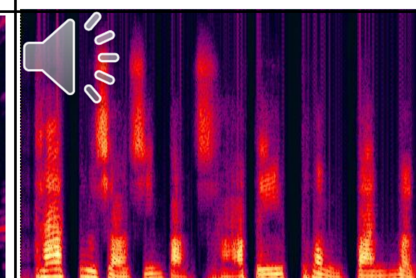
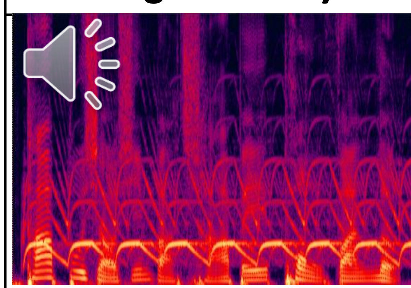
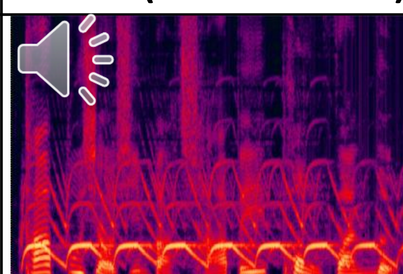
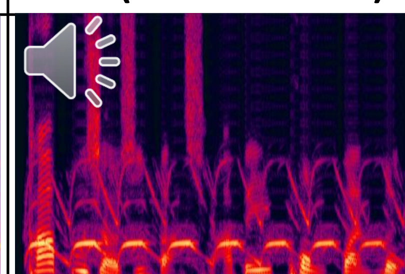
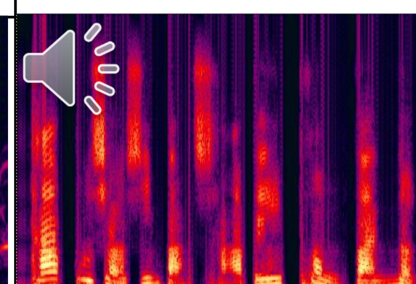
- Deep Learning based Speech Enhancement
  - **System architecture**
  - Six factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- Summary

Deep Learning Based SE System



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

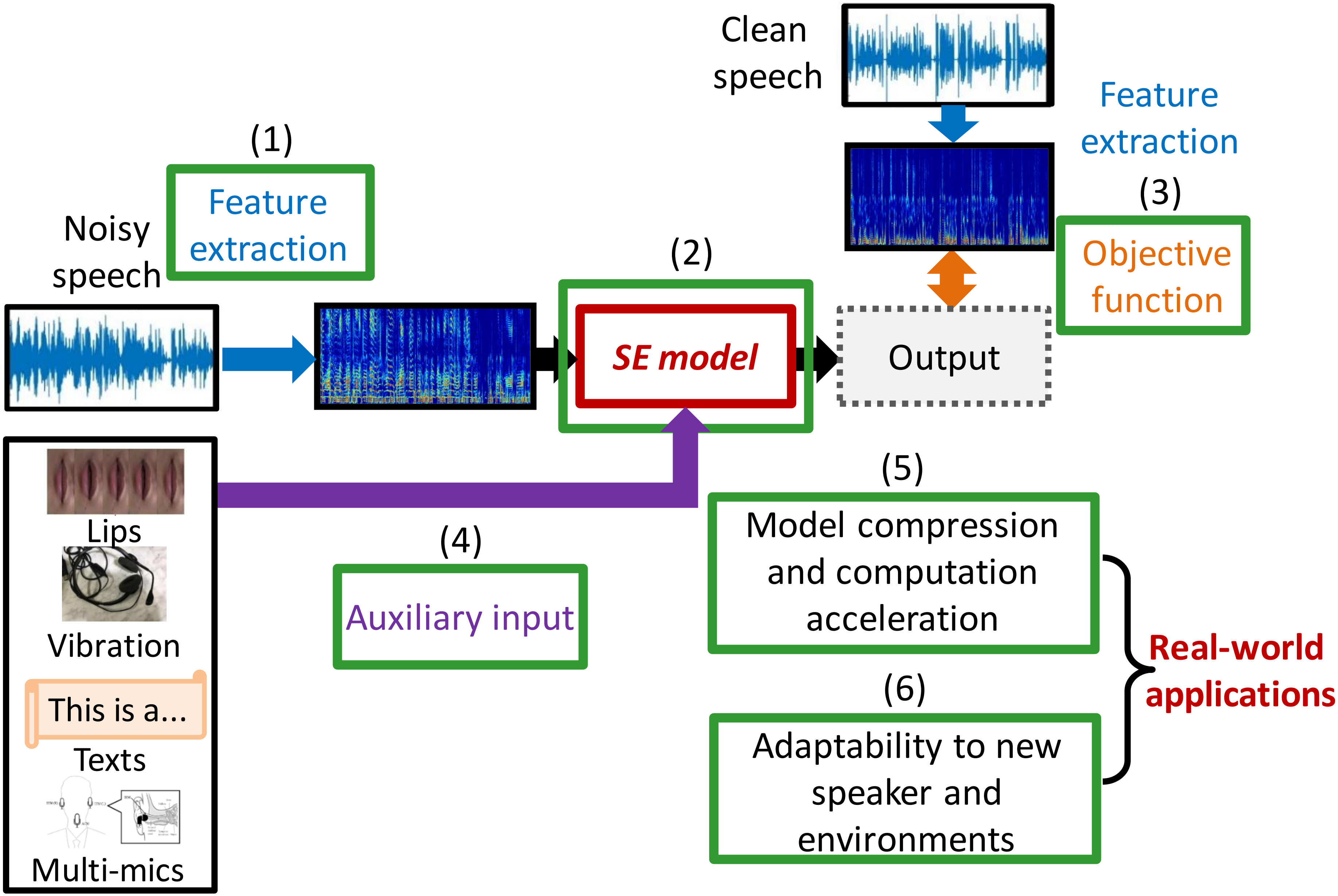
DL-based SE for Noisy Speech

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

DL-based SE for Bone-conducted Speech

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.

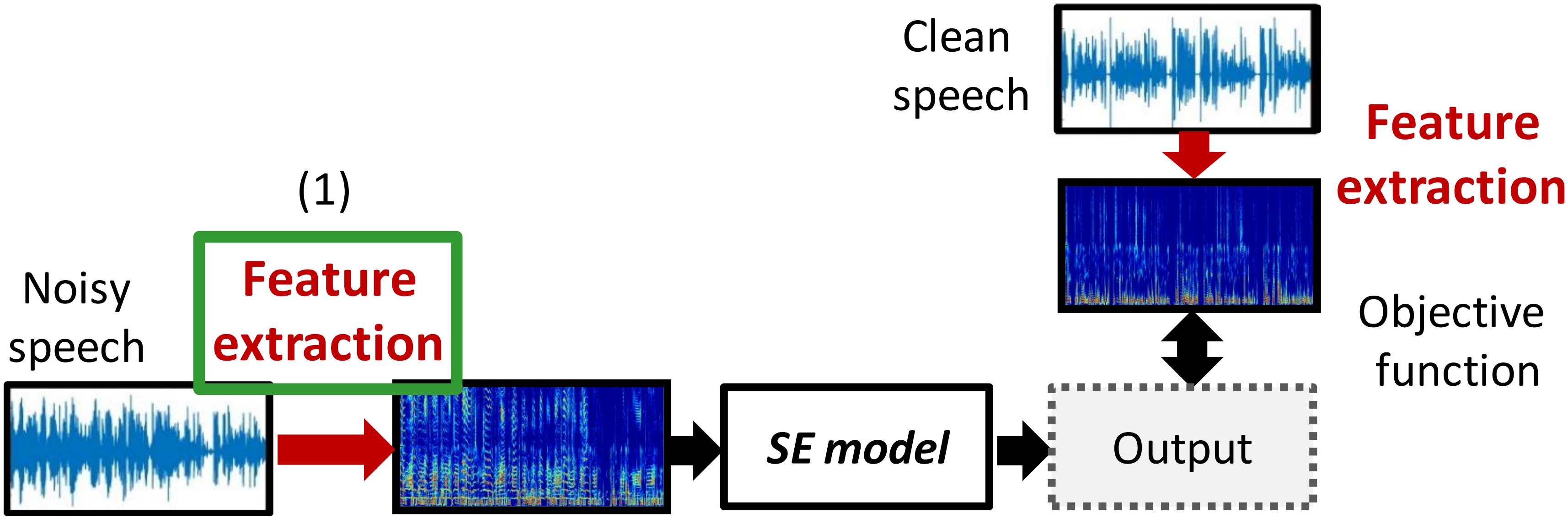
Outline

- 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

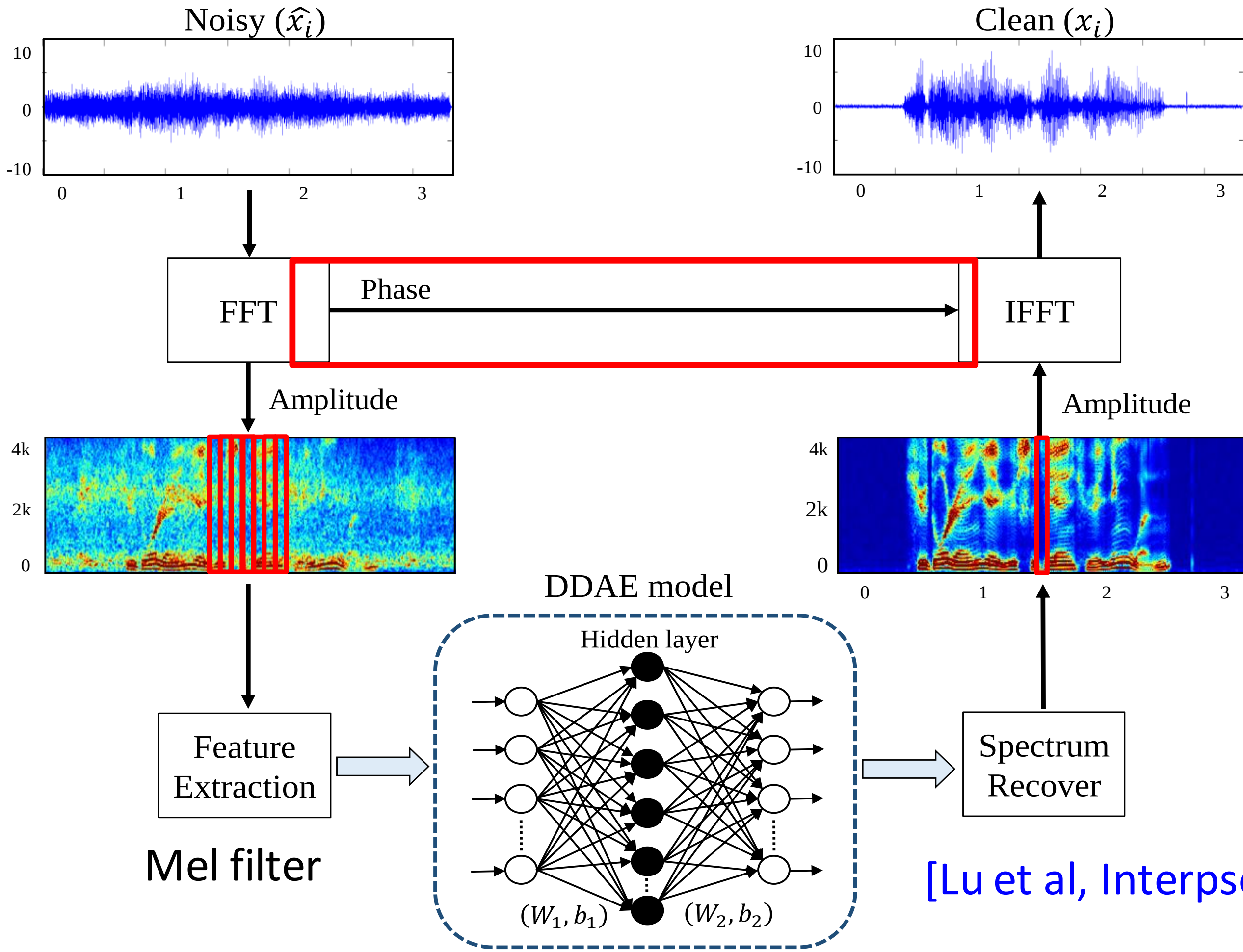
- Deep Learning based Speech Enhancement
  - System architecture
  - **Six factors need to consider**
    - ✓ **Feature types**

Input Feature Types



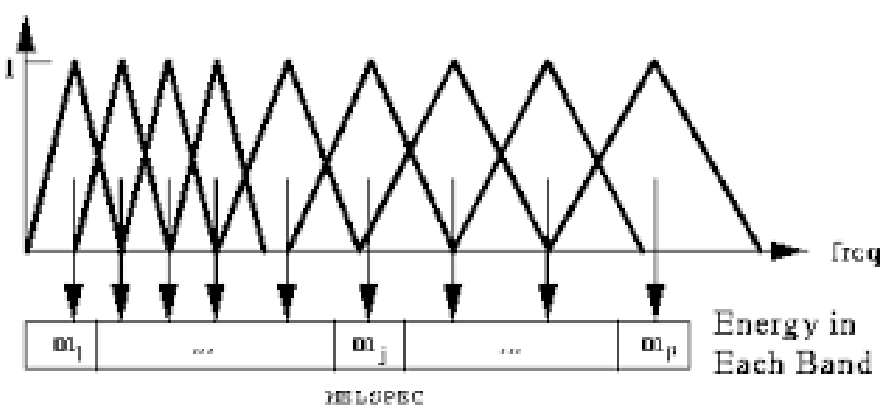
- Mel Log-power spectrum [Lu et al., Interspeech 2013, Meng et al., Interpseech 2018],
- Log-power spectrum [Xu et al., TASLP 2015, Fu et al., Interspeech 2016],
- Log1p [Chuang et al Interspeech 2020, and Lu et al., Interspeech 2020],
- Power spectrum [Fu et al., Interspeech 2016],
- Complex spectrum [Fu et al., MLSP 2017, Hu et al., arXiv 2020, Wang et al., TASLP 2020],
- Frame-wise waveform [Fu et al, APSIPA 2017],
- Utterance-wise waveform

[Fu et al, TASLP 2018, Kolbæk et al., TASLP 2020, Luo et al., TASLP 2019, Pandey et al., 2019, Luo et al., ICASSP 2020]........

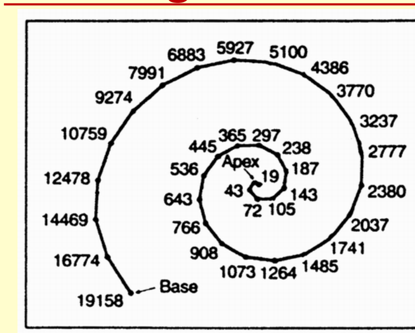


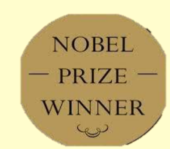
Mel filter

[Lu et al, Interpseech 2013]



Traveling wave theory





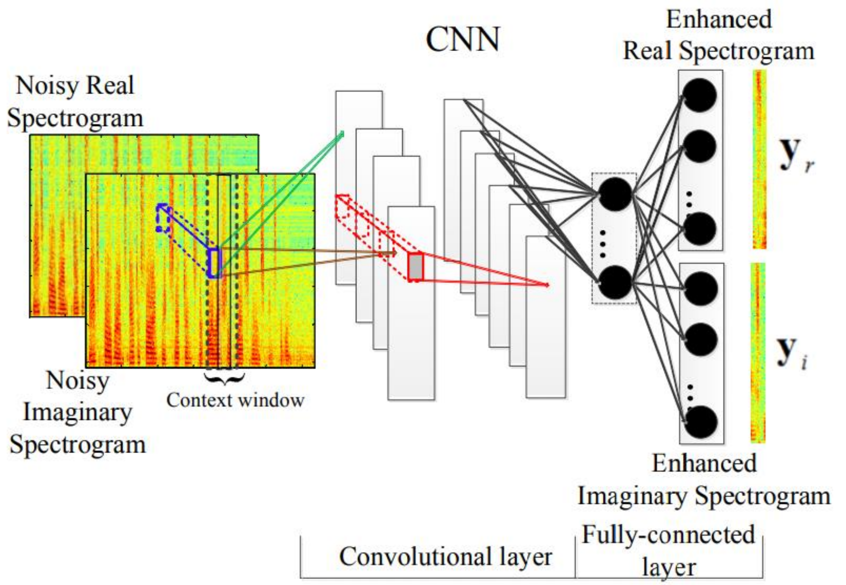
1961

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]

Noisy Real and imaginary (RI) spectrograms



Enhanced RI spectrograms

RI spectrograms are processed by a CNN model.
RI 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.

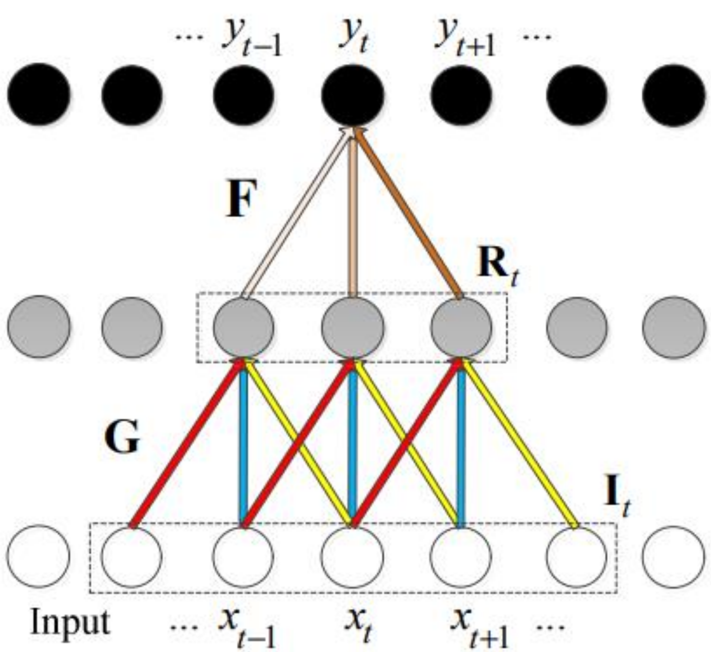
	DNN-baseline (LPS)				RI-DNN ($\alpha=1, \beta=0$)				RI-CNN ($\alpha=1, \beta=0$)			
SNR (dB)	**LSD**	**SSNR**	**STOI**	**PESQ**	**LSD**	**SSNR**	**STOI**	**PESQ**	**LSD**	**SSNR**	**STOI**	**PESQ**
12	**3.115**	-0.229	0.814	2.334	3.761	2.149	0.851	2.643	3.604	**3.042**	**0.886**	**2.741**
6	**3.404**	-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.

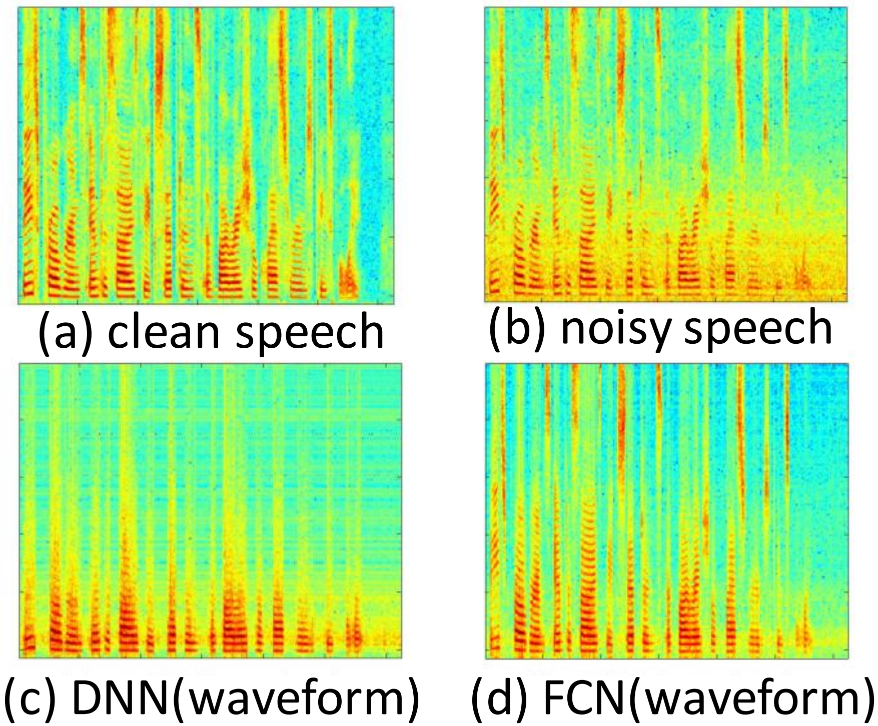
Input Feature Types

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

Local connection in FCN



Spectrograms of a TIMIT utterance:



(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:

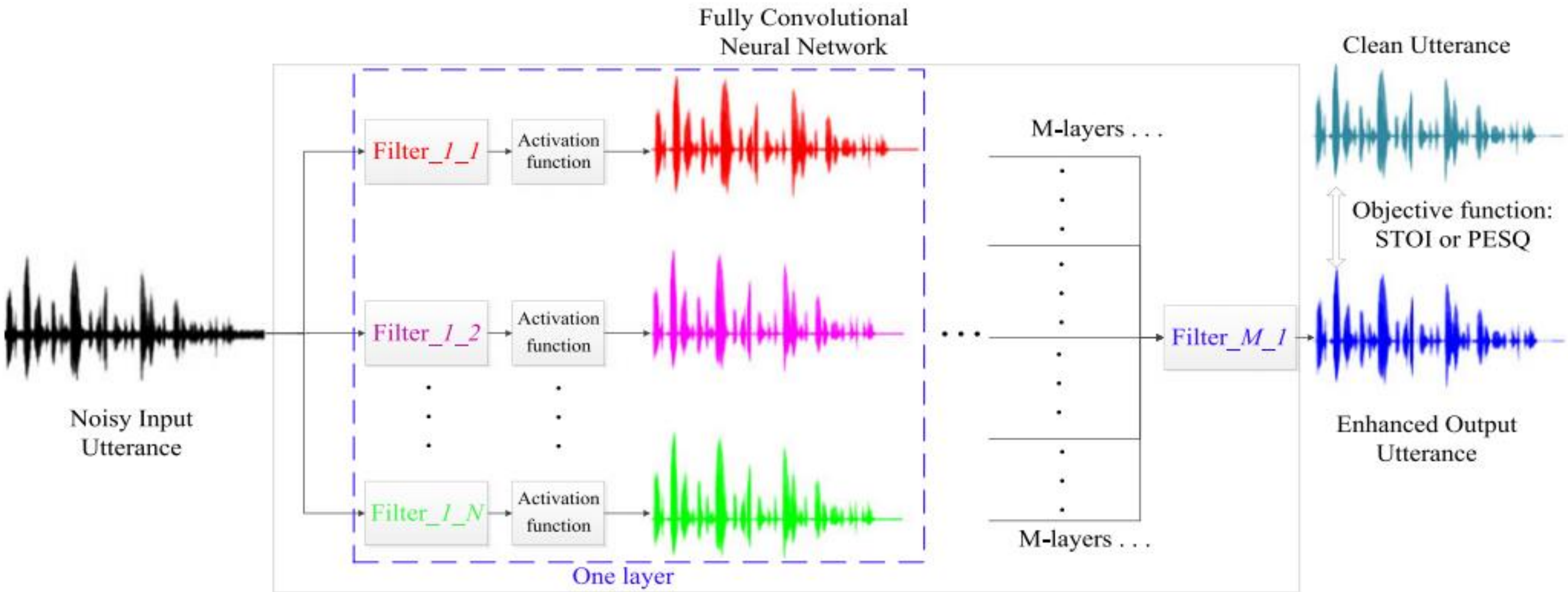
Comparison of different models and input features in terms of STOI, and PESQ.

	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.
(3) LPS with DNN underperforms the waveform-based systems.

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

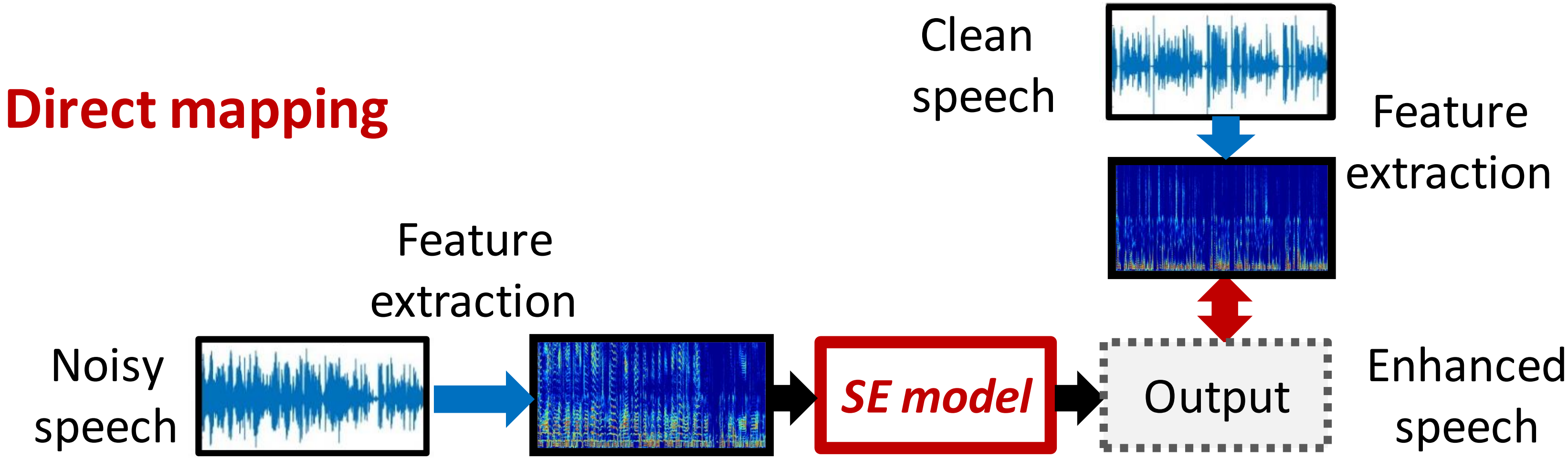
Comparison of different models and input features in terms of STOI and PESQ.

