13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

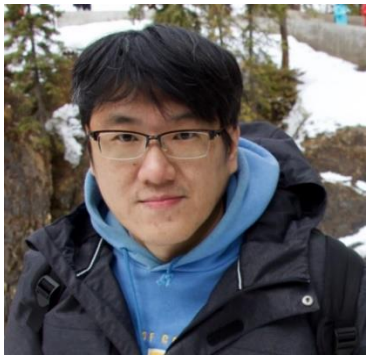
Tutorial T1

Speech Perception and Enhancement in Cochlear Implants (Part II)

Yu Tsao
Research Center for Information Technology Innovation
Academia Sinica
yu.tsao@citi.sinica.edu.tw, https://www.citi.sinica.edu.tw/pages/yu.tsao/

14/12/2021

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



– Education

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

– Work Experience

- Researcher, National Institute of Information and Communications Technology, SLC Group, Japan (2009/4-2011/9)
- Research Fellow (Professor) and Deputy Director Research Center for Information Technology Innovation (2020/9-present)

– Academia Services

- Chair, Speech, Language, and Audio (SLA) Technical Committee, APSIPA
- Distinguished Lecturer, 2019-2020, APSIPA
- Associate Editor of IEEE Signal Processing Letters
- Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing

– Lab at CITI (Academia Sinica)

Research Fellow, Deputy Director of CITI, Academia Sinica

Biomedical Acoustic Signal Processing (Bio-ASP) Lab

– Research Interests

Assisitve Speech Communication Technologies, Audio-coding, Biomedical Signal Processing, and Speech Signal Processing

Outline

1. Background
   - Traditional speech enhancement
   - Deep learning based speech enhancement
   - Goal-oriented speech enhancement
2. Deep learning based speech enhancement in cochlear implants
   - Intelligibility-oriented speech enhancement for CI speech perception
   - Integration of speech enhancement and visual cues
3. Summary

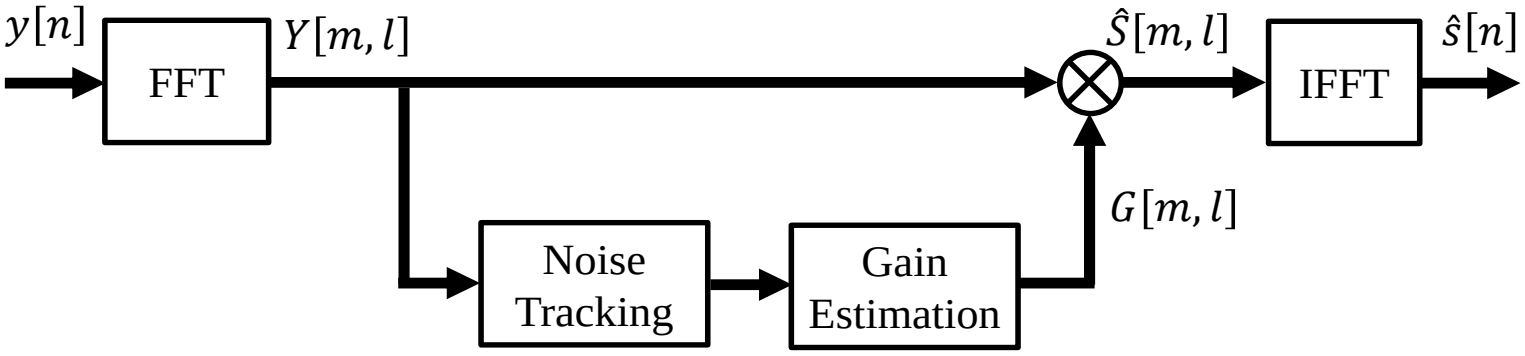
Outline

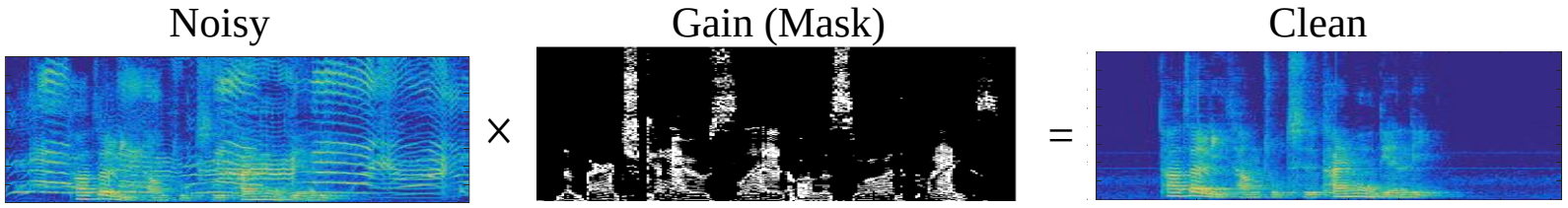
1. Background
   - **Traditional speech enhancement**
   - Deep learning based speech enhancement
   - Goal-oriented speech enhancement
2. Deep learning based speech enhancement in cochlear implants
   - Intelligibility-oriented speech enhancement for CI speech perception