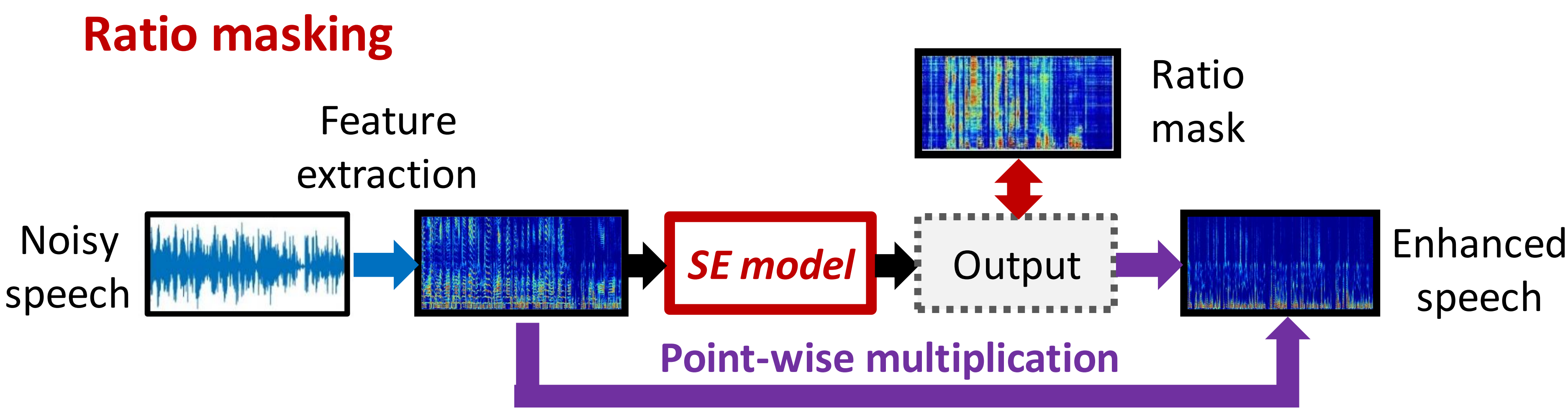
	Frame-based		Utterance-based			
	FCN (obj=MSE)		**FCN** (obj=MSE)		**FCN** (obj= STOI)	
SNR (dB)	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**
12	0.874	2.718	0.909	**2.909**	**0.931**	2.587
6	0.833	2.346	0.864	**2.481**	**0.888**	2.205
0	0.758	1.995	0.780	**2.078**	**0.814**	1.877
-6	0.639	1.719	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]

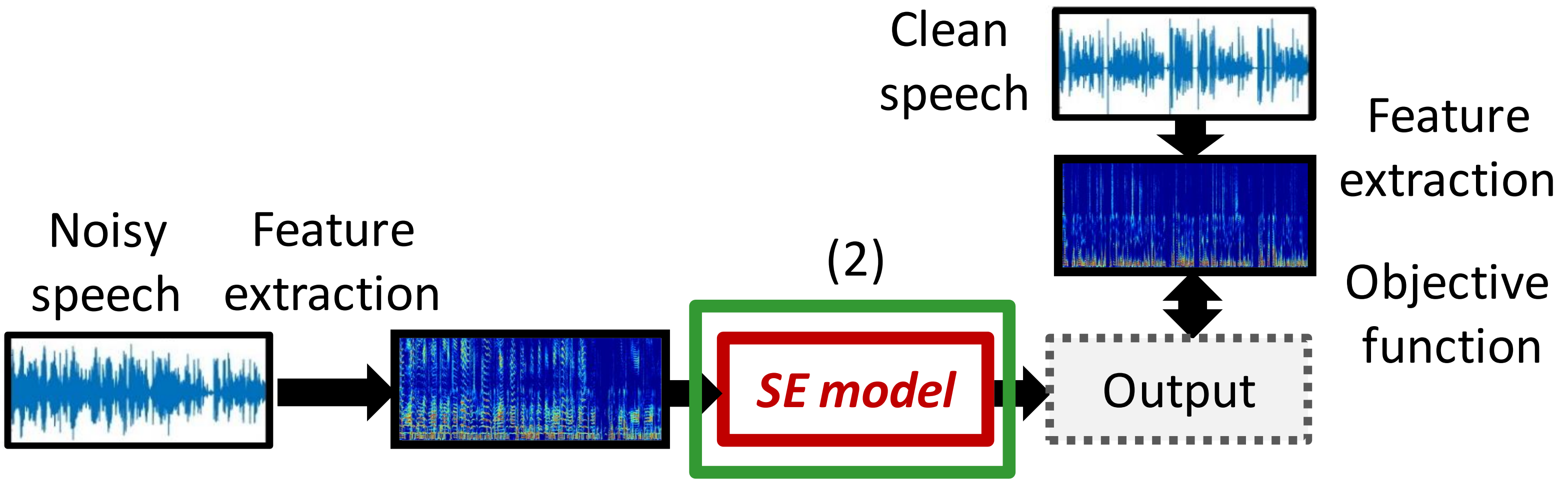




Outline

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

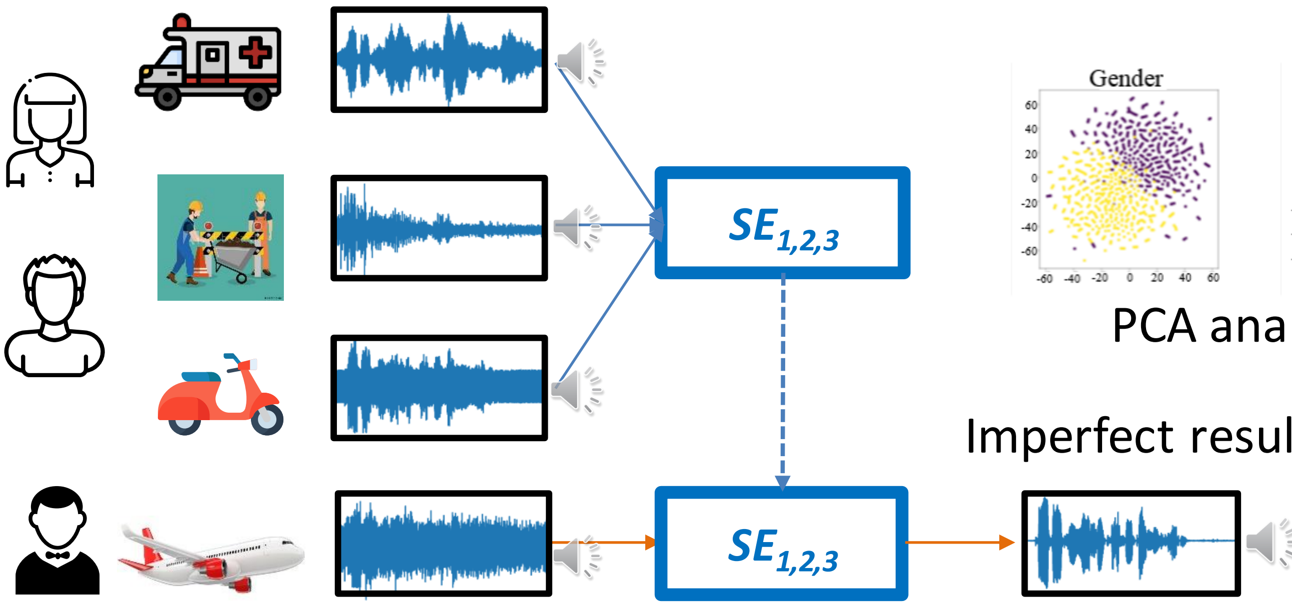
Model Types

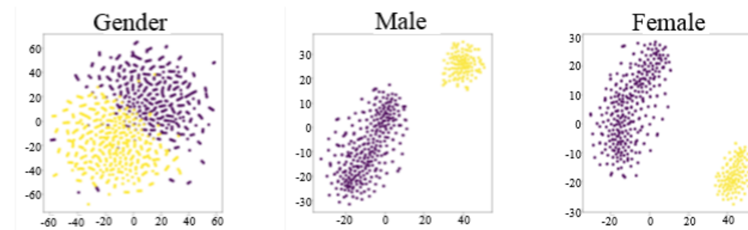


Model types:
DNN [Wang et al. NIPS 2012; Xu et al., SPL 2014], DDAE [Lu et al., Interspeech 2013], RNN (LSTM) [Chen et al., Interspeech 2015; Weninger et al., LVA/ICA 2015], CNN [Fu et al., Interspeech 2016], CRNN [Zhao et al., ICASSP 2018], FCN [Fu et al, TASLP 2018], HELM [Hussain et al., IEEE Access 2017], Vector2Vector [Qi et al., TASLP 2020], Tensor2Vector [Qi et al., ICASSP 2020], Teacher-Student [Tu et al., TASLP 2019].
Advanced architecture:
Skip connection [Tu and Zhang ICASSP 2017], Highway [Santos and Falk, NIPS workshop 2018], Densely connectied [Zhen et al., ICASSP 2019], Attention mechanism [Hao et al., ICASSP 2019], U-Net architecture [Pascual et al., Interspeech 2017], Complex parameters [Y.-S. Lee et al., ICASSP 2017]. Transformer [Kim et al., ICASSP 2020, Fu et al., APSIPA 2020], Ensemble learning [Le Roux, WASPAA 2013, ICASSP 2017, Chazan et al., WASPAA 2017, Zhang et al., TASLP2016, Yu et al., TASLP 2020].

Model Types (Ensemble Learning)

- DAE Multibranched Encoder (DAEME) [Yu et. al., TASLP, 2020]



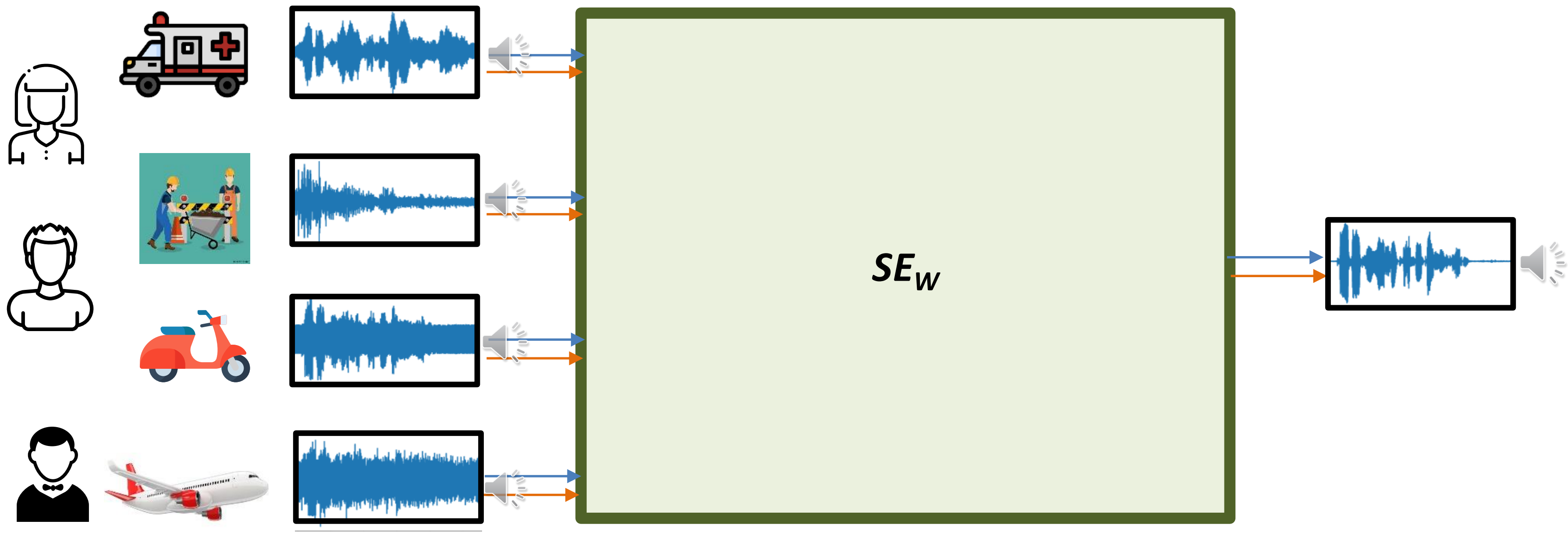


PCA analysis on noisy data

Based on the study in [Kolbæk et al., TASLP 2017], three major factors that affect the SE performance notably:
(1) Speaker (2) Noise type; (3) Signal-to-noise ratio (SNR).

Model Types (Ensemble Learning)

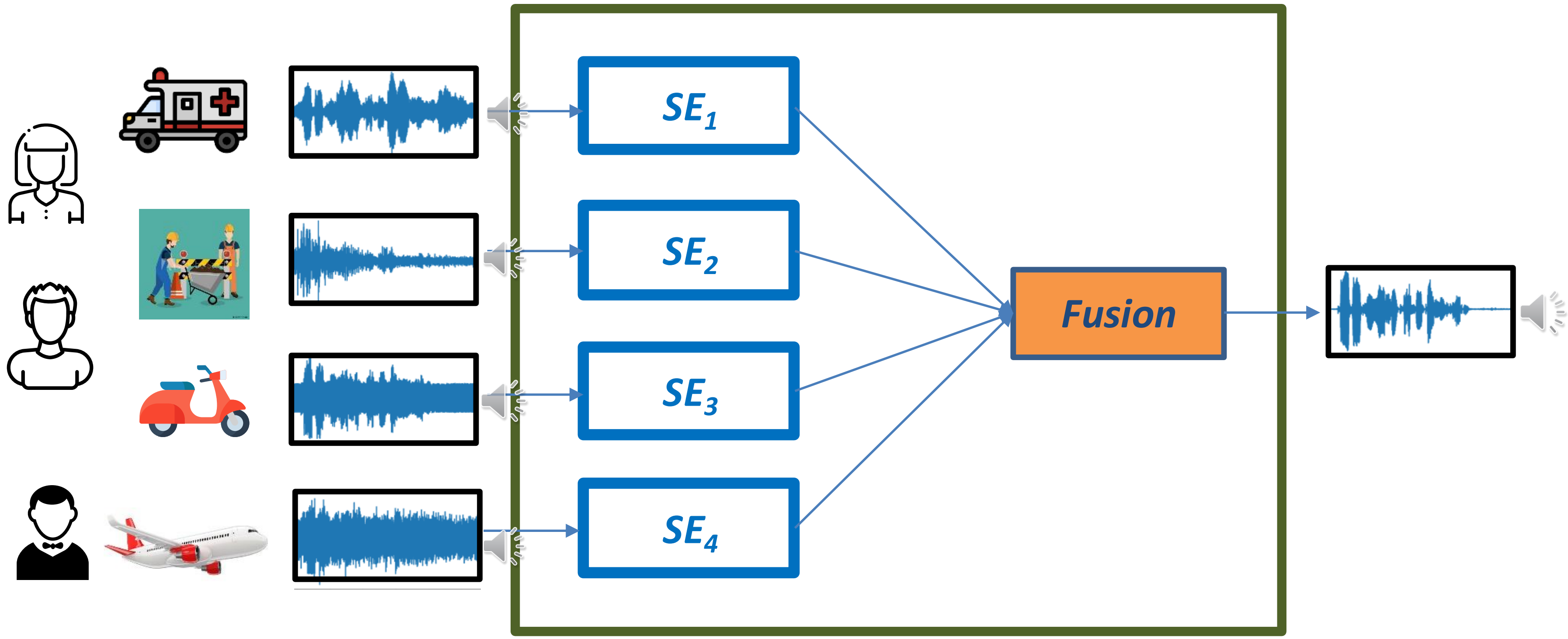
- DAEME [Yu et. al., TASLP, 2020]



(1) Training a gigantic SE model can be a potential solution.
(2) Such approach may not be suitable/feasible for the conditions where computation resources and data are limited.

Model Types (Ensemble Learning)

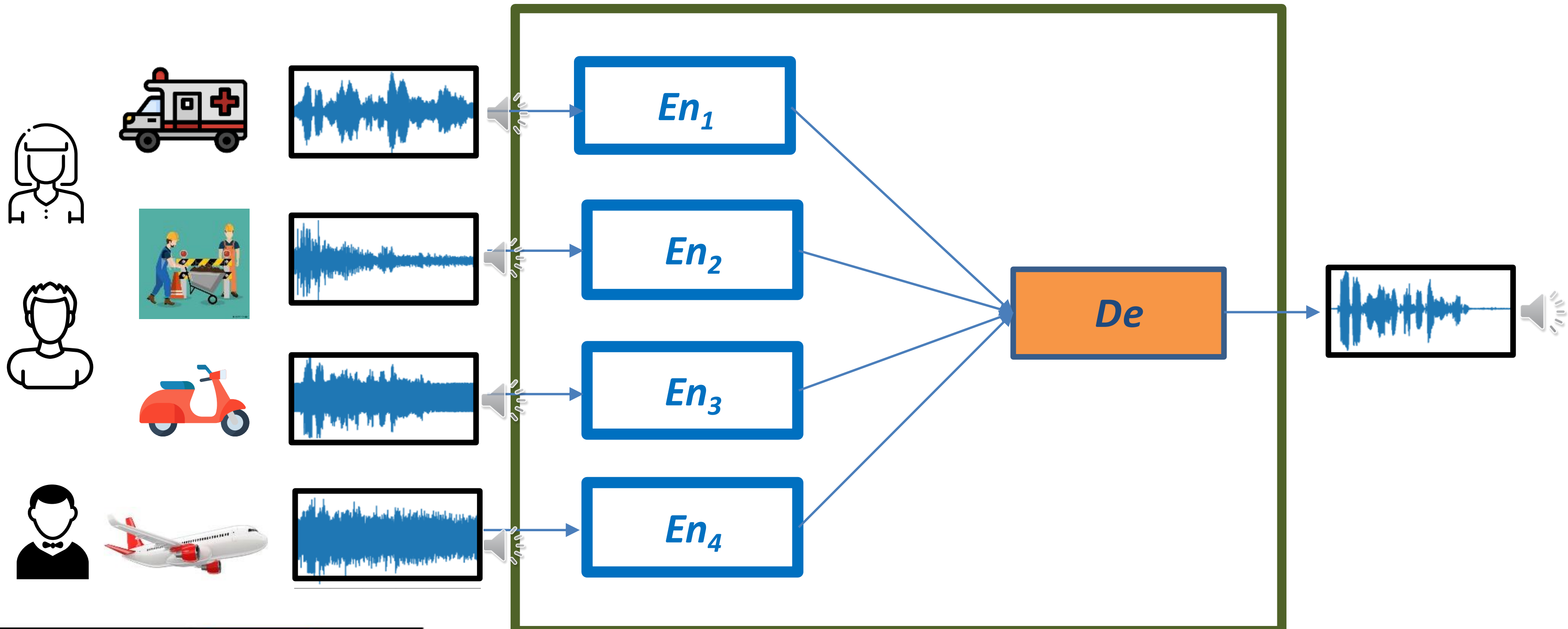
- DAEME [Yu et. al., TASLP, 2020]

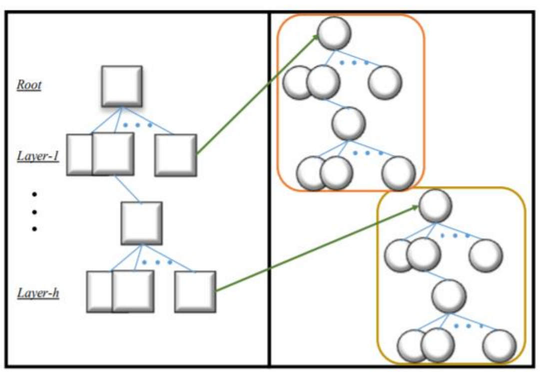


(1) The proposed DAEME is based on the ensemble learning criterion.
(2) When training ensemble models, we intend to implement a "conditional overfitting" strategy, which aims to train each component model to overfit to (or perfectly match) its training data.

Model Types (Ensemble Learning)

- DAEME [Yu et. al., TASLP, 2020]

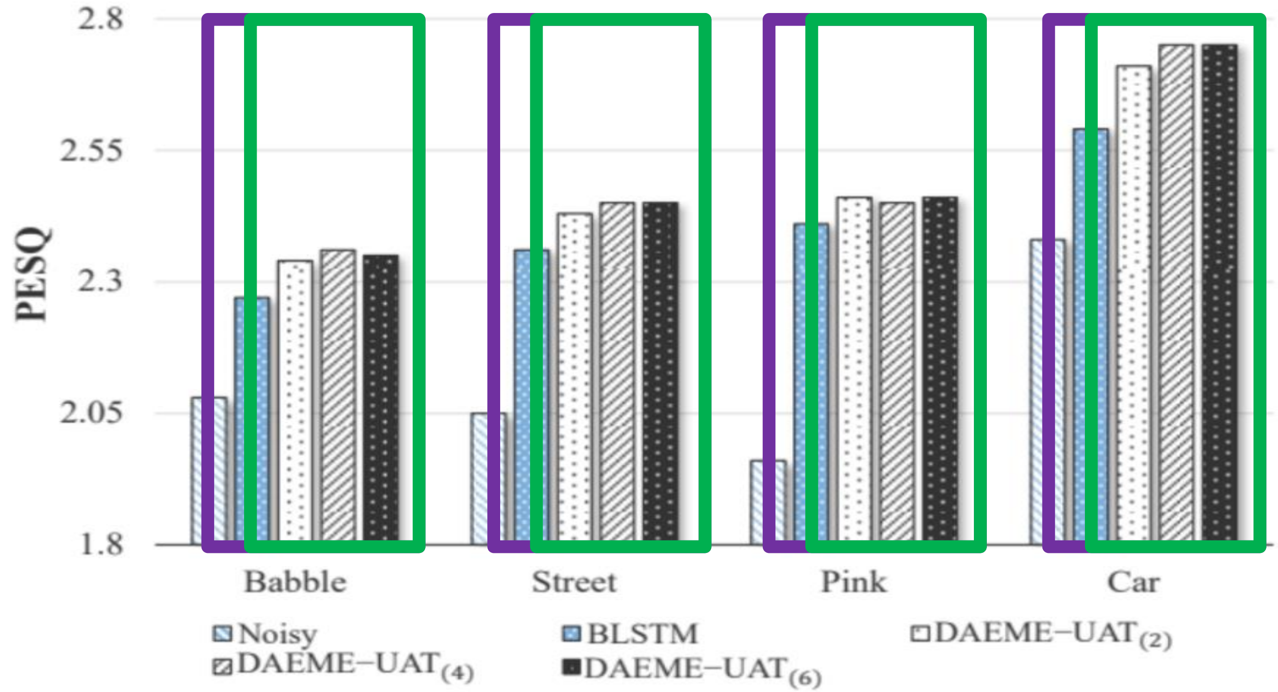




(1) Good flexibility and interpretability to combine different types of Encoder and Decoder.
(2) An utterance-attribute tree (UAT) can be used to guide the design of the multi-branched encoders.

Model Types (Ensemble Learning)

- DAEME [Yu et. al., TASLP, 2020]



(1) As compared to the SE model with a single encoder (original BLSTM system), DAEME achieves better performance

(2) When we have more SE models in the encoders (2, 4, 6), higher PESQ/STOI scores can be obtained*.

*STOI results are reported in [Yu et. al., TASLP, 2020]

Outline

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


聞


聽

大學日：心不在焉，聽而不聞

Hear but pay no attention; listen but not hear

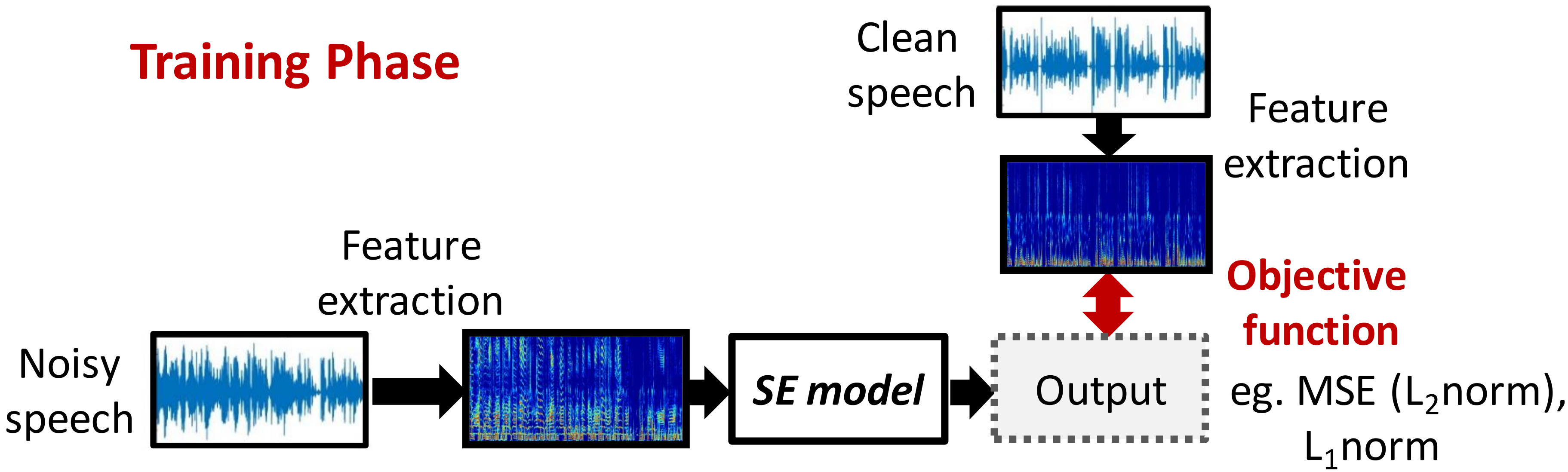
Intelligibility and Quality are different

Objective Function

Speech perception
- Quality
- Intelligibility

Automatic speech recognition
- Recognition accuracy

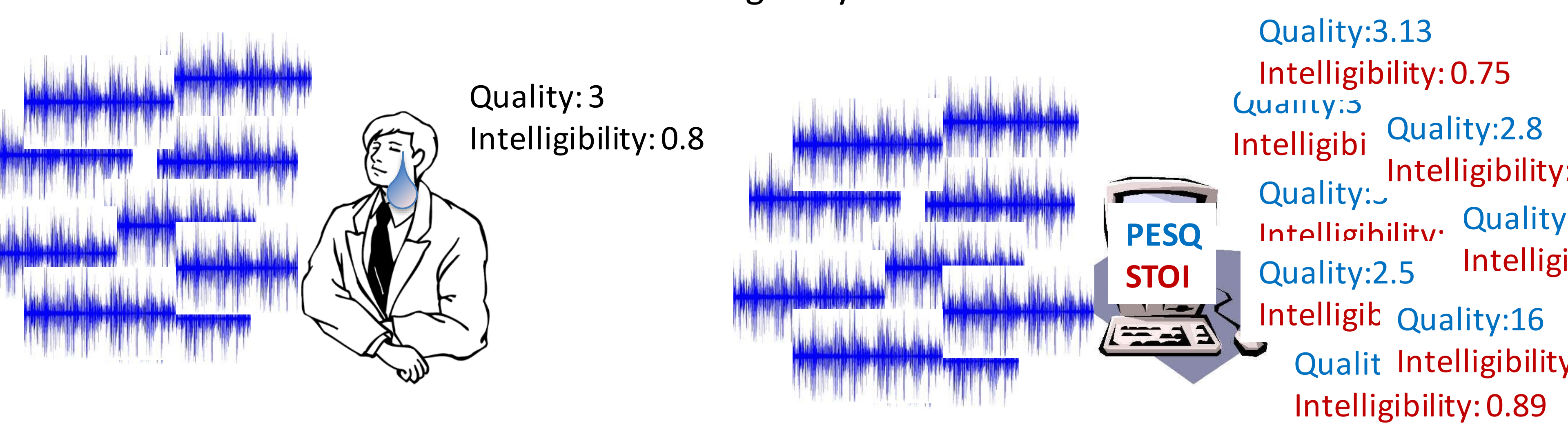
Training Phase



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



- We derived objective function based on STOI and PESQ.
- We have proposed two solutions: (1) Direct optimization on STOI[1]; (2) Generative adversarial tainting (GAN) to optimize PESQ and STOI[2].

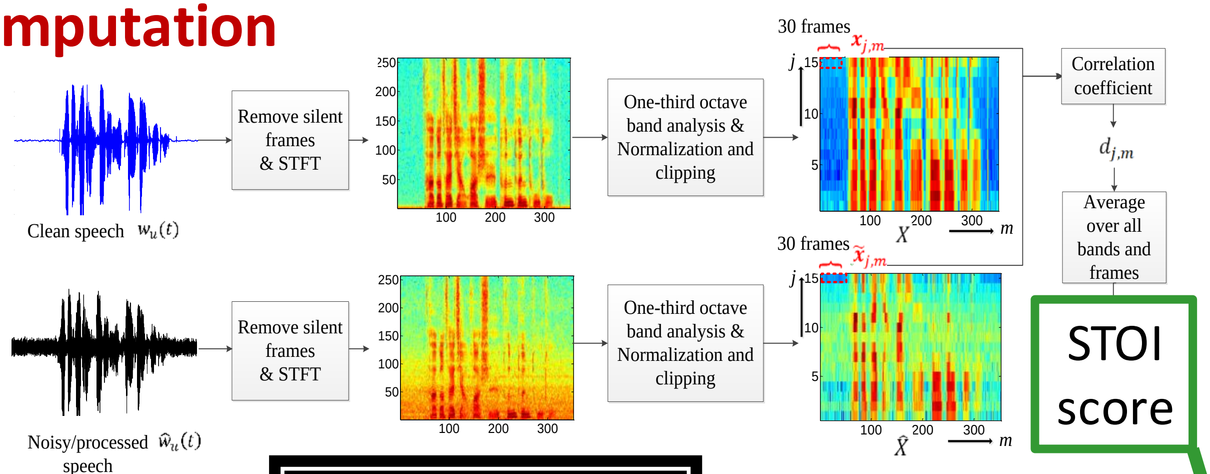
➢ "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks" IEEE TASLP 2018.

➢ "Metric GAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," ICML 2019

Objective Function

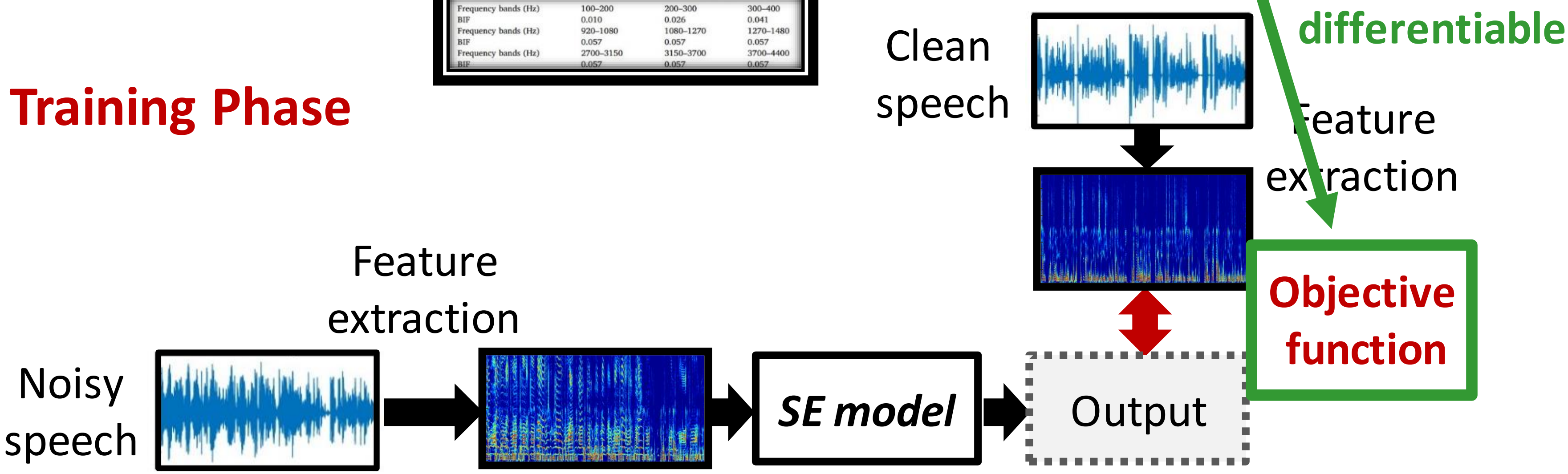
- STOI-based Objective Function [Fu et al, TASLP 2018]

STOI Computation



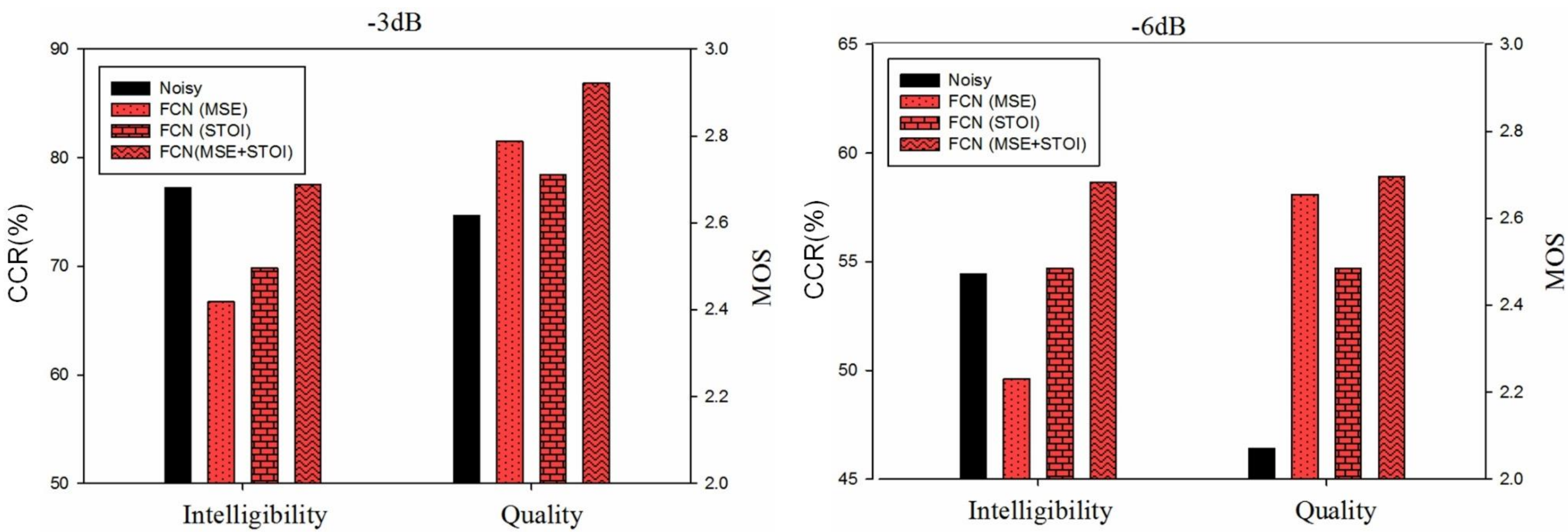
Frequency bands (Hz)	100–200	200–300	300–400
BIF	0.010	0.026	0.041
Frequency bands (Hz)	920–1080	1080–1270	1270–1480
BIF	0.057	0.057	0.057
Frequency bands (Hz)	2700–3150	3150–3700	3700–4400
BIF	0.057	0.057	0.057

Training Phase



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.

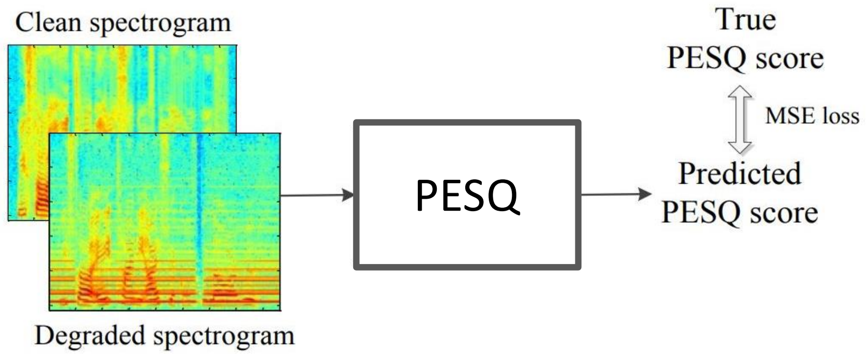
Objective Function

- 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 2020].

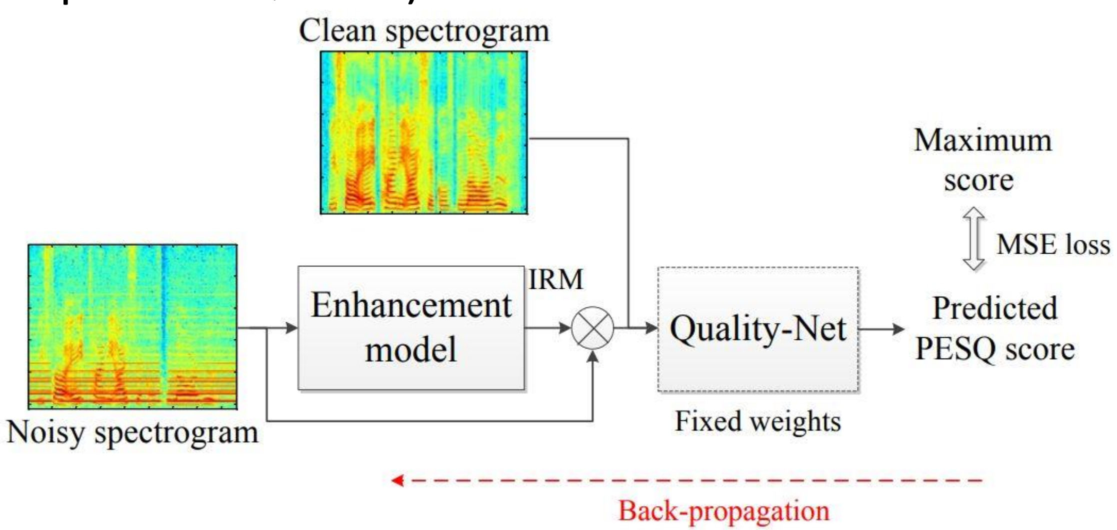
Objective Function

- PESQ-based Objective Function [Fu et al, IEEE SPL 2020]

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

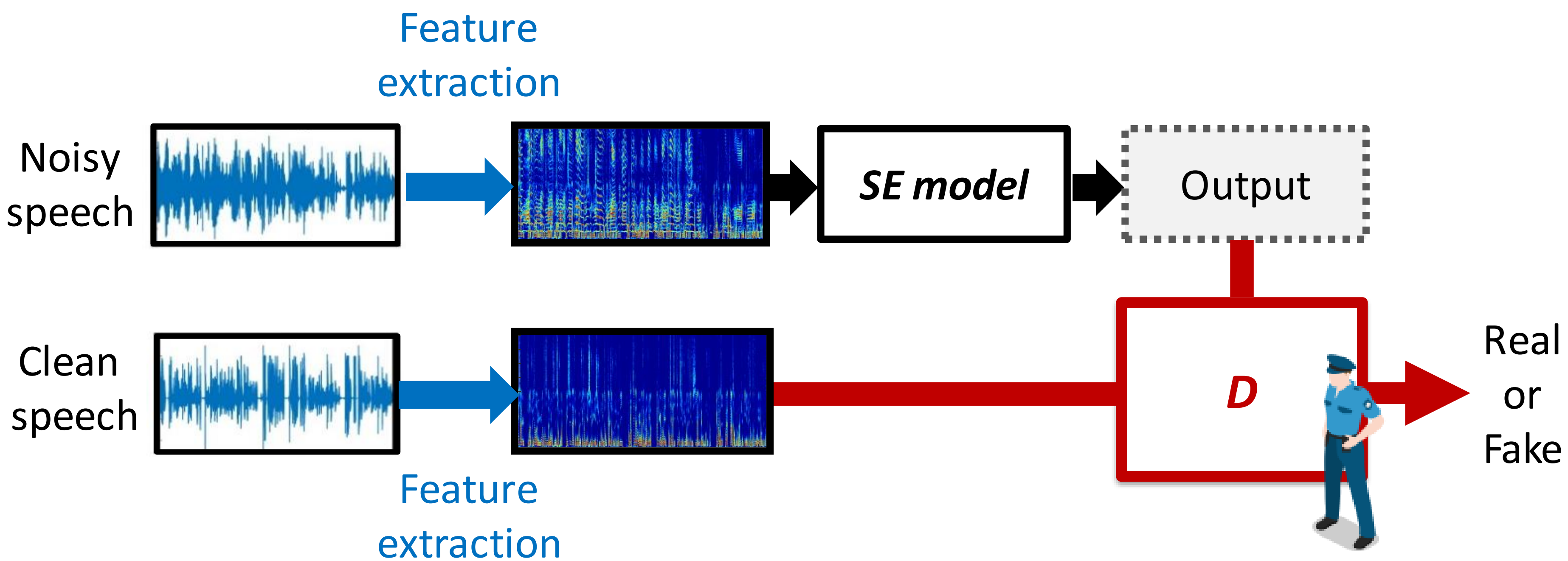


Stage 2: train the SE model based on the Quality-Net (input: paired clean and noisy speech; output: PESQ score)



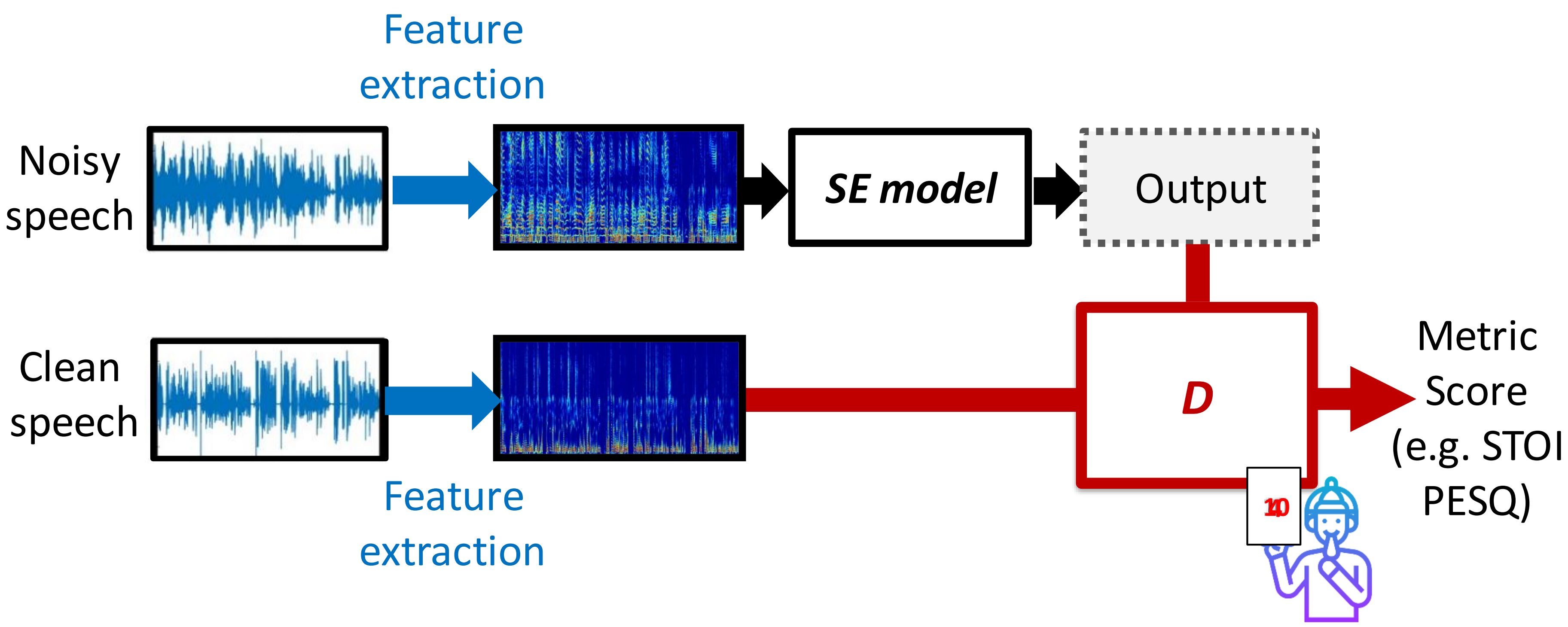
Objective Function

- 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:

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

$0 \leq Q'(G(x),y) < 1$ 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

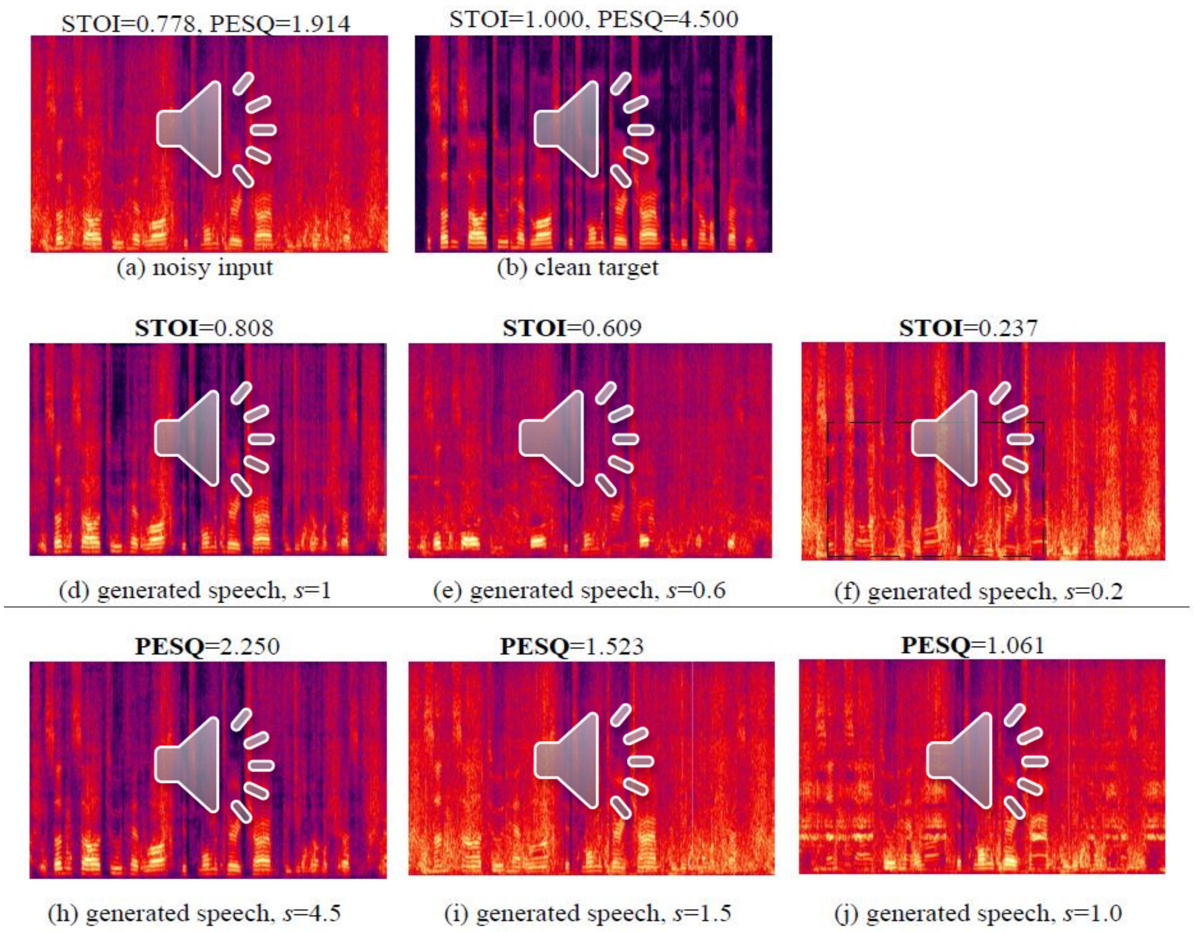
	Noisy		IRM (L1)		IRM (CGAN)		PE policy grad*(P)		MetricGAN (P)		MetricGAN (S)	
SNR (dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
12	2.375	0.919	2.913	0.935	2.879	0.936	**2.995**	0.927	2.967	0.936	2.864	**0.939**
6	1.963	0.831	2.52	0.878	2.479	0.876	2.595	0.869	**2.616**	0.881	2.486	**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



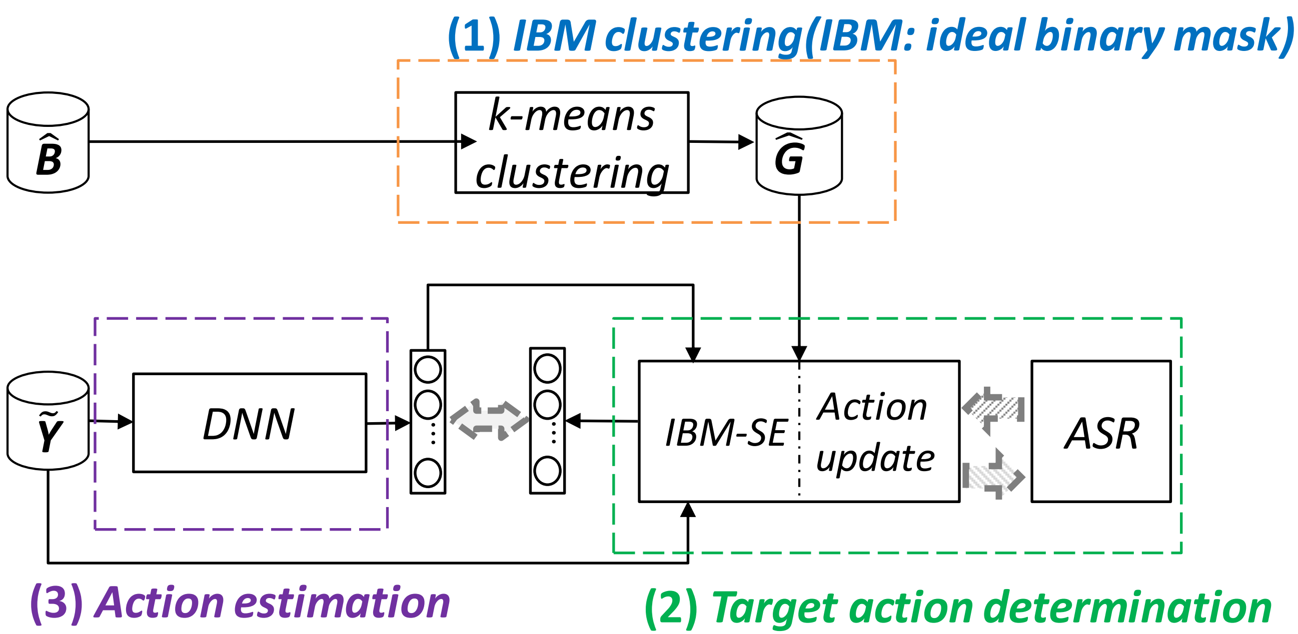
We can specify a metric score to either **increase** or **decrease** the speech **quality** or **ineligibility**.

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

Objective Function

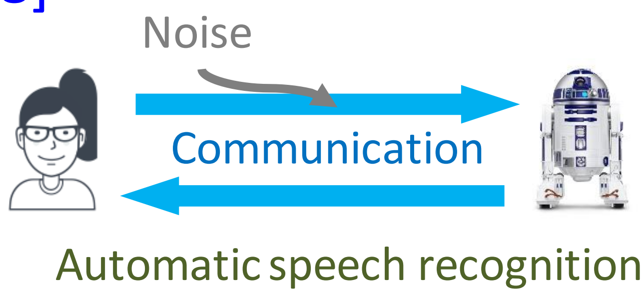
- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]

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

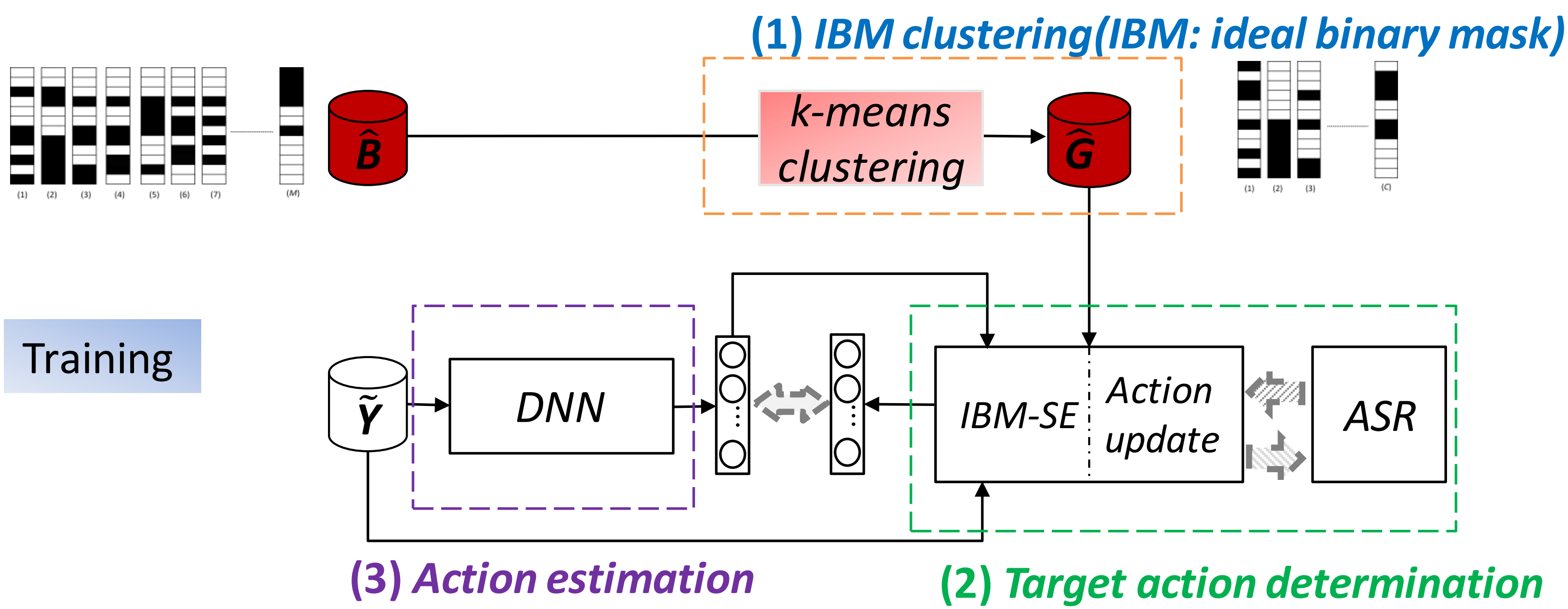


Objective Function

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



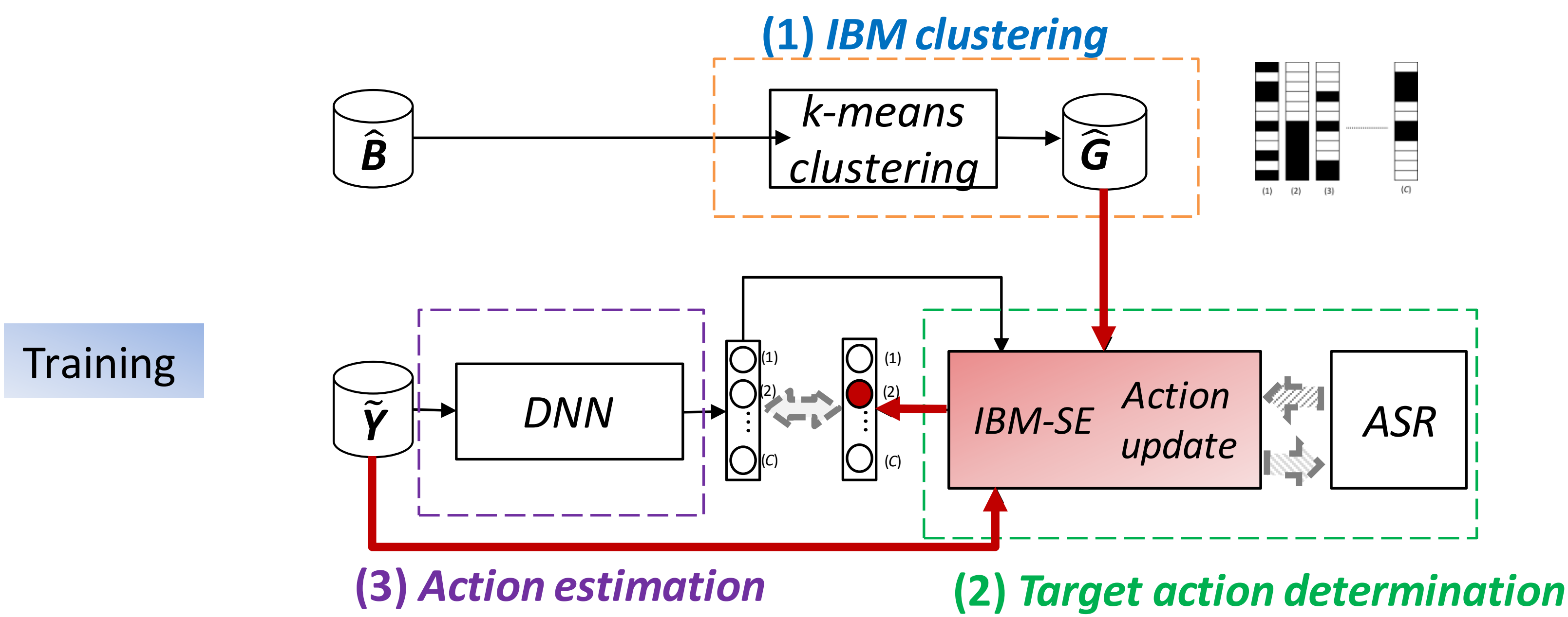
- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system



Objective Function

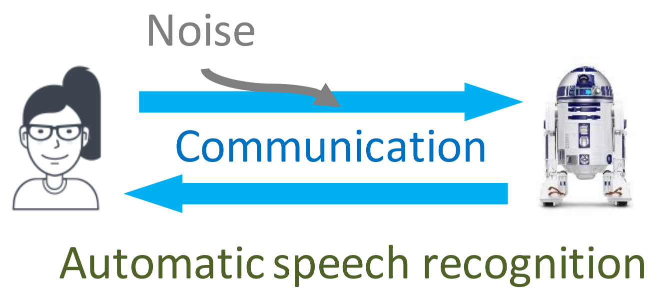
- Reinforcement learning (RL) with ASR-based rewards [Shen et al., ICASSP 2018]

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



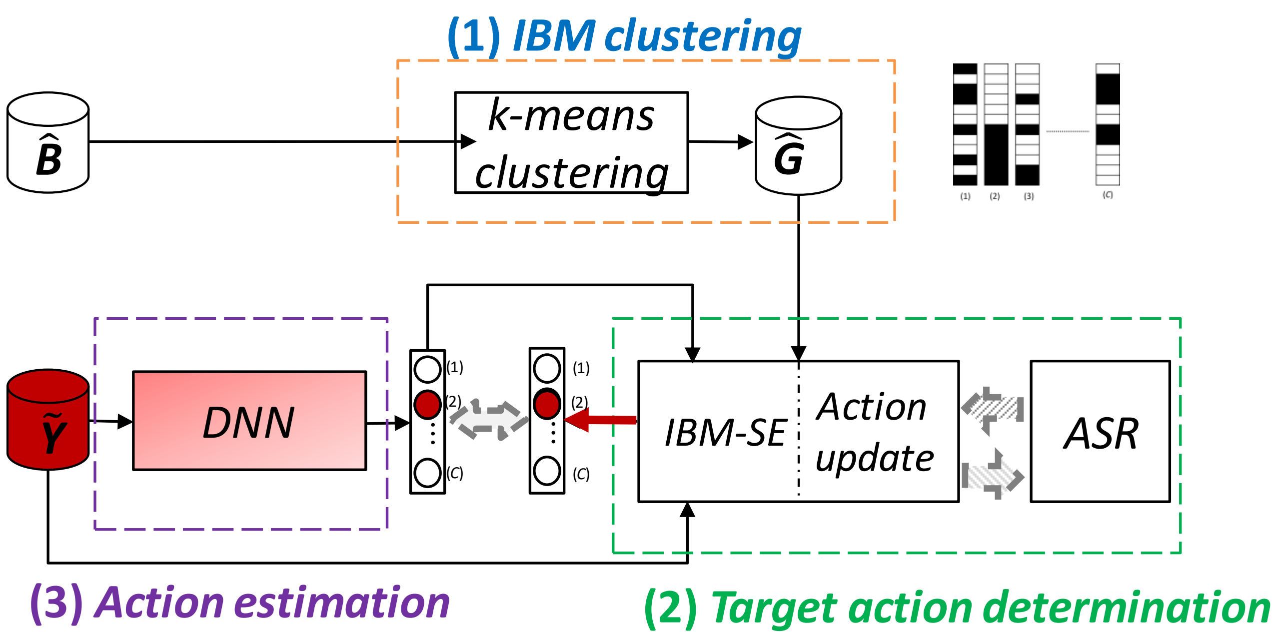
Objective Function

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

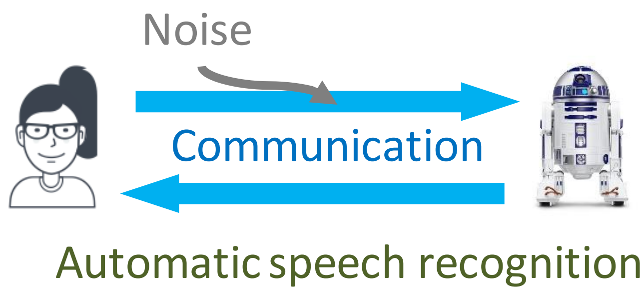


- Problem: complex correlation of acoustic features and recognition results
- Proposed solution: reinforcement learning based speech enhancement system

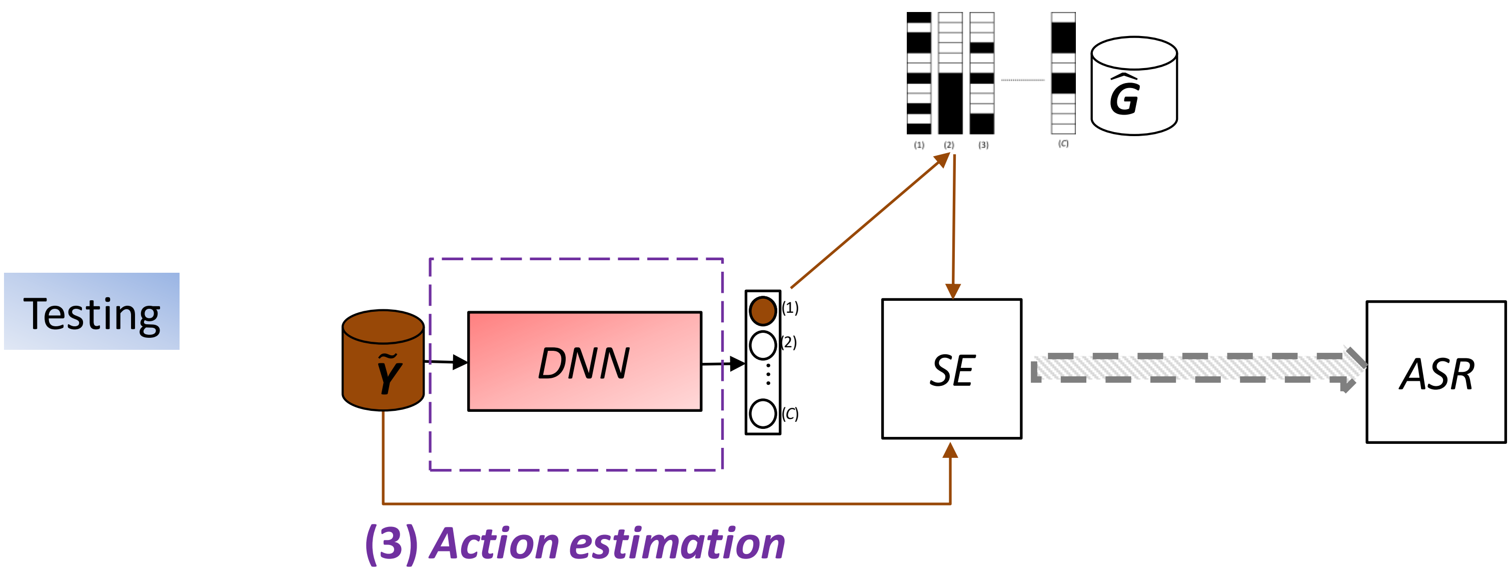
Training



Objective Function



- 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.

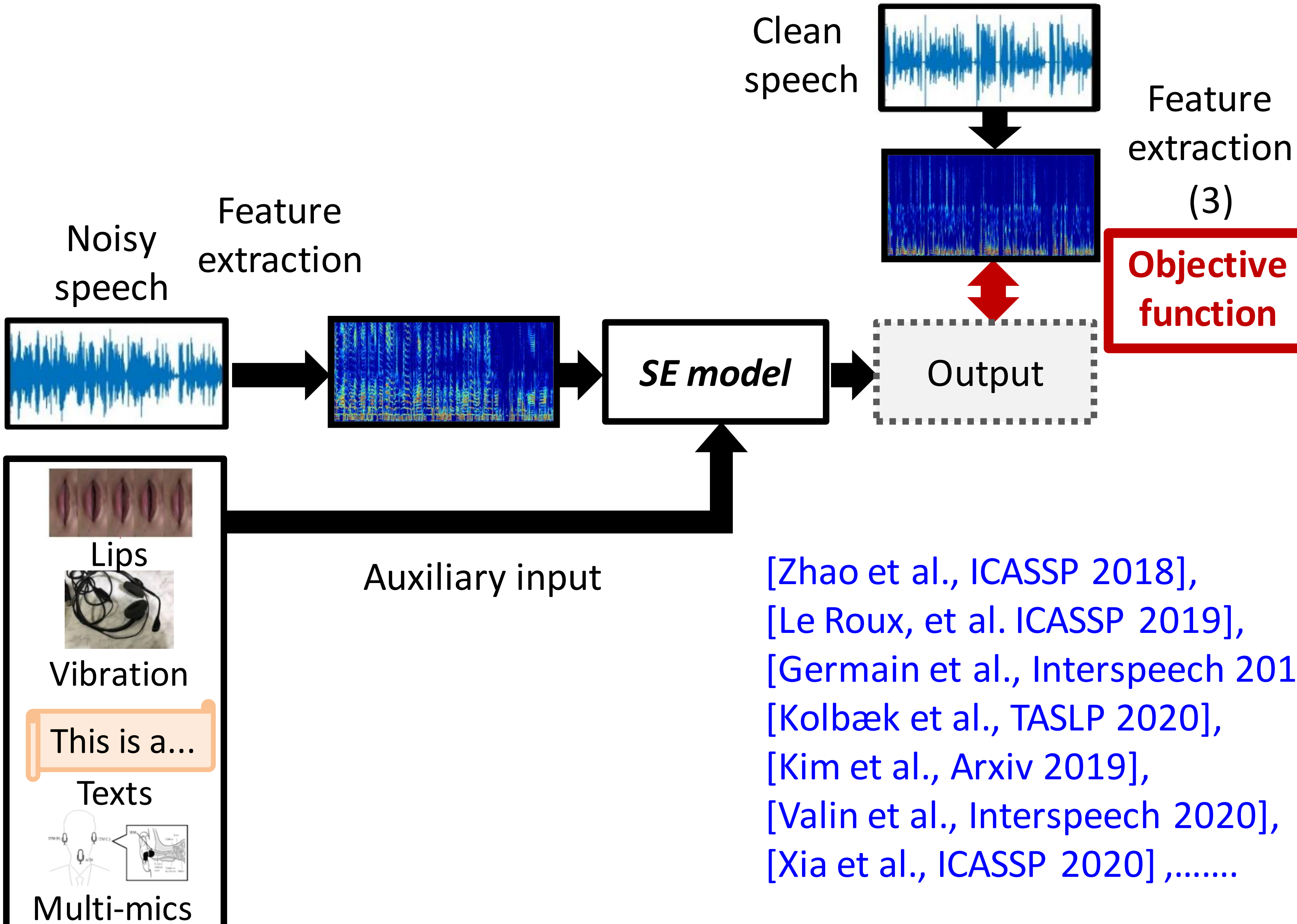
SNR	*Noisy*	$1nnSE$	$RLSE_1$	$RLSE_2$
5 dB	56.14	73.09	55.60	**49.18**
0 dB	81.40	85.79	77.20	**65.75**

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**

Speech recognition accuracy-based objective function improves ASR performance **and objective measures (human listening).**

Deep Learning Based SE System

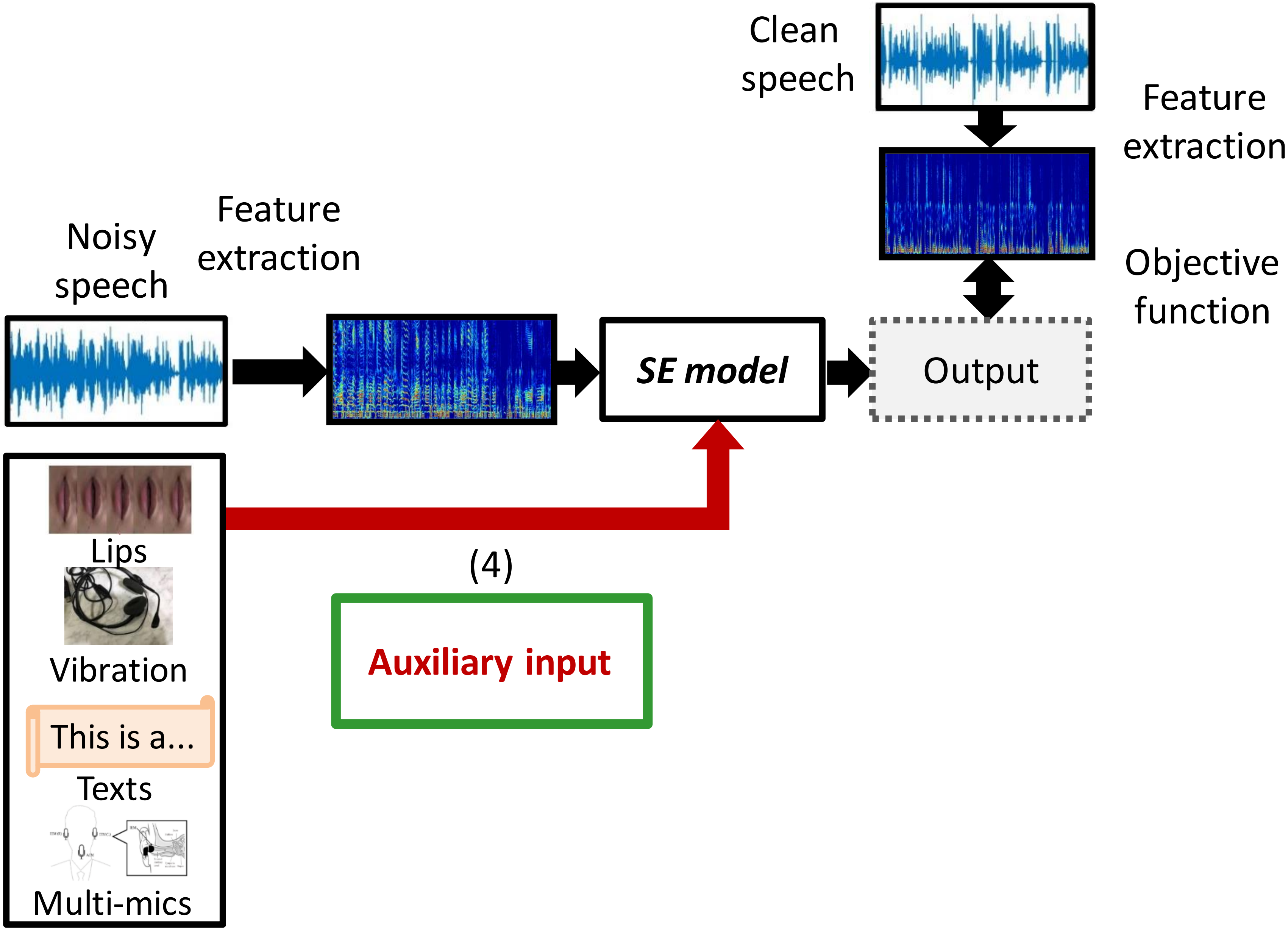


[Zhao et al., ICASSP 2018],
[Le Roux, et al. ICASSP 2019],
[Germain et al., Interspeech 2019],
[Kolbæk et al., TASLP 2020],
[Kim et al., Arxiv 2019],
[Valin et al., Interspeech 2020],
[Xia et al., ICASSP 2020] ,.......

Outline

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

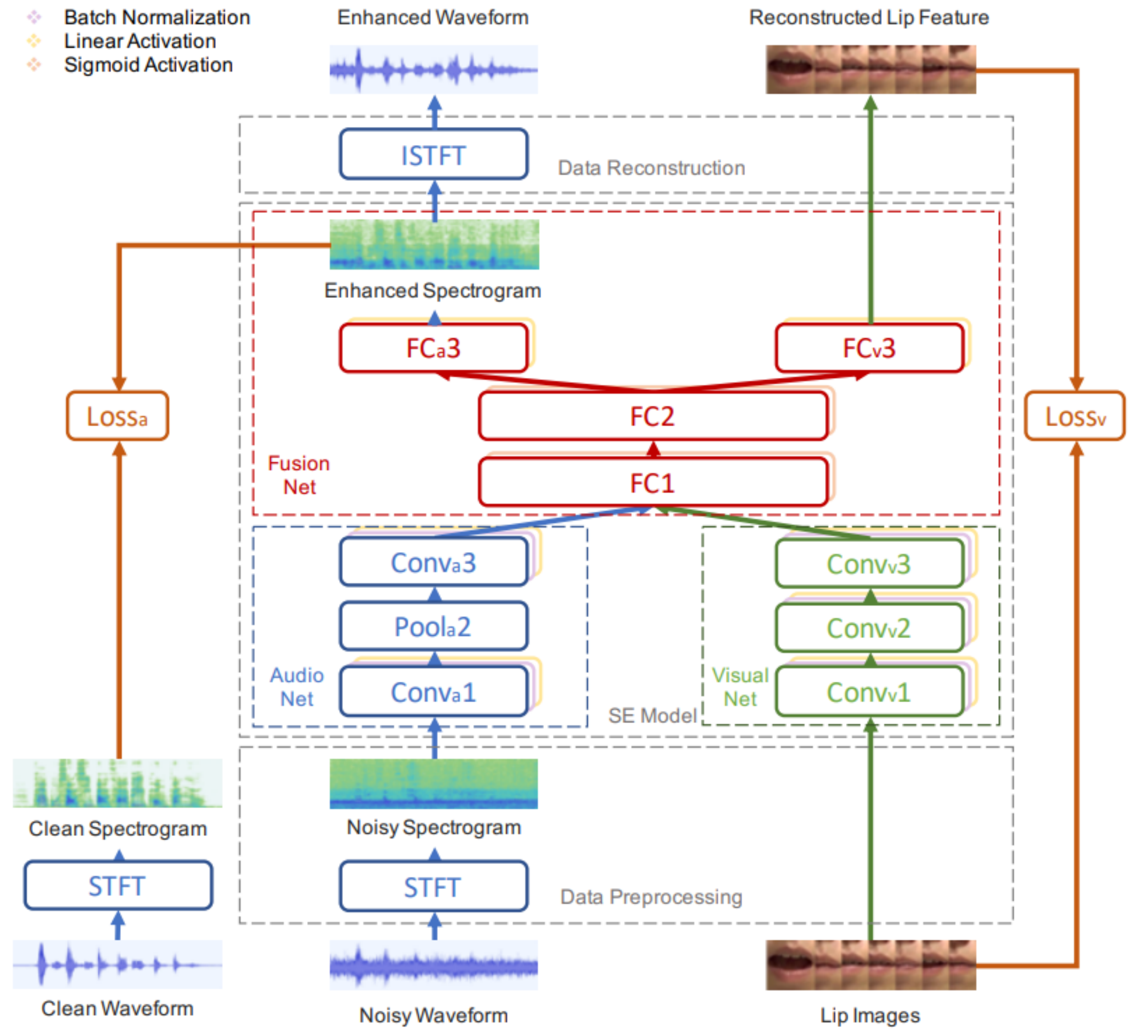
Auxiliary Input



Multimodal SE (Visual)

- Audio-visual SE [Hou et al., TETCI 2018, Sadeghi et al. TASLP 2020]

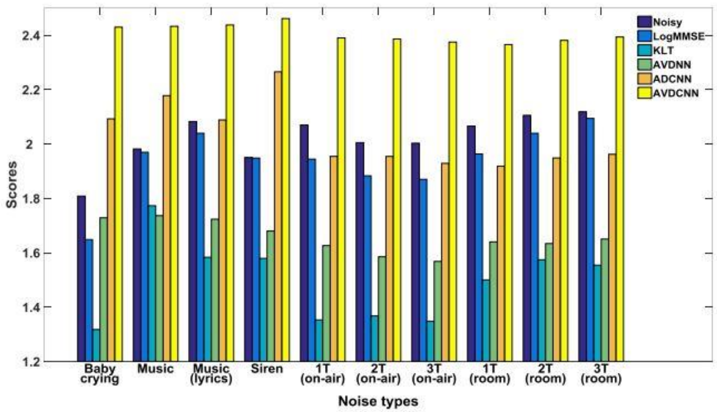
  - Additional parts
    - Lip images
    - Visual Net
    - FCv3
  - Visual target: image



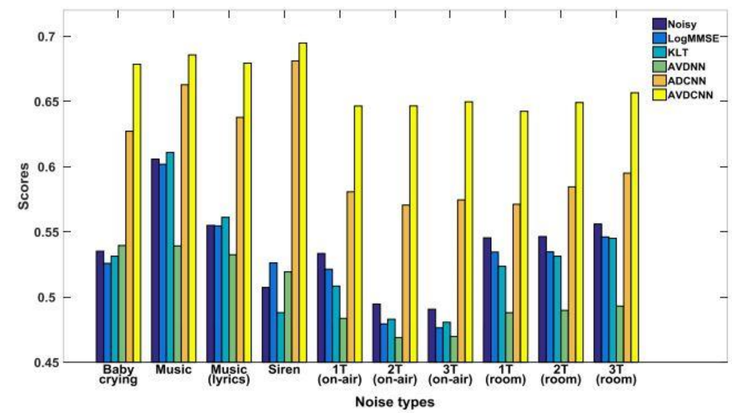
Multimodal SE (Visual)

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

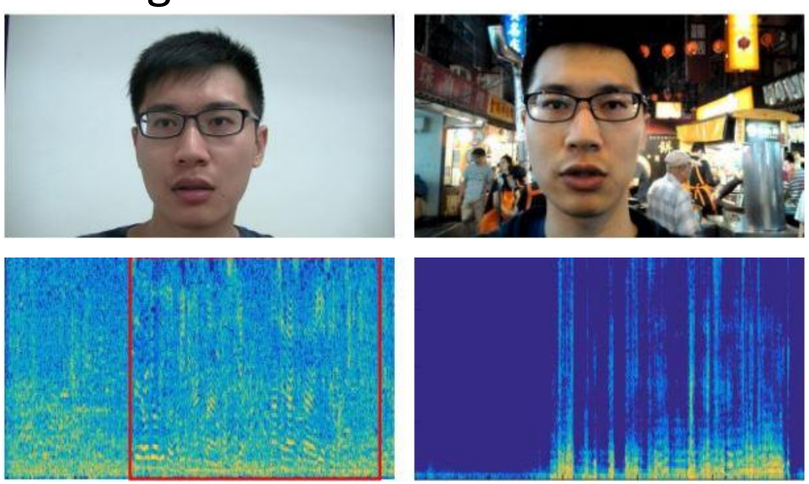
The PESQ scores



The STOI scores



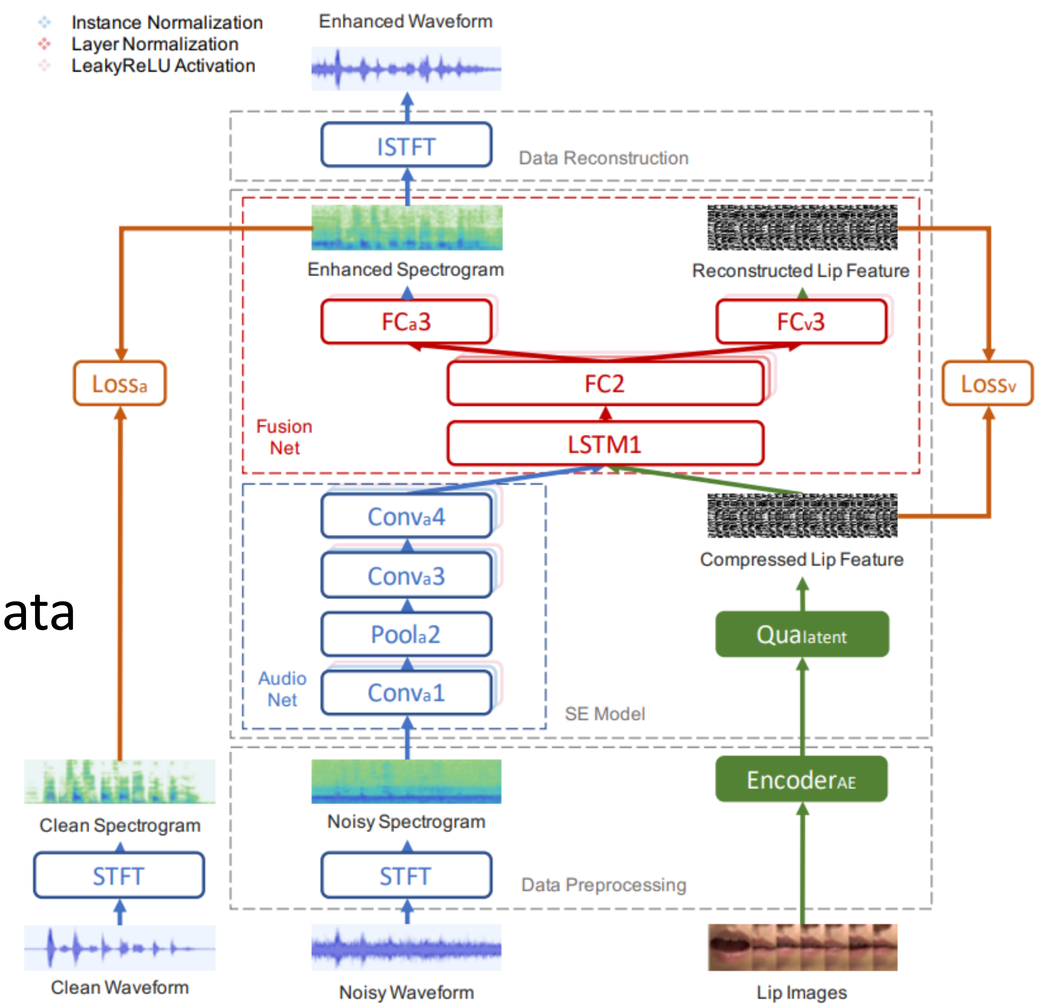
Testing in the real-world conditions



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

Multimodal SE (Visual)

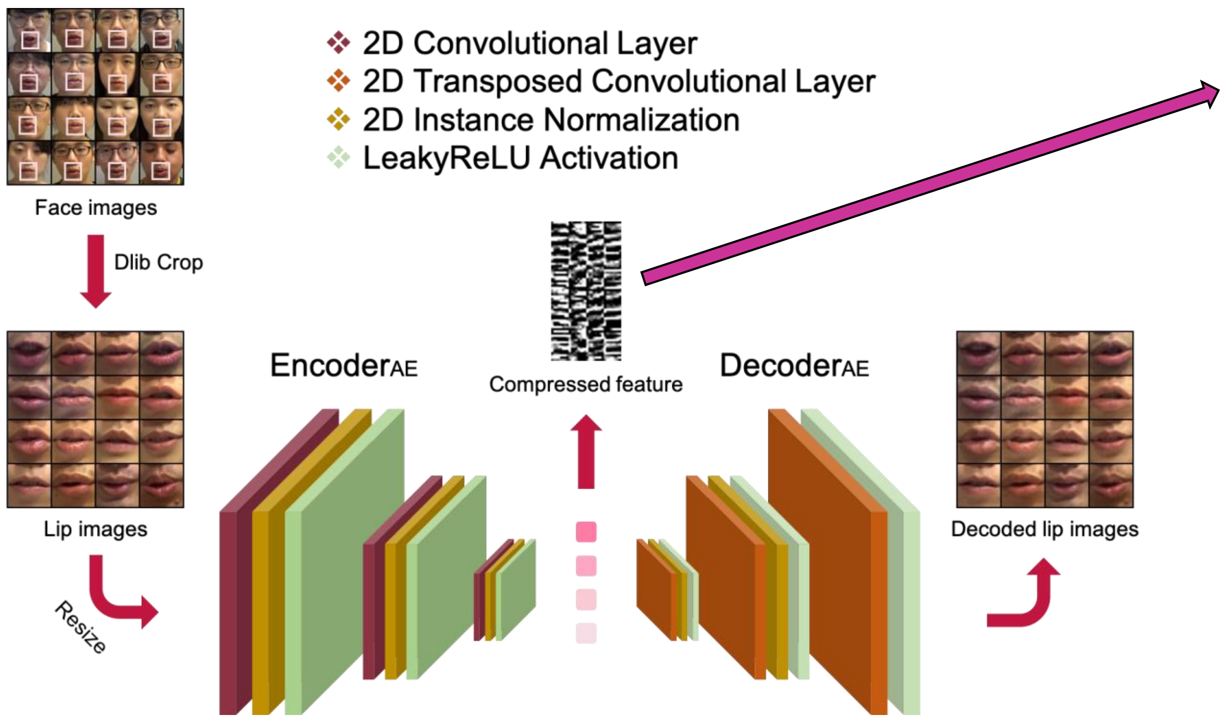
- Lite Audio-visual SE [Chuang et al., Interspeech 2020]



- Issue (1): size of images
- Issue (2): privacy issue
- LAVSE (Lite AVSE)
- EncoderAE replace visual net
- Qualatent further compress data

Multimodal SE (Visual)

- Lite Audio-visual SE [Chuang et al., Interspeech 2020]



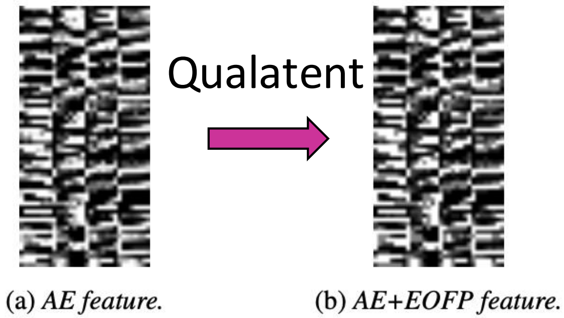


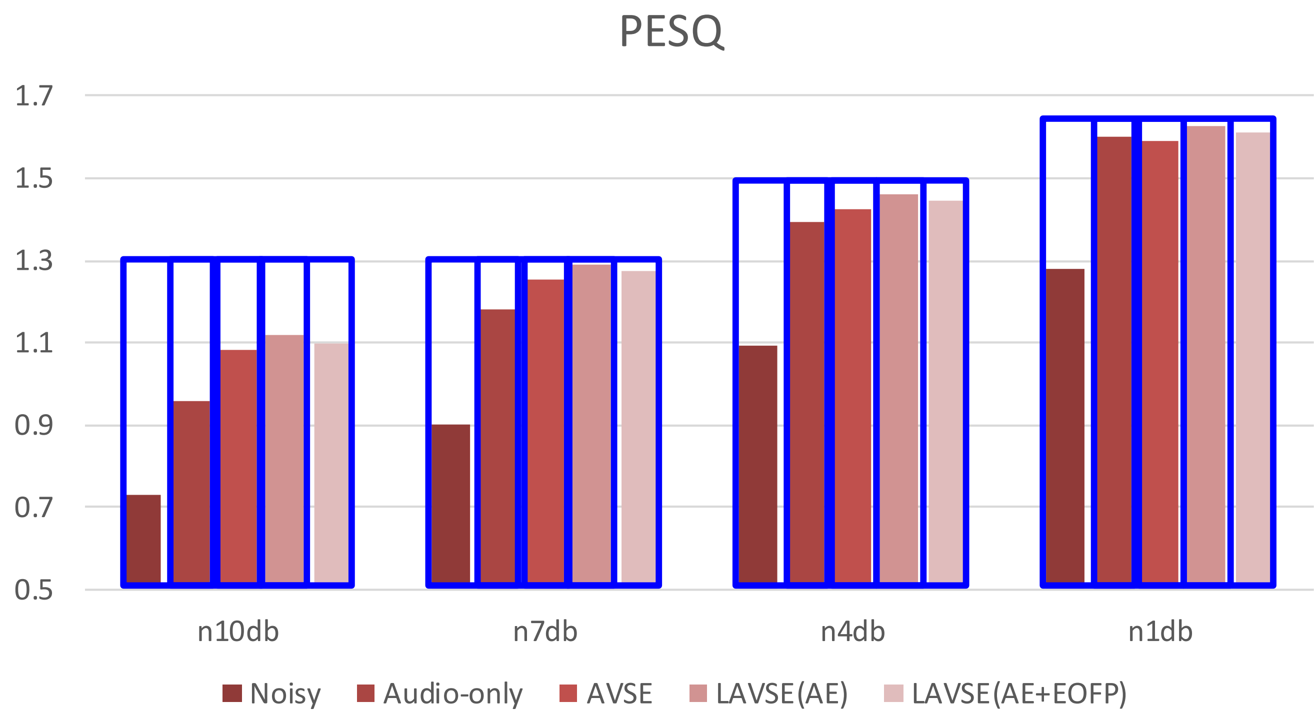
Figure 6: *Visual latent features of lips.*

EOFP (exponent-only floating point quantized) [Hsu et al., SLT 2018]

1. EncoderAE representation enhances the privacy.
2. Qualatent further compress data.

Multimodal SE (Visual)

- Lite Audio-visual SE [Chuang et al., Interspeech 2020]

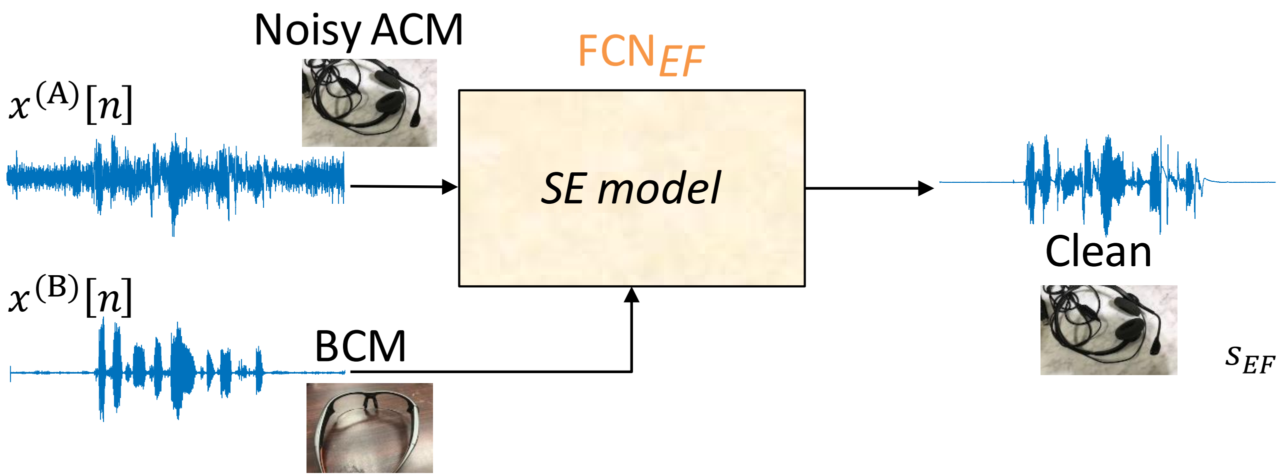


1. Lite AVSE outperforms original AVSE.
2. AVSE+EOFP slightly underperforms AVSE with a notable reduction of 48 times on the visual features.

TMSV dataset: https://drive.google.com/drive/folders/1B-eJs1yYVf0qHrYOWrtxYs3a8inPHm1K

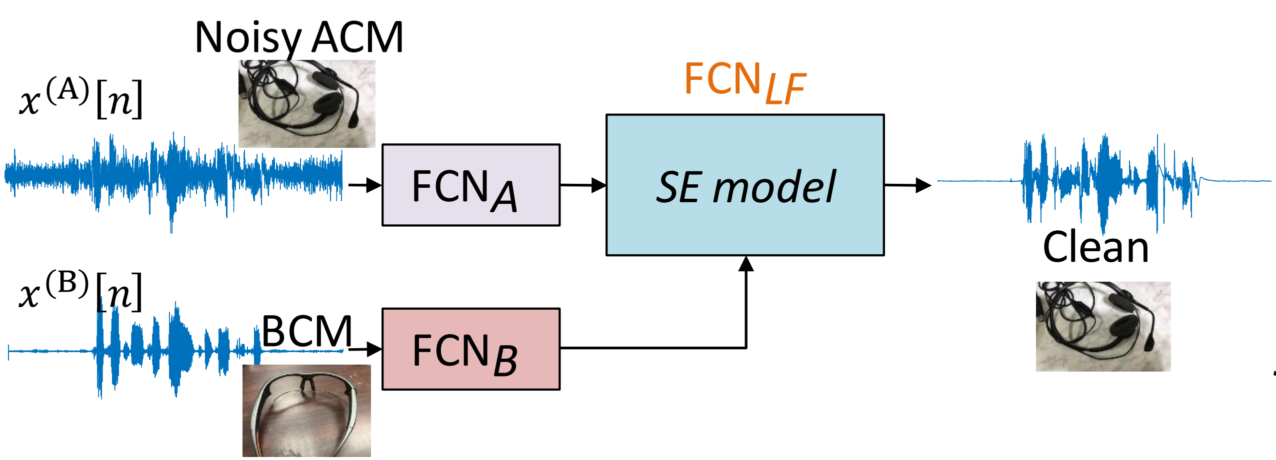
Multimodal SE (Bone-conducted)

- BCM-ACM versus BCM or ACM only [Yu et al., SPL 2020]
  - The input of FCN_{EF} combines both noisy and BCM signals



$$s_{EF} = FCN_{EF}\{x^{(A)}[n], x^{(B)}[n]\}$$

  - The input of the $Fusion$ function is processed noisy and BCM signals



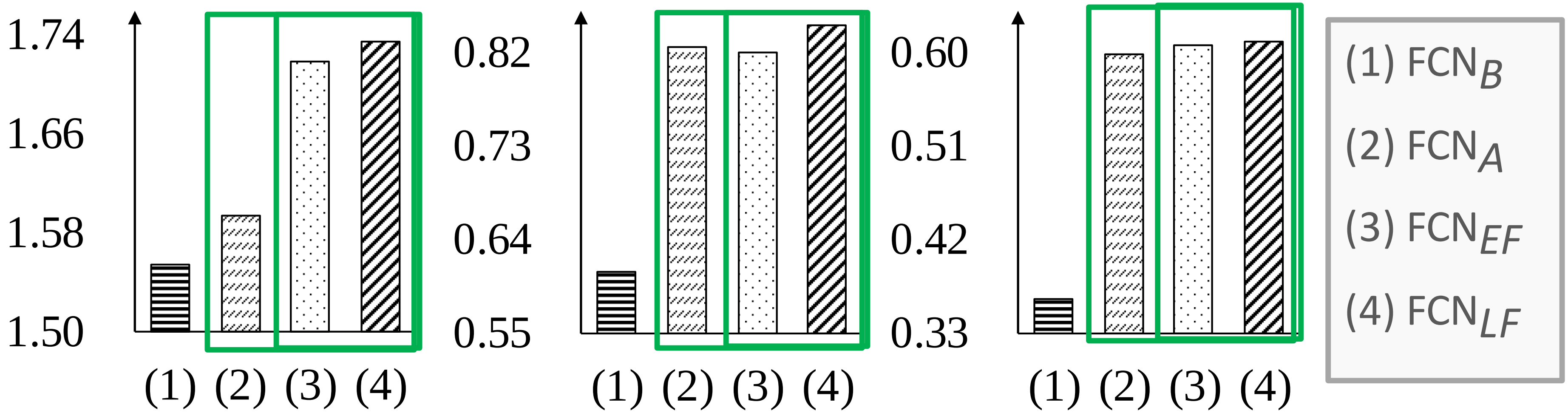
$$s^{(A)} = FCN_A\{x^{(A)}[n]\}$$

$$s^{(B)} = FCN_B\{x^{(A)}[n]\}$$

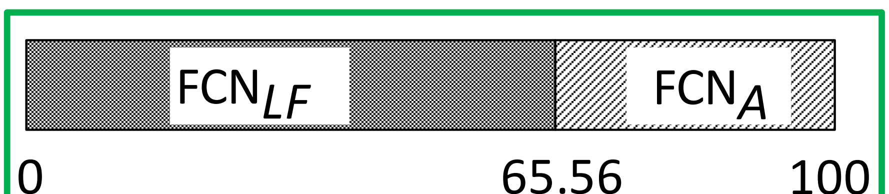
$$s_{LF} = FCN_{LF}\{s^{(A)}[n], s^{(B)}[n]\}$$

Multimodal SE (Bone-conducted)

- BCM-ACM versus BCM or ACM only [Yu et al., SPL 2020]



(a) PESQ (b) STOI (c) ESTOI



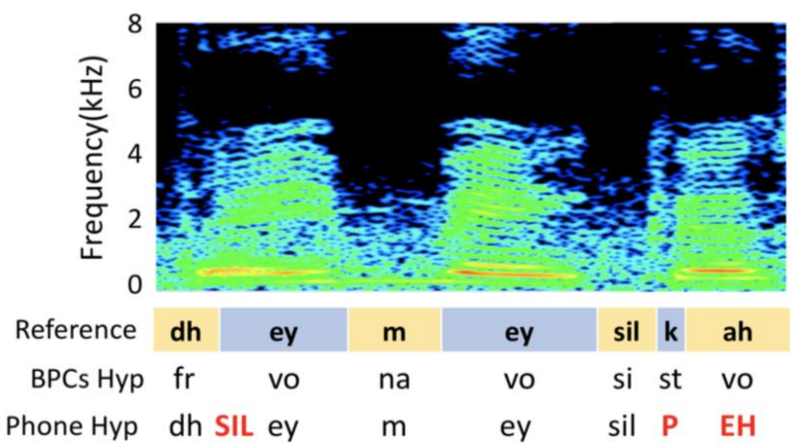
The results (in percentage, %) for the AB test that compares FCN_{LF} and FCN_A.

$(p = 0.00088 < 0.01)$

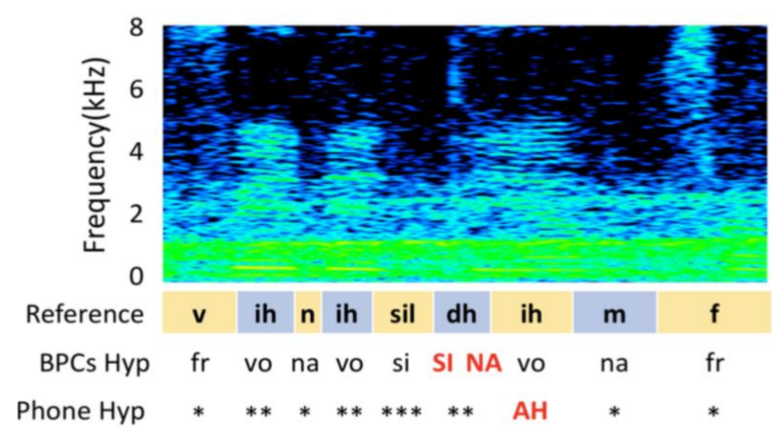
(1) BCM information improves the SE performance in terms of PESQ, STOI, ESTOI and listening tests.
(2) Late fusion outperforms early-fusion.

Multimodal SE (Text)

- Broad Phone Classes (BPC) SE [Lu et al., Interspeech 2020]
  - Main idea
    - In noisy conditions, knowing **speech contents** facilitates listeners to more effectively **retrieve pure speech signals**.
    - **Phone recognizer** can be used to obtain phonemes (text) information.
    - Recognized phonemes may be **erroneous** and thus **misguide** the SE process.
    - We used the broad phone class (BPC) instead, which is built by: **place of articulatory** and **manner of articulatory** and **data-driven criterion**
  - Recognition results



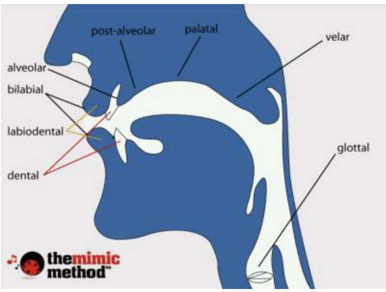
(a) *Spectrogram and recognition result at 10dB SNR*



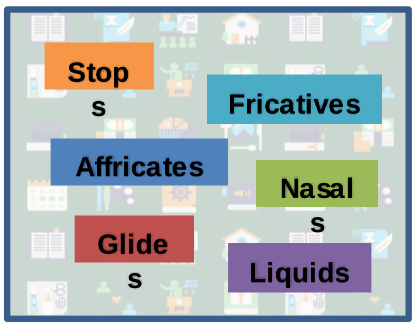
(b) *Spectrogram and recognition result at 0dB SNR level*

Multimodal SE (Text)

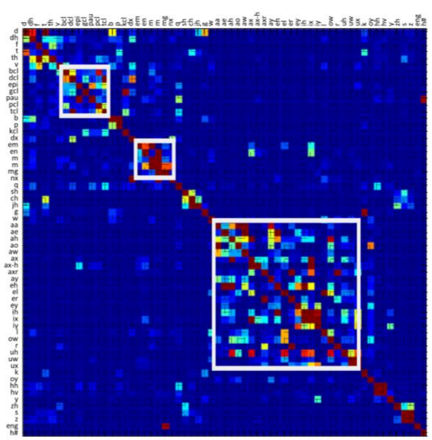
- Broad Phone Classes (BPC)-SE [Lu et al., Interspeech 2020]



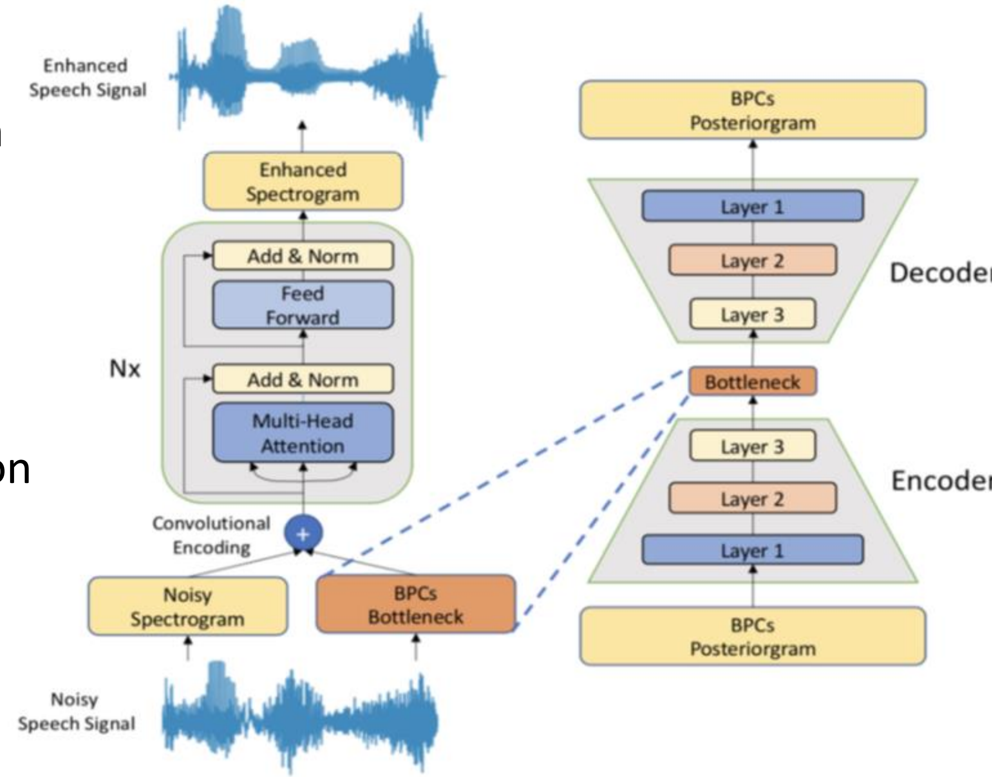
Place of articulation
(PBPCs)



Manner of articulation
(MBPCs)



Data-driven BPCs
(DBPCs)



1. Train an recognizer (BPC speech recognizer) to estimate BPCs in each frame.
2. Know which phone/BPC help the SE model to generate better enhanced speech.

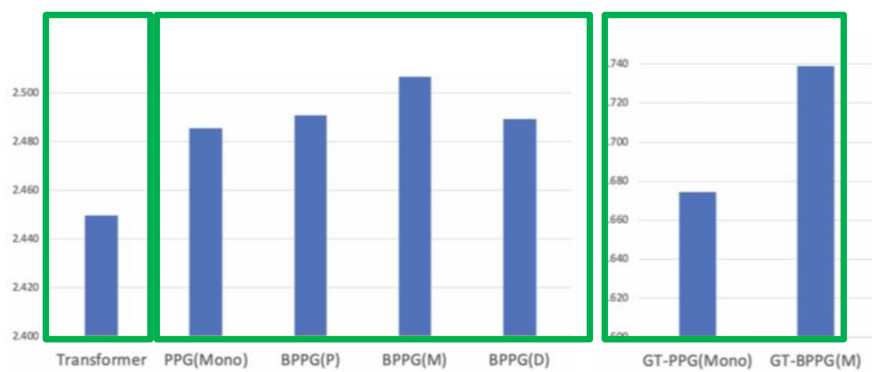
Multimodal SE (Text)

- Broad Phone Classes SE [Lu et al., Interspeech 2020]

The STOI scores

SNR	Noisy	LSTM	Transformer	PPG(Mono)	Broad Phone Class			Ground Truth	
					BPPG(P)	BPPG(M)	BPPG(D)	GT-PPG(Mono)	GT-BPPG(M)
-5	0.595	0.548	0.620	0.616	**0.629**	**0.627**	**0.628**	0.679	**0.708**
0	0.701	0.686	0.755	0.759	**0.765**	**0.765**	**0.763**	0.796	**0.808**
5	0.800	0.815	0.851	0.859	**0.860**	**0.861**	**0.859**	0.876	**0.879**
10	0.880	0.900	0.912	0.917	**0.918**	**0.918**	**0.917**	0.924	**0.925**
15	0.935	0.946	0.948	0.950	**0.951**	**0.950**	**0.951**	0.953	**0.953**
Avg	0.782	0.779	0.817	0.820	**0.824**	**0.824**	**0.823**	0.846	**0.855**

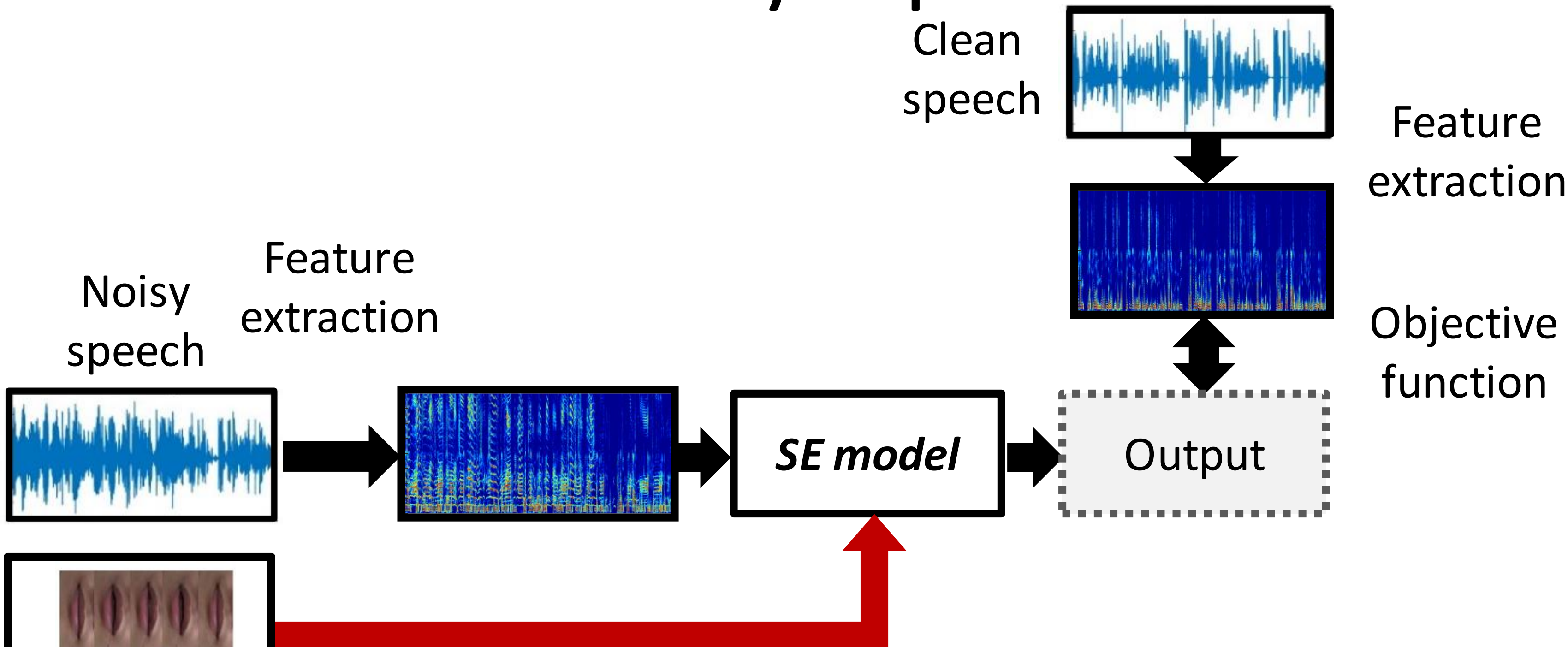
The PESQ scores

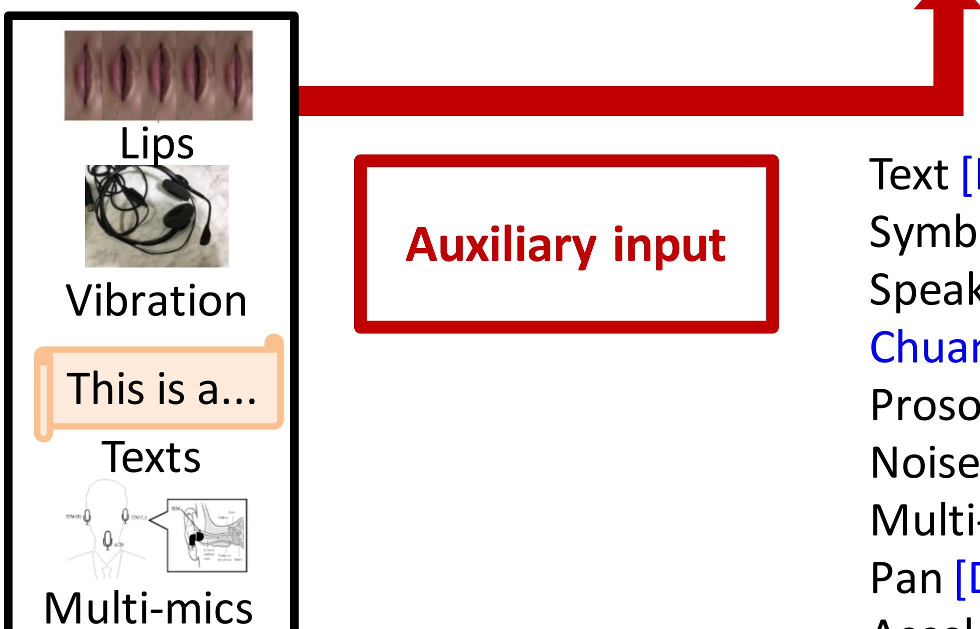


(a) *baseline model and proposed method* (b) *ground truth*

(1) Both Mono(phone) and BPC based PPGs improve the SE performance.
(2) BPC is more robust against different SNR ratios than Mono.

Auxiliary Input





Text [Kinoshita et al., In terspeech 2015],
Symbolic [Liao et al., Interspeech 2019],
Speaker Identity [Koizumi et al., ICASSP 2020; Chuang et al., Interspeech 2019],
Prosodic features [Lin et al., APSIPA 2019],
Noise token [Li et al., Interspeech 2020],
Multi-mic [Liu et al., TASLP 2020],
Pan [Du et al., ICASSP 2020],
Acceler. [Tagliasacchi et al, Interspeech 2020].

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.

weights
32-bit float

-2.743	-1.536	-0.401	1.229
1.008	-0.705	0.474	-0.401
-0.910	1.589	0.813	-1.894
0.351	-0.887	0.958	0.315

16×32-bit =(512) bits

K-means →

cluster 2-bit	value
0	0.334
1	1.383
2	-0.597
3	-1.690

Cluster index
2-bit

3	3	2	1
1	2	0	2
2	1	0	3
0	2	1	0

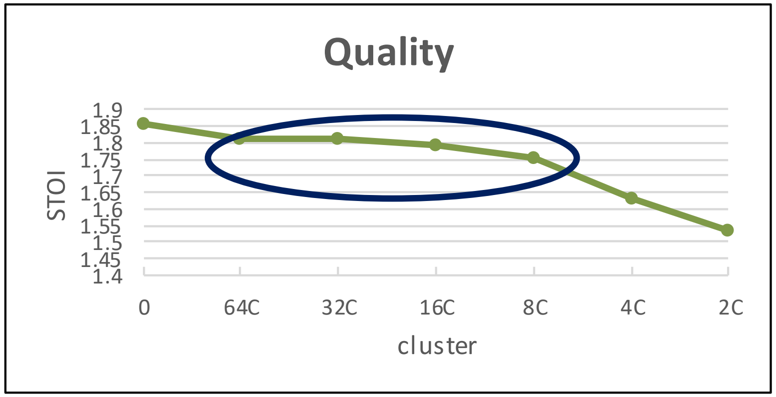
4×32-bit + 16×2-bit =(160) bits

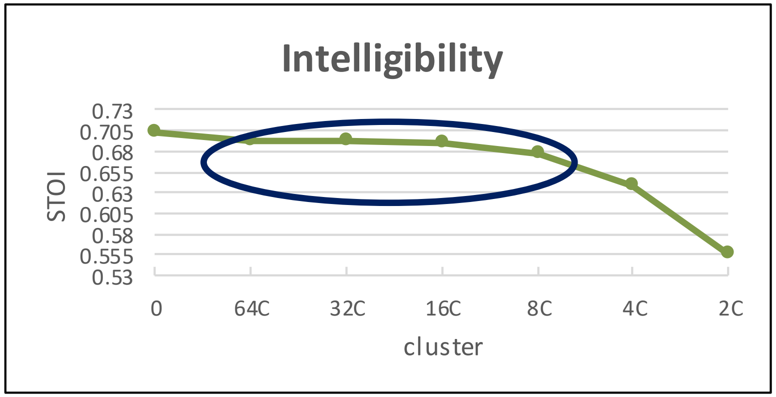
Model Compression (WS-SE)

- 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
Original	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

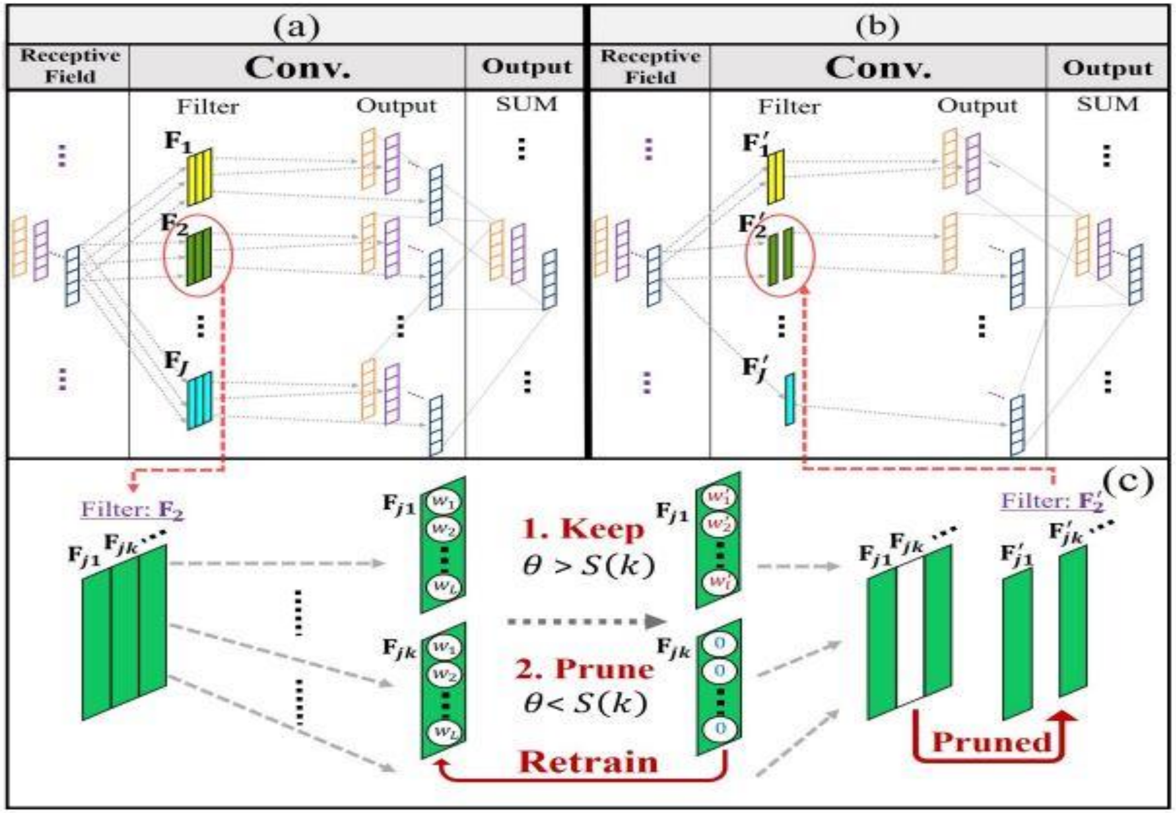




(1) Performance does not change much when the cluster number increases from 0 to 16.
(2) However, the performance drops significantly when K> 16.

Model Compression

- Parameter Pruning (PP)
  - The goal is to removing redundant parameters in an SE model.
  - Computing a sparsity score for each channel.
  - Removing channels with high sparsity scores.



Model Compression

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

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

$$M = \frac{\sum_{n,w} |k_{nw}|}{N \times W}$$

N: number of channels
W: number of weights

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

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

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

With a lower threshold, more parameters will be pruned.

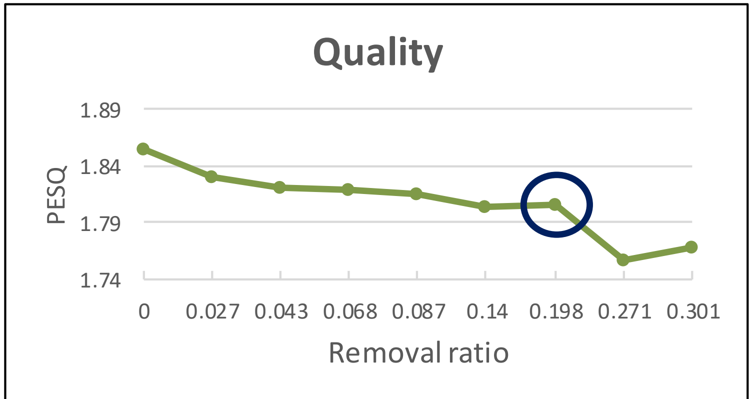
Model Compression (PP-SE)

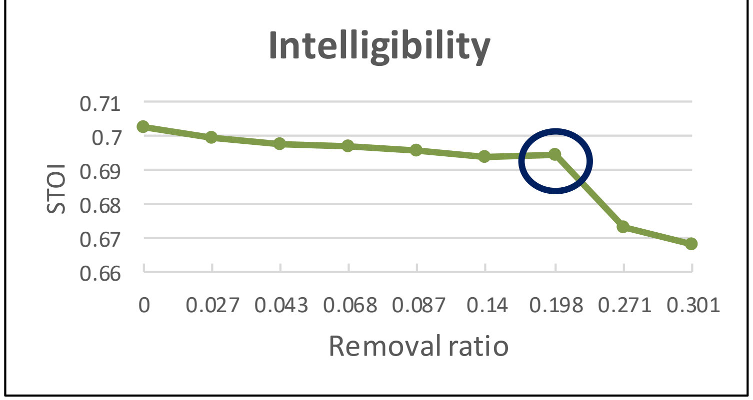
- The results of PP

A Threshold Θ is specified

If $sparsity > \Theta$, the channel 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

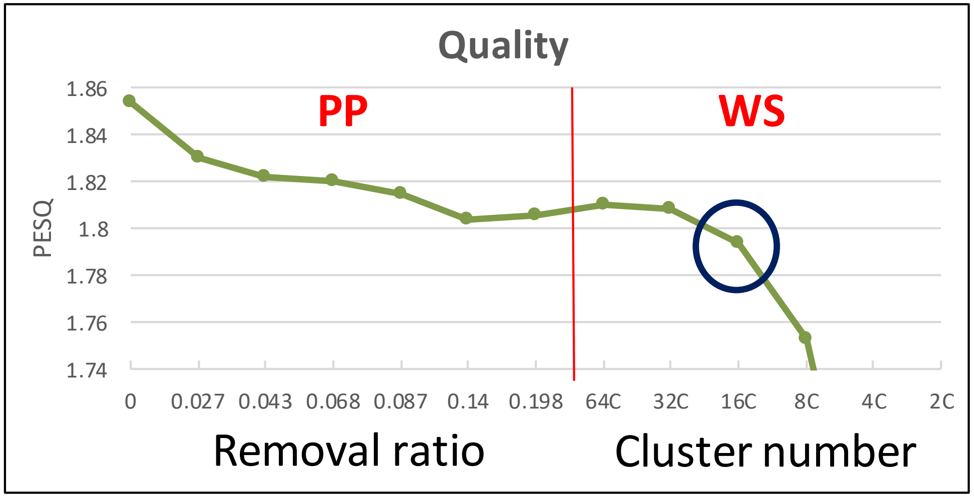


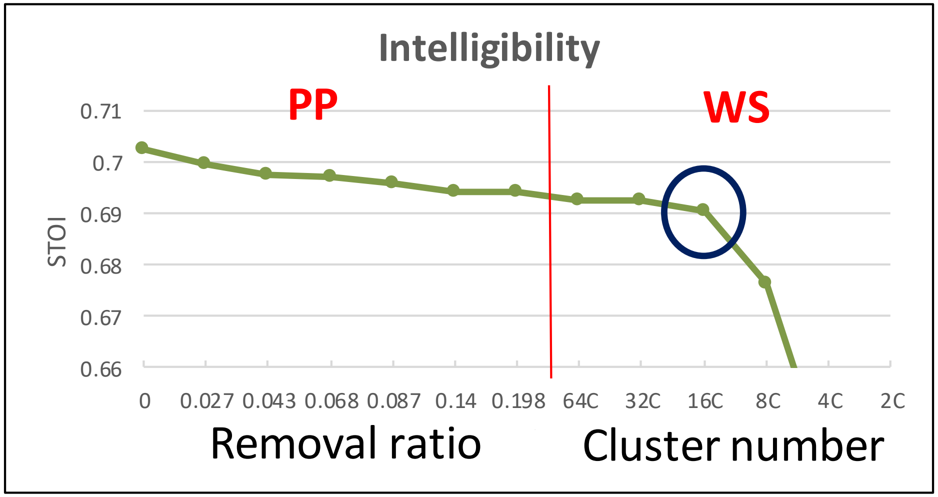


A notable performance drop when Threshold <0.7.

Model Compression (PP+WS SE)

- The results of PP+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)

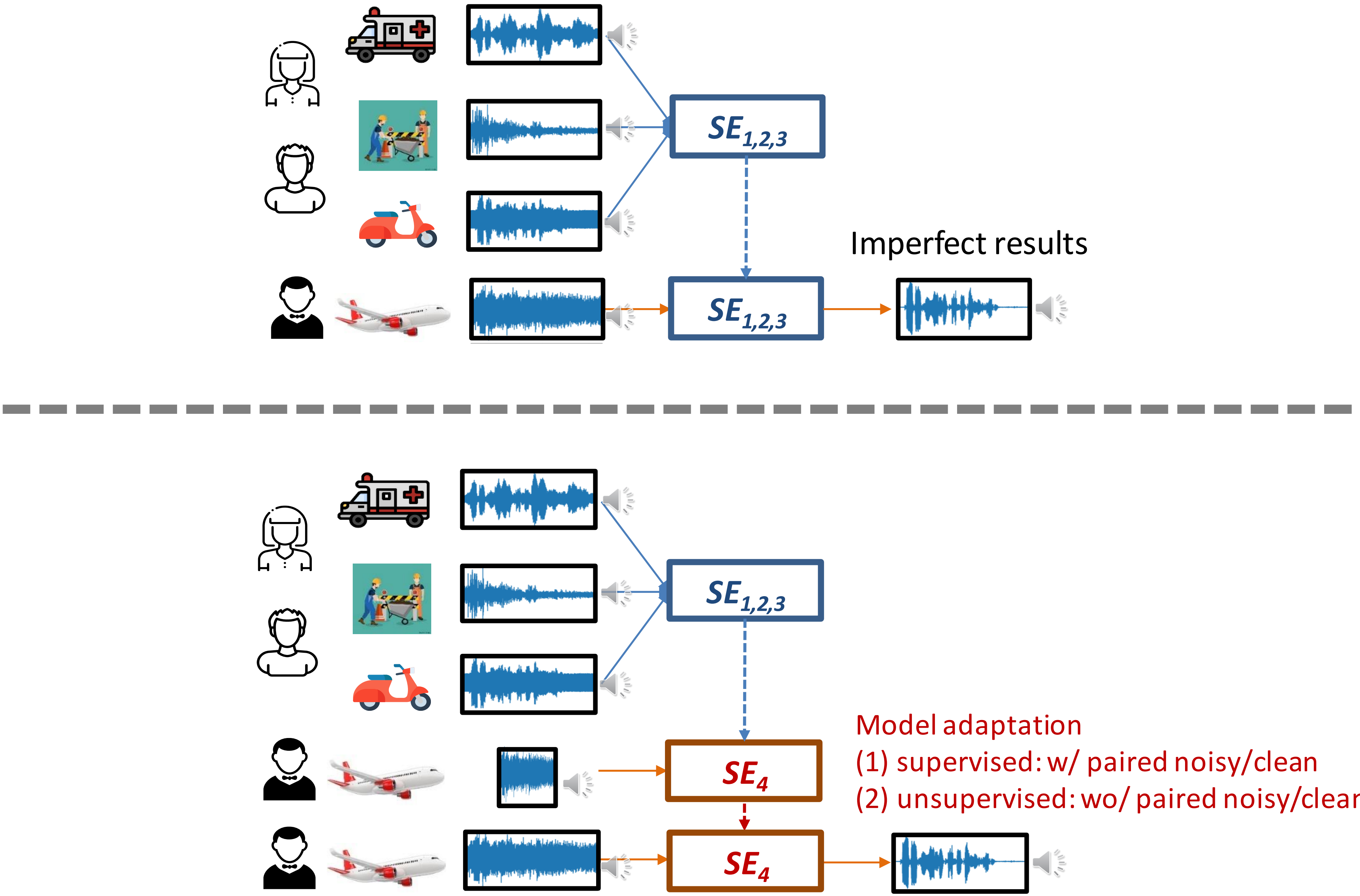




(1) The **model size** of the compressed model is only **9.76%** as compared to the original model.
(2) The **computation cost** is reduced by **20%**.

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**

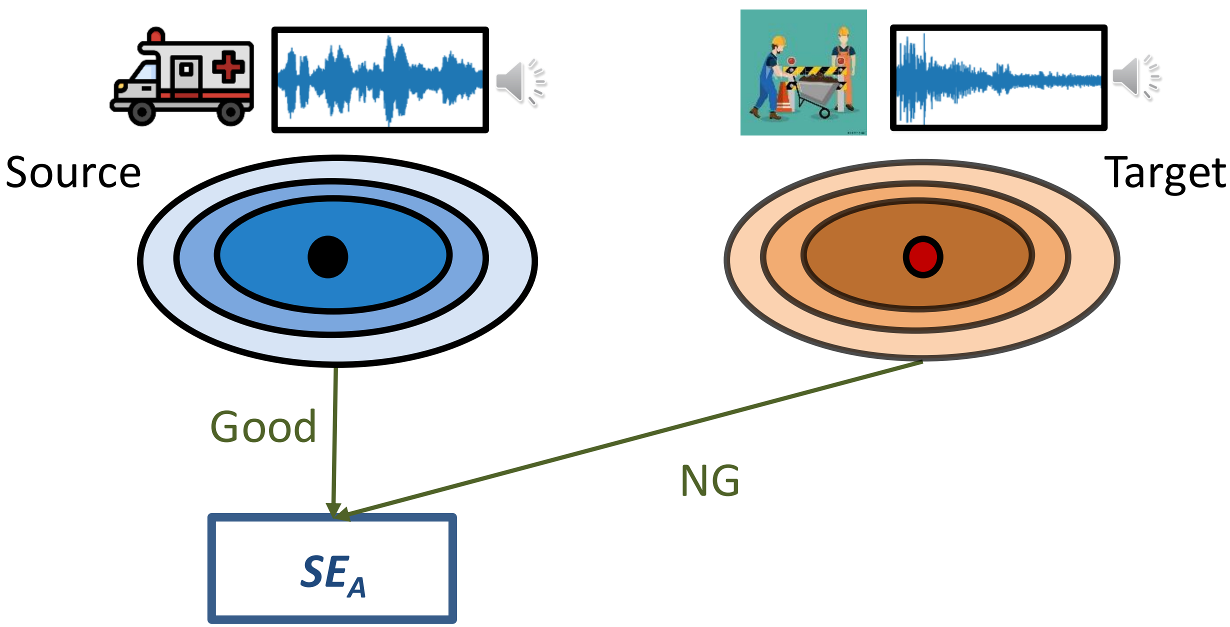


$SE_{1,2,3}$
$SE_{1,2,3}$
Imperfect results
$SE_{1,2,3}$
Model adaptation
(1) supervised: w/ paired noisy/clean
(2) unsupervised: wo/ paired noisy/clean
SE_4
SE_4

Model Adaptation

- SE using Regularized Incremental Learning (SERIL) [Lee et al., Interspeech 2020]

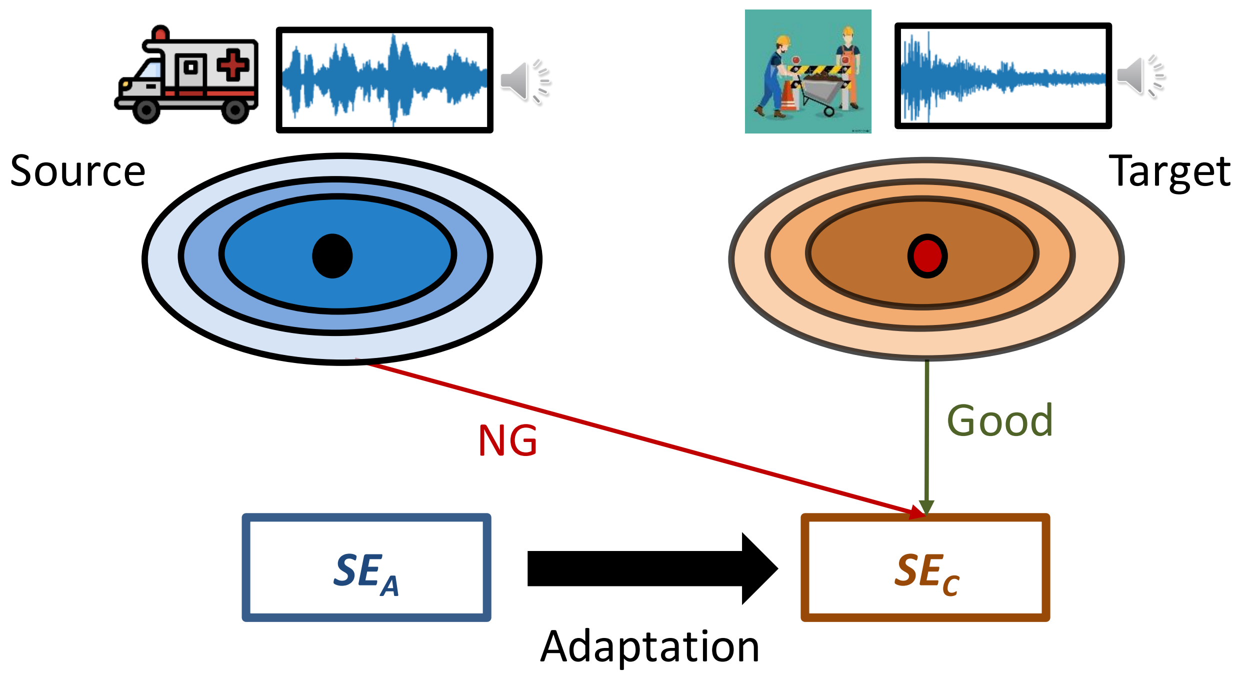
➢ For supervised model adaptation:



Noise/speaker mismatch may cause poor SE performance.

Model Adaptation

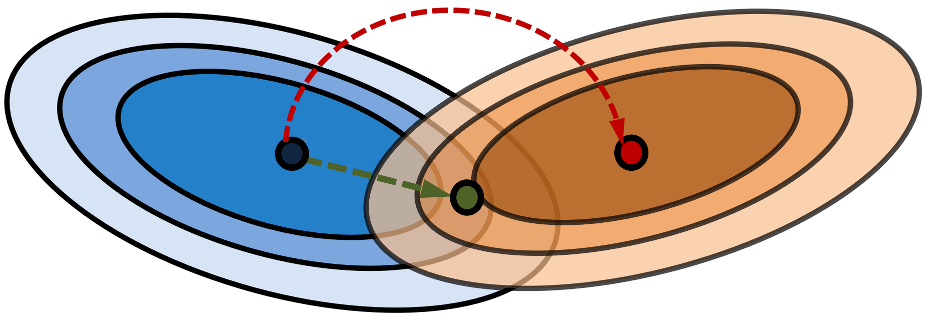
- SE using Regularized Incremental Learning (SERIL) [Lee et al., Interspeech 2020]
  - For supervised model adaptation:



(1) A direct adaptation may cause a catastrophic forgetting issue.
(2) The SERIL approach is proposed for SE adaptation.

Model Adaptation

- SERIL [Lee et al., Interspeech 2020]



Rather then direct adaptation, SERIL adopts proper constraints.

$$L(\theta) = L_{old}(\theta) + L_{new}(\theta)$$

Not available

Constraints

From target data

Solution 1 Curvature strategy [Kirkpatrick et al., PNAS 2017, Schwarz et al., ICML 2018]

Solution 2: Path optimization approach [Zenke et al., ICML2017]

SERIL uses a combined approach [Chaudhry et al., 2018]

Model Adaptation

- SERIL [Lee et al., Interspeech 2020]

Original: training set

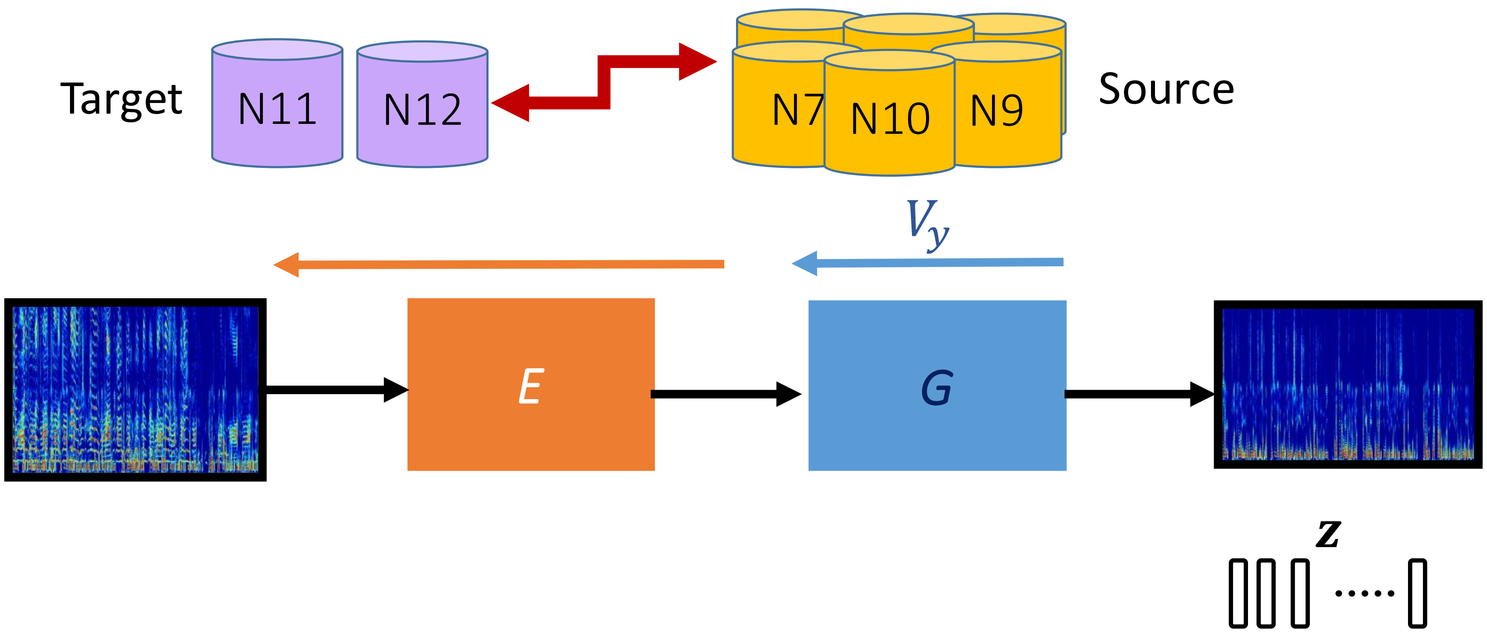
N: Unprocessed
P: Original Model.
F: Direct adaptation
R: SERIL

Metric	M	original	cough	door moving	foot-steps	clap
PESQ	N	2.266	2.041	1.864	1.868	1.474
	P	**2.708**	2.118	2.059	2.015	1.603
	F	2.406	2.204	2.339	2.133	**2.948**
	R	2.461	**2.375**	**2.581**	**2.381**	2.936
STOI	N	0.816	0.788	0.743	0.778	0.789
	P	**0.869**	0.798	0.779	0.799	0.801
	F	0.811	0.816	0.825	0.829	0.923
	R	0.826	**0.839**	**0.859**	**0.855**	**0.931**

(1) Original model achieves the best in the original testing set.
(2) Direct adaptation suffers from the catastrophic forgetting issue.
(3) SERIL consistently improves permeance for all noise types.

Model Adaptation

- Noise-adaptive DAT (NADAT) [Liao et al., Interspeech 2019]
  - For unsupervised model adaptation:



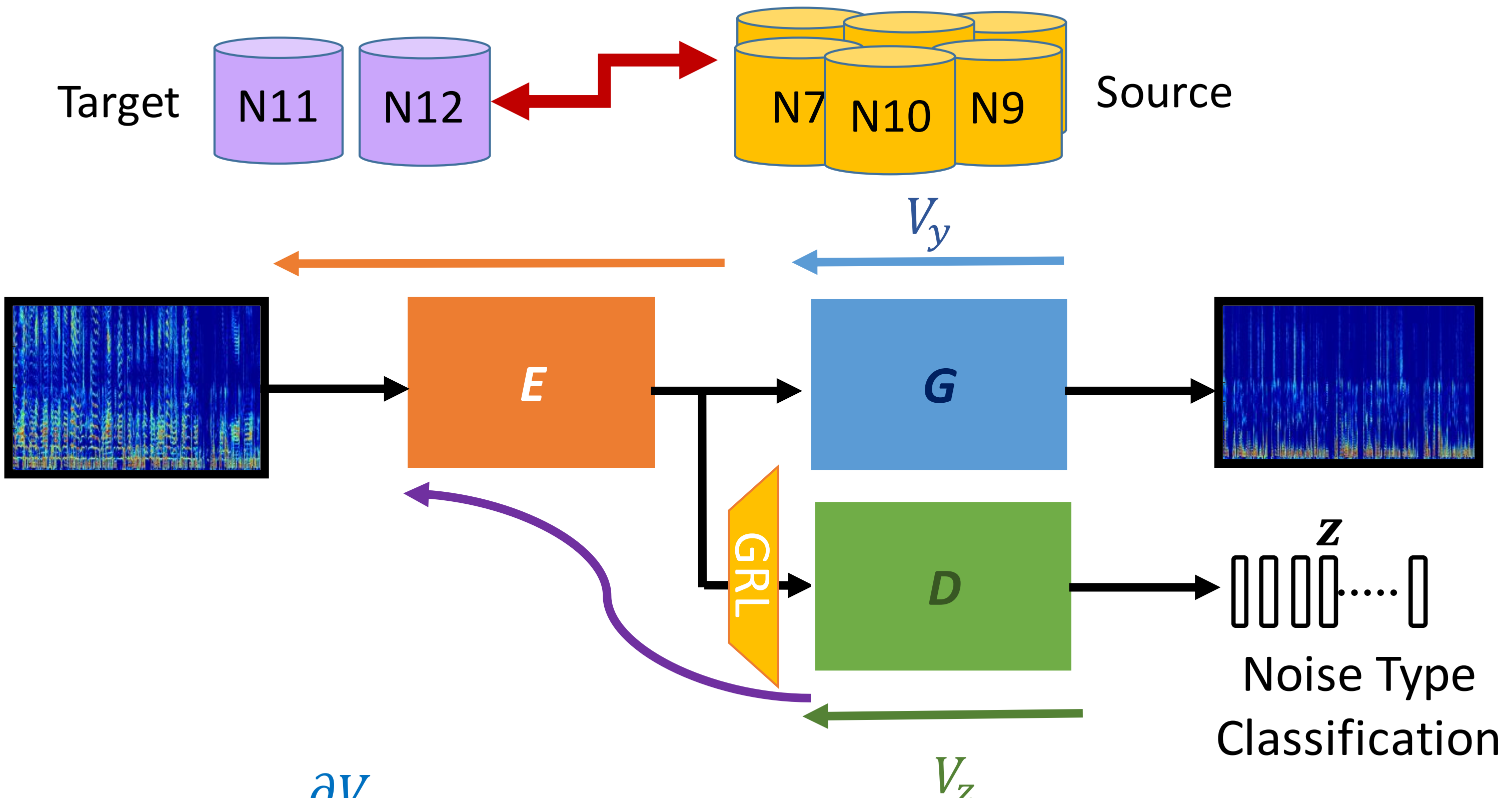
$$\theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G}$$ Min reconstruction error

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

Min reconstruction error

Model Adaptation

- NADAT [Liao et al., Interspeech 2019]



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

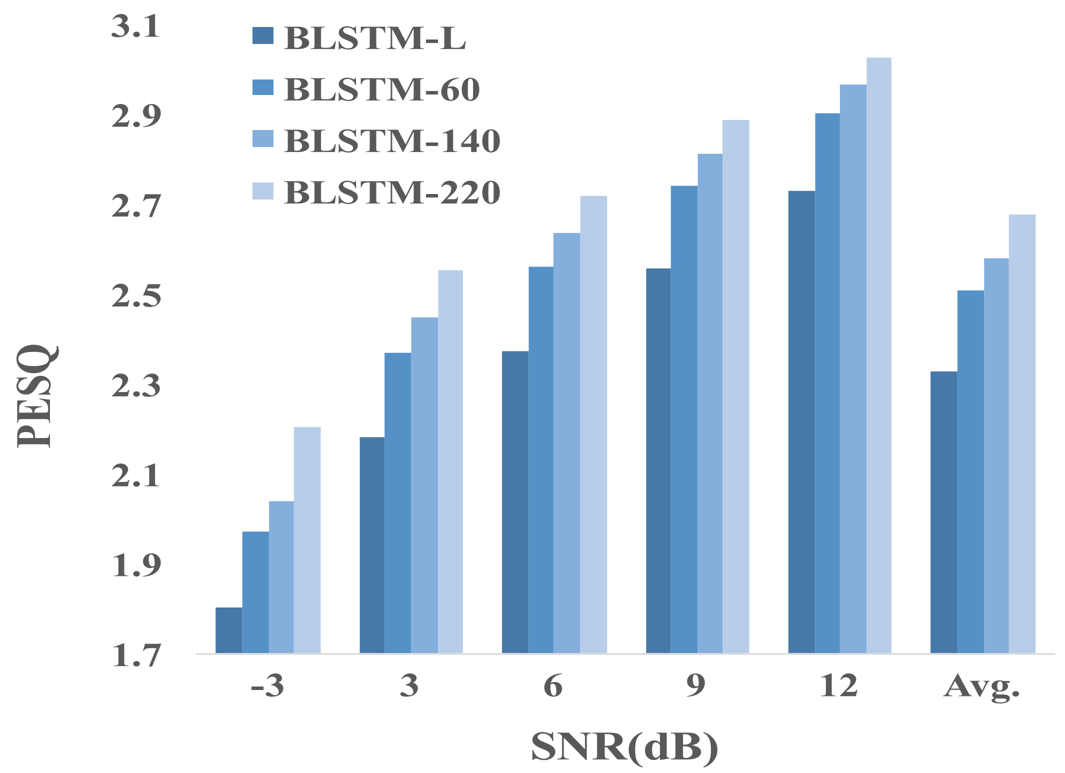
$\theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D}$ Max domain accuracy

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

Min reconstruction error and Min domain accuracy

Model Adaptation

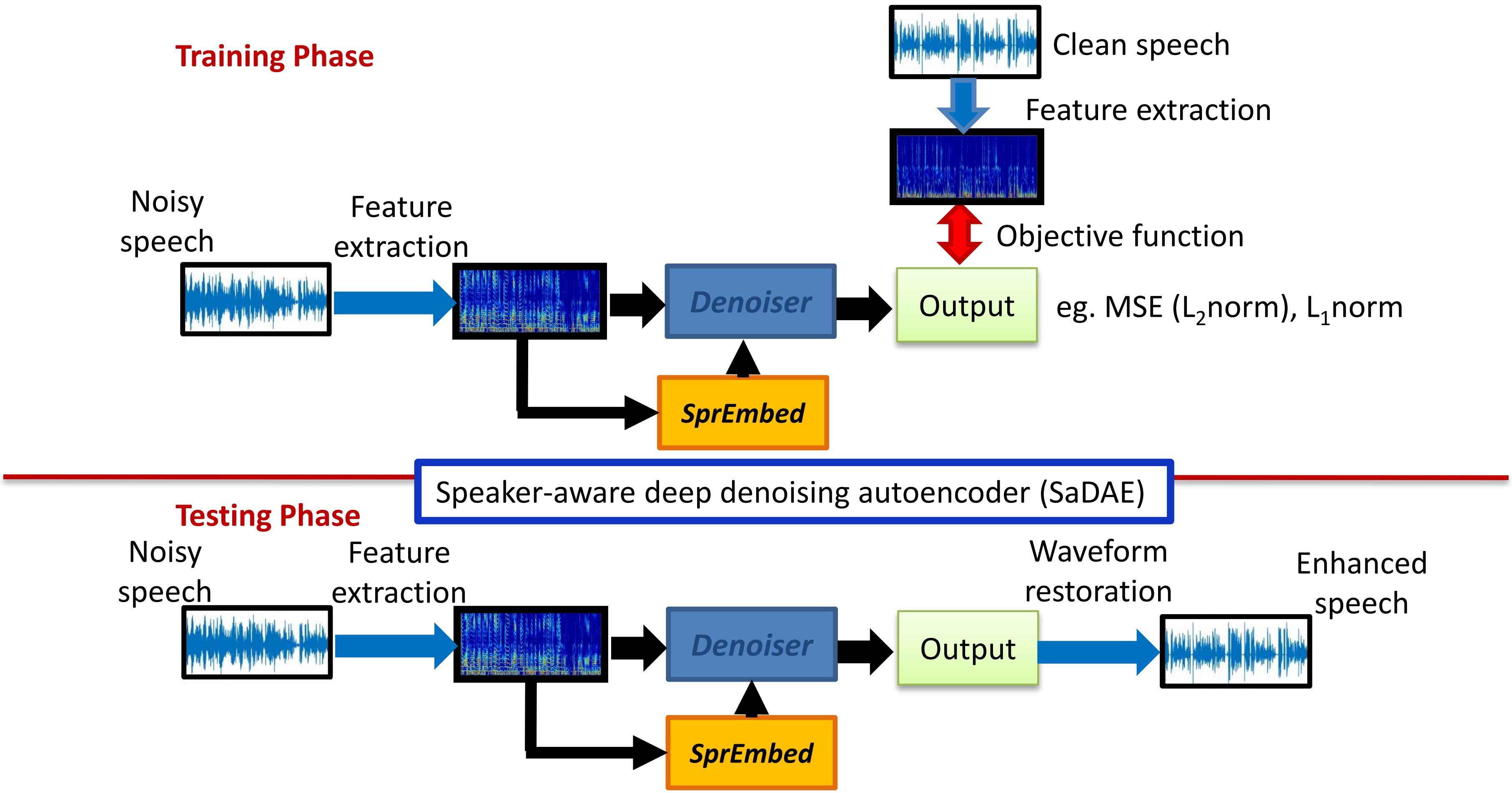
- Adapting to new noise type (Baby cry)



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

Speaker Adaptability

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



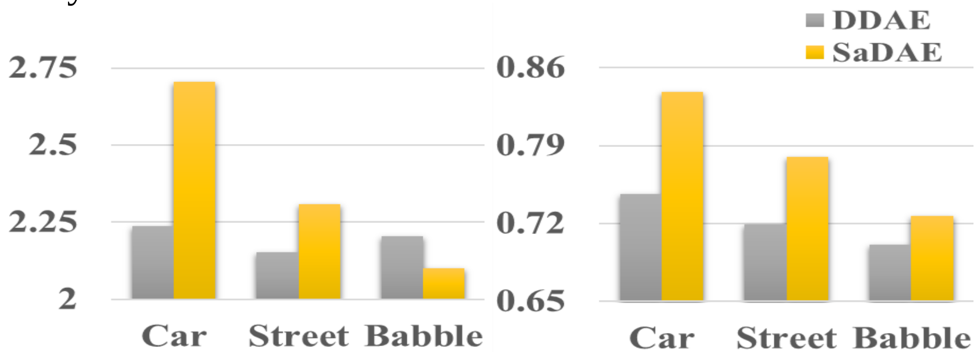
Speaker Adaptability (SaDAE)

- 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.



(a)PESQ (b)STOI

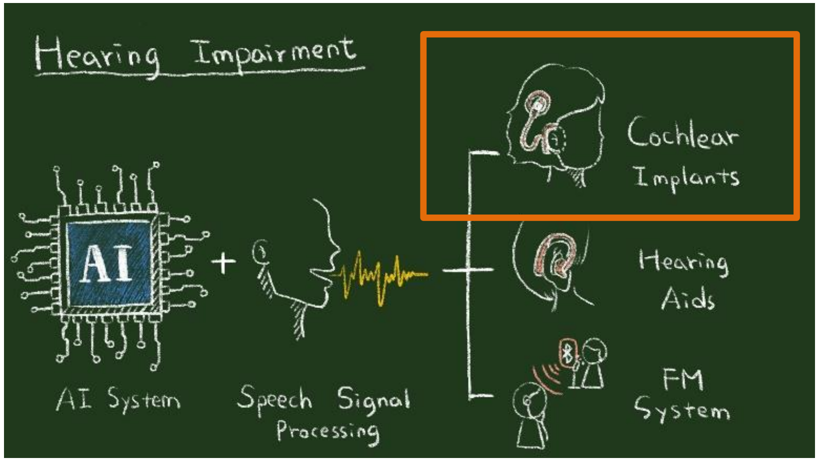
SaDAE outperforms conventional DDAE for both PESQ and STOI.

Outline

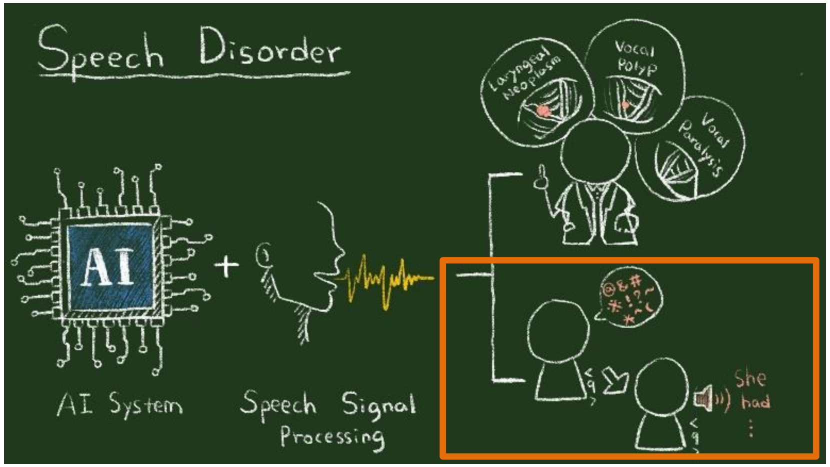
- Deep Learning based Speech Enhancement
  - System architecture
  - Five factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- **Assistive Voice Communication Technologies**

Assistive Voice Communication

- Assistive listening



- Assistive speaking



Cochlear Implant

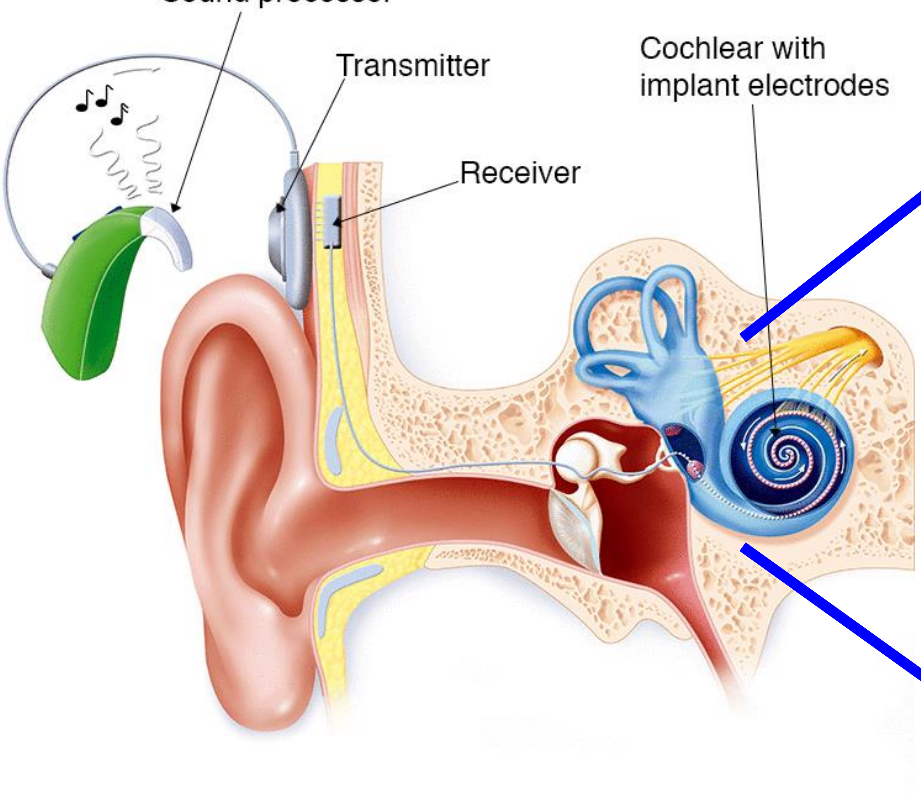
Transmitter coil

Speech processor:

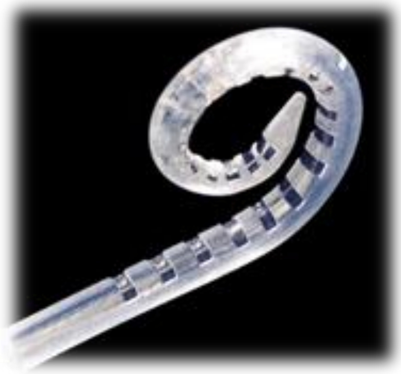
1. Microphone.
2. DSP chip.
3. Battery
4. Others…

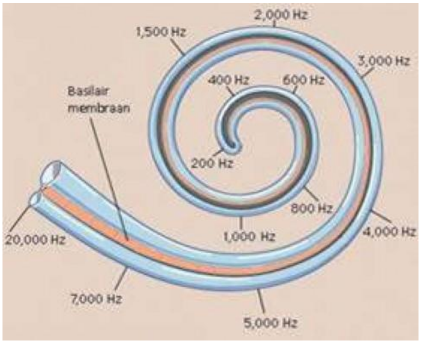
Source from:
https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cochlear-implant-surgery

Cochlear Implant



Electrodes





Traveling wave theory (Nobel Prize 1961)

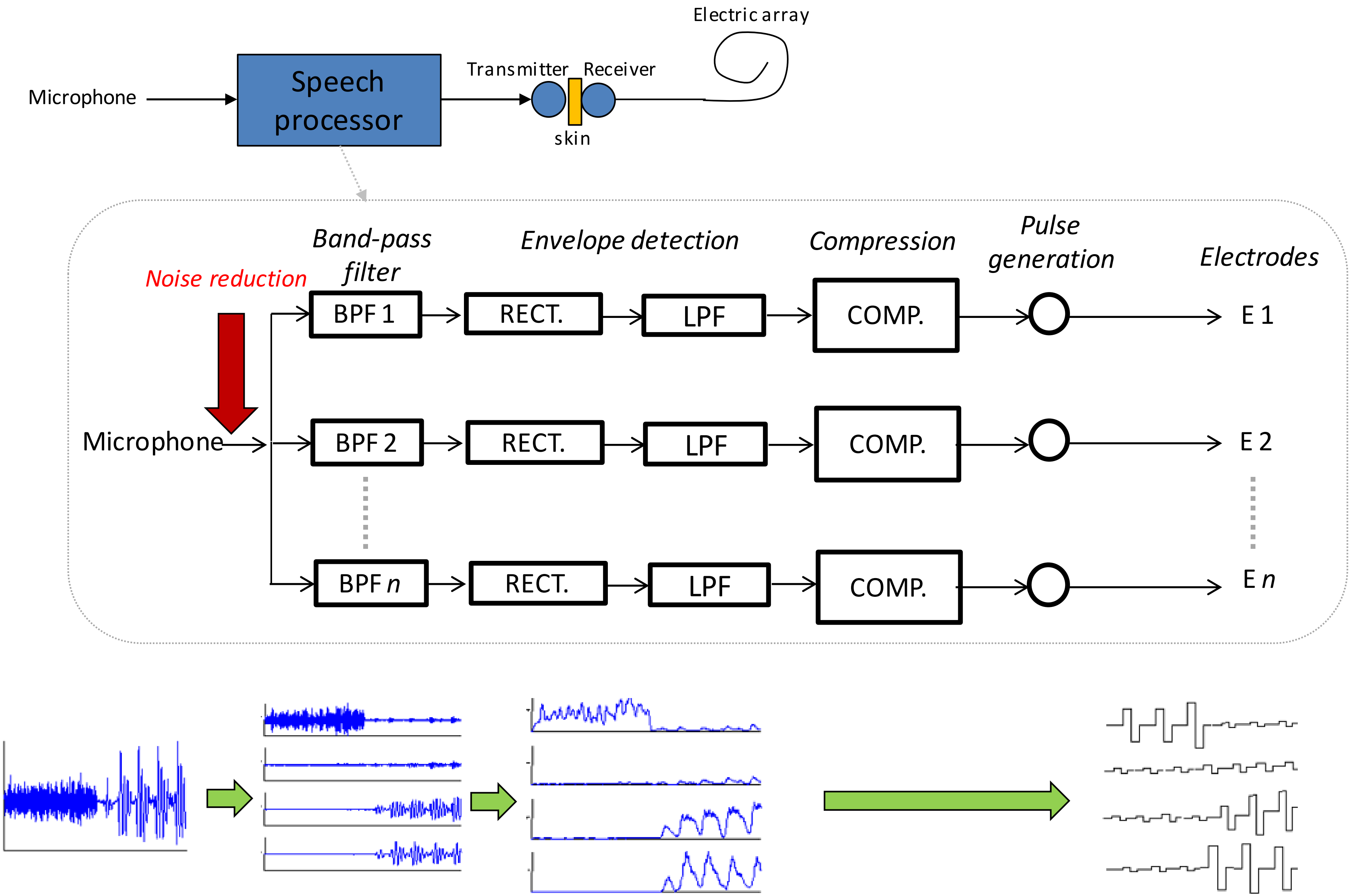
Source from:
https://www.healthdirect.gov.au/cochlear-implant
http://www.yanthia.com/online/projlets/spear3/index.html
https://medium.com/@mosaicofminds/maps-in-the-brain-f236998d544f

SE for Cochlear Implant

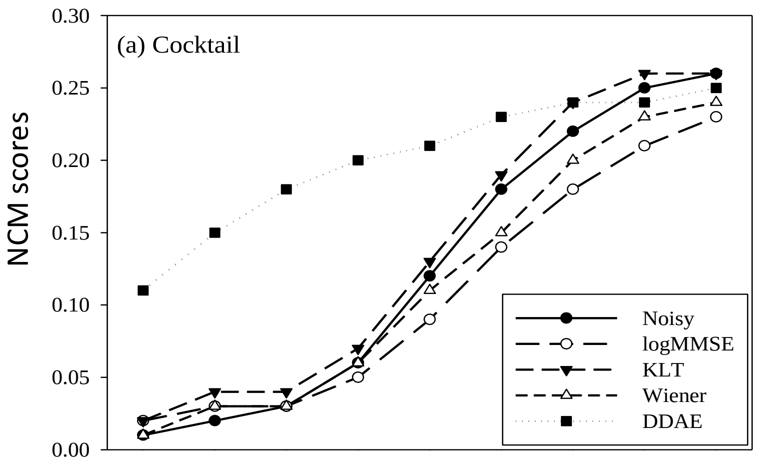
- The tremendous progress of CI technologies in the past three decades has enabled many CI users to enjoy **high level** of speech understanding **in quiet**.
- For most CI users, however, the performance of speech understanding **in noise** still remains **challenging**.
  - F. Chen, Y. Hu, and M. Yuan, "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.



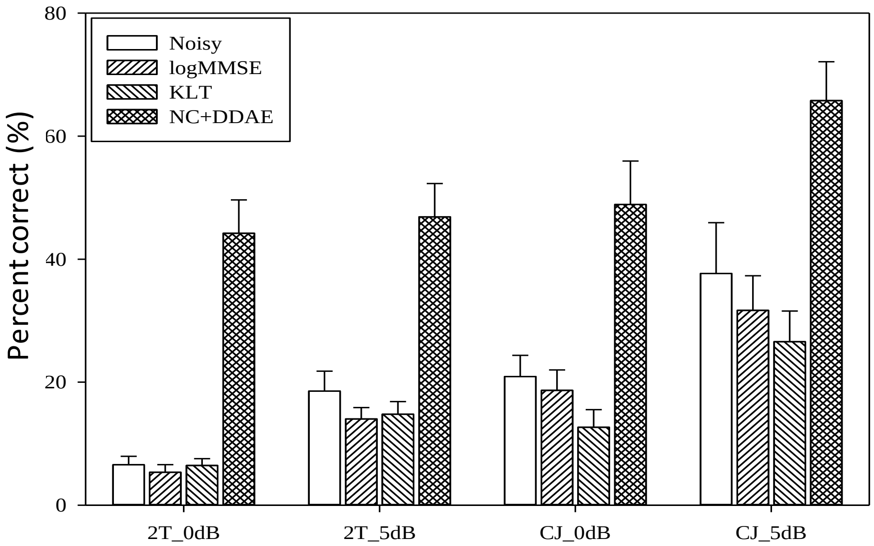
SE for Cochlear Implant



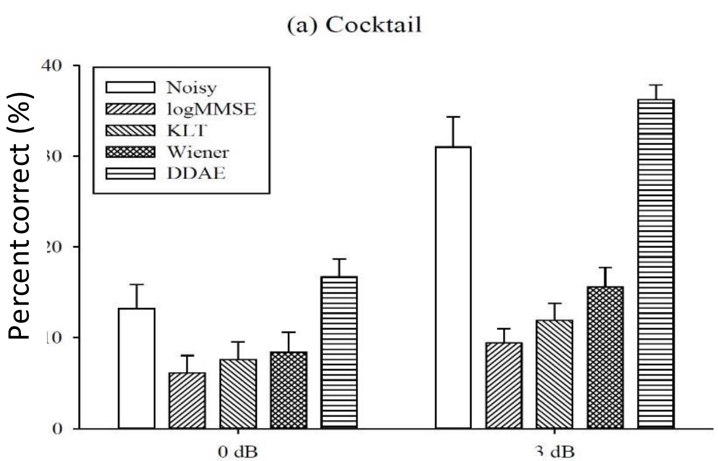
Testing Results



Objective evaluation (NCM)



Clinical trial: 9 CI subjects.

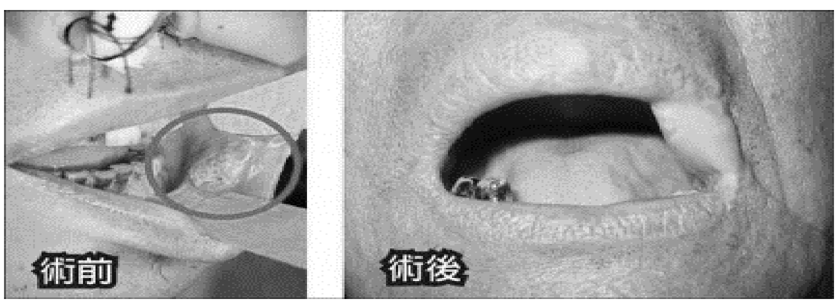


Vocoder results: 10 normal hearing subjects.

- 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.
- R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, "A Study of Joint Effect on Denoising Techniques and Visual Cues to Improve Speech Intelligibility in Cochlear Implant Simulation," to appear in IEEE Transactions on Cognitive and Developmental Systems.

SE for Speaking Disorder

- **Task:** improving the speech intelligibility of surgical patients.
- **Target:** oral cancer (top five cancer for male in Taiwan).



Before After

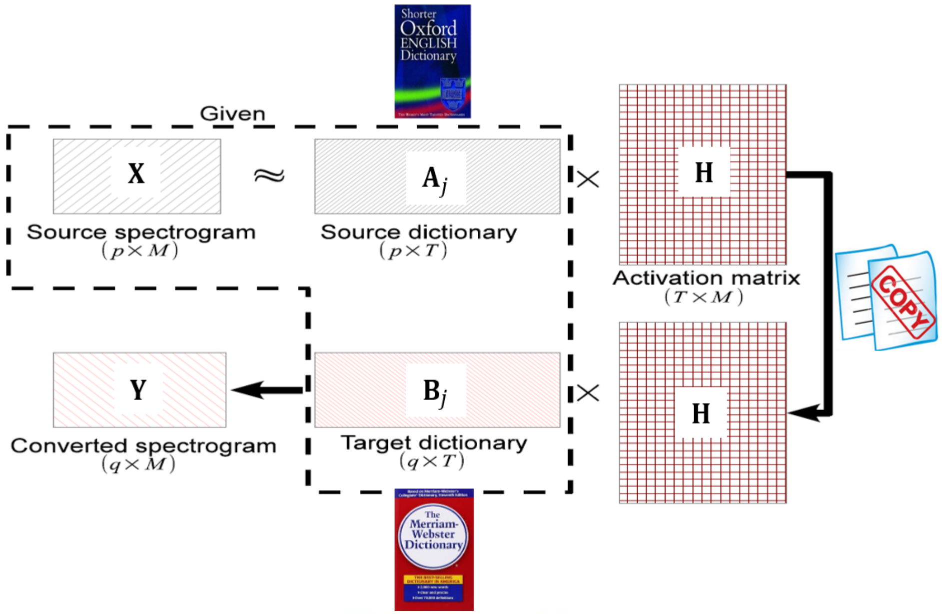
Liberty Times Ltd..

Before After

Taipei Veterans General Hospital

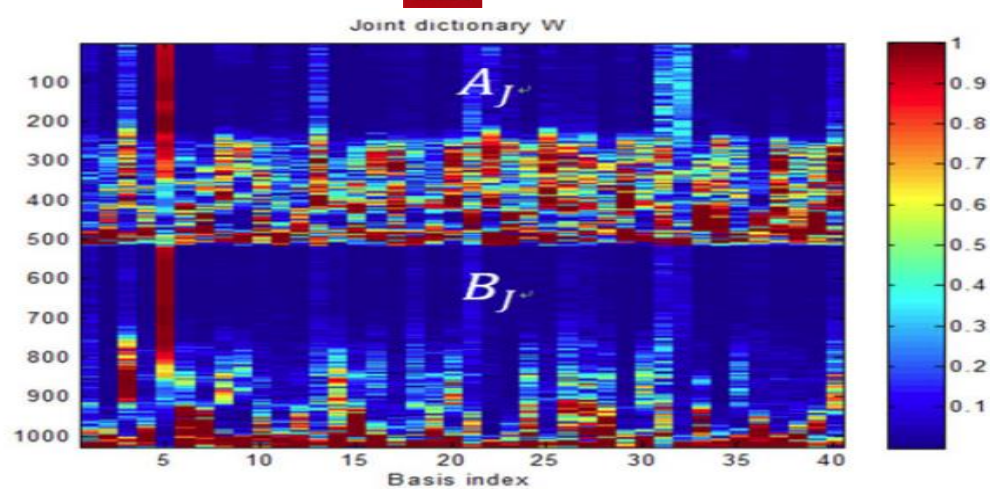
SE for Speaking Disorder

- Proposed: joint training of source and target dictionaries with non-negative matrix factorization (NMF):



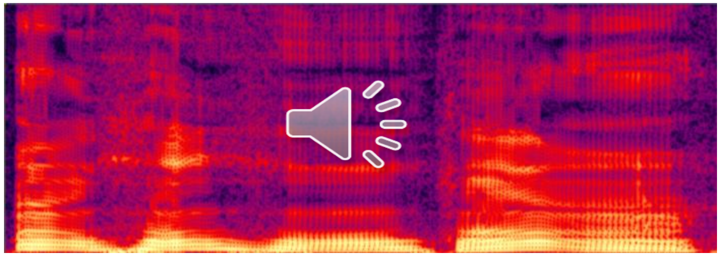
After Surgery

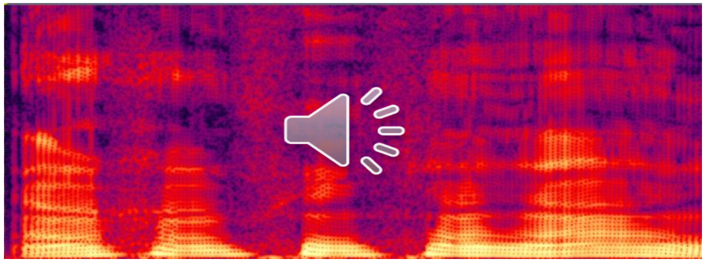
Before Surgery



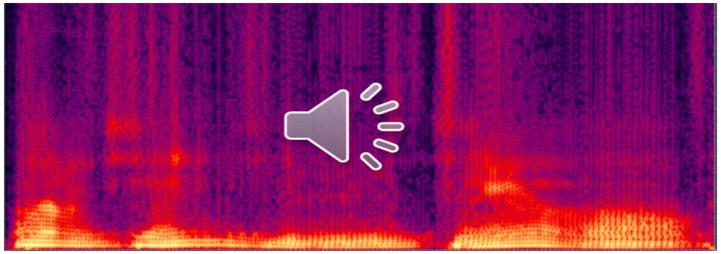
Testing Results

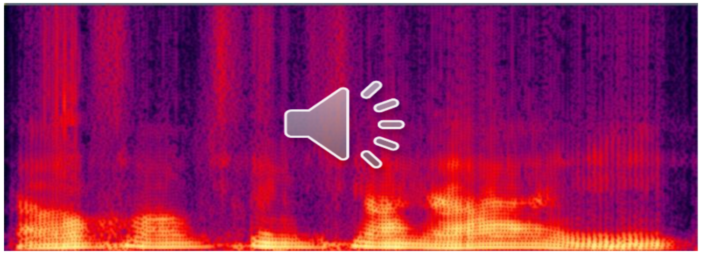
Original:





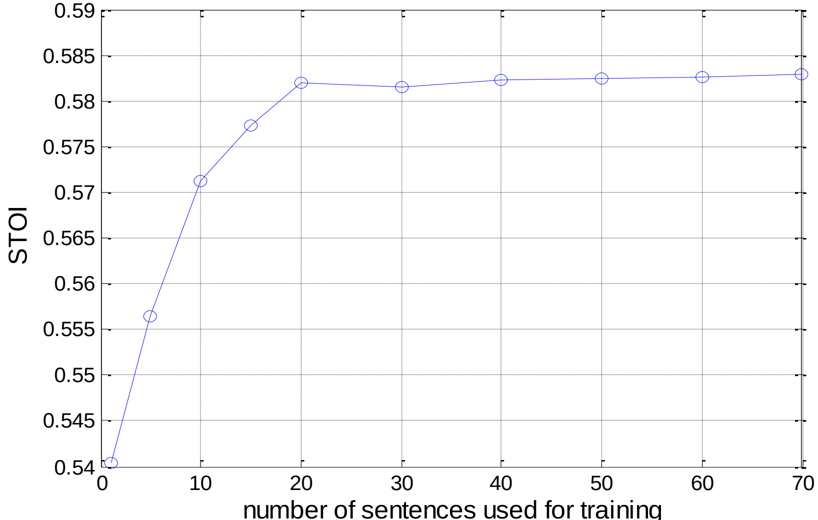
After Conversion:





衛生紙給我

遙控器在哪裡



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

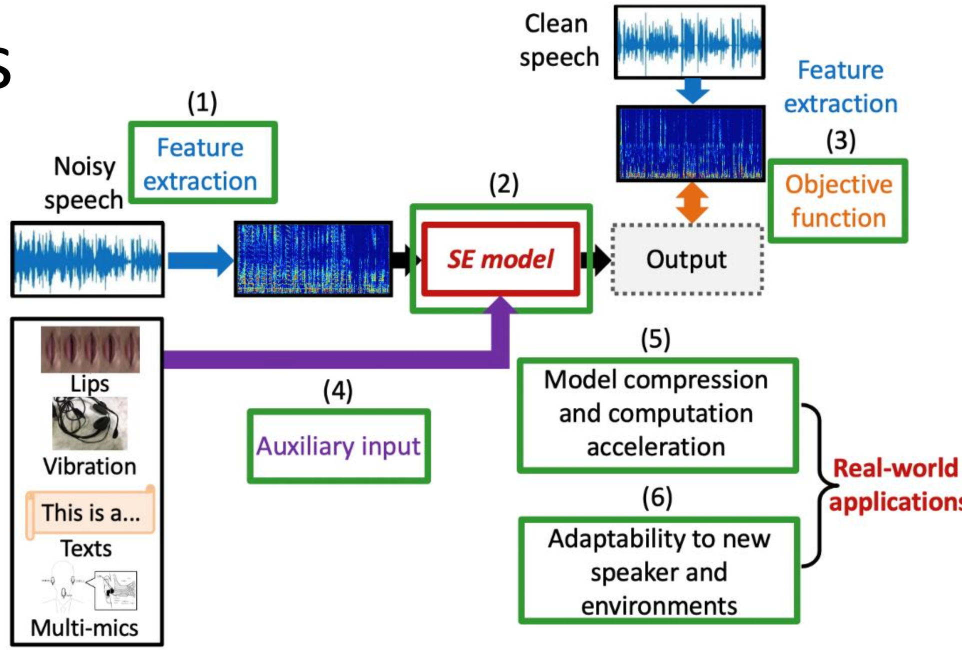
GAN-based solution [Chen et. al., Interspeech 2019]

Outline

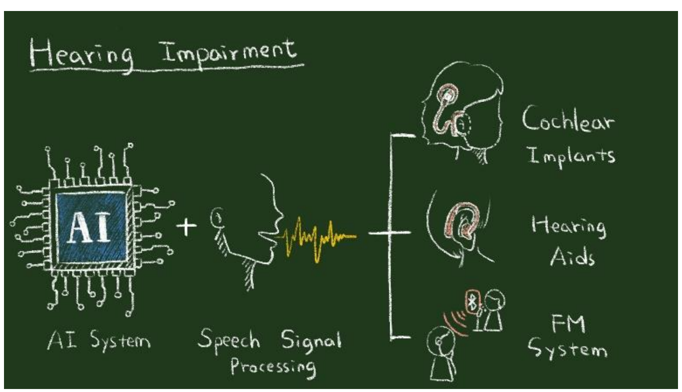
- Deep Learning based Speech Enhancement
  - System architecture
  - Six factors need to consider
    - ✓ Feature types
    - ✓ Model types
    - ✓ Objective function
    - ✓ Auxiliary input
    - ✓ Model compression
    - ✓ Increasing adaptability
- Assistive Voice Communication Technologies
- **Summary**

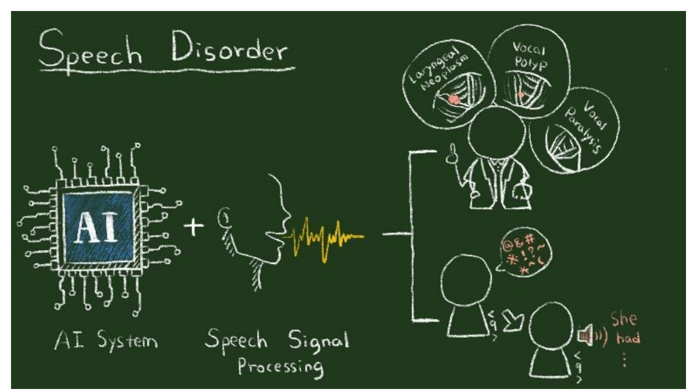
Summary

- Six Factors



- Assistive Voice Communication Technologies





Other Related Works

- Unpaired Speech Enhancement
  - Adversarial training [Mimura et al., ASRU 2017, Meng et al., Interpseech 2018, Xiang and Bao, TASLP 2020]
  - Variational autoencoder [Sadeghi et al, TASLP2020]
  - Noisy2Noisy [Alamdari et al., AC 2020]
  - Self-supervised [Zezario et al., ICASSP 2020]
- Post-filtering
- Other Modalities
- Meta-learning
- Mask-based Speech Enhancement

Resources

[1] https://bio-asplab.citi.sinica.edu.tw/Opensource.html#SE (Codes+Papers, from BioASP Lab)

[2] https://bio-asplab.citi.sinica.edu.tw/Opensource.html#Dataset (Dataset, from BioASP Lab)

[3] https://github.com/nanahou/Awesome-Speech-Enhancement (Codes+Papers)

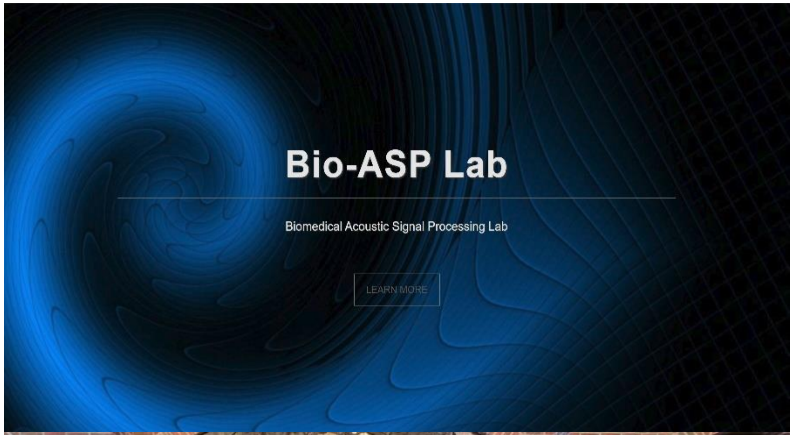
[4] https://paperswithcode.com/task/speech-enhancement (Codes+Papers)

[5] https://github.com/mpariente/asteroid (Codes+Papers)

CITISEN: A Deep Learning-Based Speech Signal-Processing Mobile Application

GitHub: https://github.com/yuwchen/CITISEN
Paper: https://arxiv.org/pdf/2008.09264.pdf
Youtube:
https://www.youtube.com/watch?v=BUfY64TCXi4&feature=youtu.be&fbclid=IwAR0snLN2wBLi5aU8xTdtPJsU5z2ujvt3ow6jHMtTbKldJsBwoaNsAGoCKUM

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 C.-H. 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.
- Pandey, Ashutosh, and DeLiang Wang. "A new framework for CNN-based speech enhancement in the time domain." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.7 (2019): 1179-1188.

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.
- J. Kim, M. El-Khamy, and J. Lee, T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement. ICASSP 2020.
- S.-W. Fu, et al, Boosting Objective Scores of Speech Enhancement Model through MetricGAN Post-Processing, APSIPA 2020.
- M. Kim, “Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders,” in Proc. ICASSP, 2017.
- S. E. Chazan, J. Goldberger, and S. Gannot, “Deep recurrent mixture of 1experts for speech enhancement,” in Proc. WASPAA, 2017.
- X.-L. Zhang and D. Wang, A deep ensemble learning method for monaural speech separation, IEEE/ACM Trans. Audio, Speech Lang. Process., 1089 vol. 24, no. 5, pp. 967–977, May 2016.
- Z. Meng, J. Li, and Y. Gong., Adversarial feature-mapping for speech enhancement. arXiv preprint arXiv:1809.02251, 2018.
- Z. Meng, J. Li, and Y. Gong., Cycle-consistent speech enhancement., arXiv preprint arXiv:1809.02253, 2018.
- Kolbæk, Morten, Zheng-Hua Tan, and Jesper Jensen. "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2016): 153-167.
- M. Kolbæk, et al., On loss functions for supervised monaural time-domain speech enhancement, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 825-838.

References

- Y.-H. Tu, J. Du, and C.-H. Lee. Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.12, 2019: 2080-2091.
- Z.-Q. Wang, P. Wang, and D. Wang. Complex Spectral Mapping for Single-and Multi-Channel Speech Enhancement and Robust ASR. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1778-1787.
- Y. Hu, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020).
- J.-M. Valin, et al. A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech. arXiv preprint arXiv:2008.04259 (2020).
- Germain, Francois G., Qifeng Chen, and Vladlen Koltun. "Speech denoising with deep feature losses. arXiv preprint arXiv:1806.10522, 2018.
- J. Kim, E.K. Mostafa, and J. Lee. End-to-end multi-task denoising for joint SDR and PESQ optimization. arXiv preprint arXiv:1901.09146, 2019.
- J. Le Roux, et al. SDR–half-baked or well done? ICASSP, 2019.
- Y. Zhao, et al. Perceptually guided speech enhancement using deep neural networks. ICASSP 2018.
- Y. Xia, et al. Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement, ICASSP 2020.
- Y. Koizumi, et al. Speech enhancement using self-adaptation and multi-head self-attention. ICASSP 2020.
- Z. Du, et al. Pan: Phoneme-Aware Network for Monaural Speech Enhancement. ICASSP 2020.
- H. Li and J. Yamagishi. "Noise Tokens: Learning Neural Noise Templates for Environment-Aware Speech Enhancement. arXiv preprint arXiv:2004.04001 (2020).
- C.-F. Liao, et al. Incorporating symbolic sequential modeling for speech enhancement. arXiv:1904.13142, 2019.
- K. Kinoshita, et al. Text-informed speech enhancement with deep neural networks. Interspeech 2015.
- S.-Y. Chuang, et al. Lite Audio-Visual Speech Enhancement. arXiv:2005.11769, 2020.
- Y.-J. Lu, et al. Incorporating broad phonetic information for speech enhancement, arXiv 2020.
- C.-C. Lee, et al., SERIL: Noise Adaptive Speech Enhancement using Regularization-based Incremental Learning. arXiv:2005.11760, 2020.

References

- Y. Luo, and N. Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." IEEE/ACM transactions on audio, speech, and language processing 27.8 (2019): 1256-1266.
- Y. Luo, Z. Chen, and T. Yoshioka. "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation." ICASSP 2020.
- Y. Xiang and C. Bao. A Parallel-data-free Speech Enhancement Method using Multi-Objective Learning Cycle-consistent Generative Adversarial Network. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2020).
- M. Mimura, S. Sakai, and T. Kawahara, Cross-domain Speech Recognition Using Nonparallel Corpora with Cycle-consistent Adversarial Networks, ASRU, 2017.
- M. Sadeghi, et al., Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1788-1800.
- N. Alamdari, A. Arian, and K. Nasser, Improving Deep Speech Denoising by Noisy2Noisy Signal Mapping, Applied Acoustics 2020.
- R. E. Zezario, et al., Self-Supervised Denoising Autoencoder with Linear Regression Decoder for Speech Enhancement, ICASSP 2020.
- W.-C. Lin, et al. Investigation of Neural Network Approaches for Unified Spectral and Prosodic Feature Enhancement, APSIPA 2019.
- M. Tagliasacchi, et al. "SEANet: A Multi-modal Speech Enhancement Network." arXiv:2009.02095, 2020.

Thank You Very Much for Your Attention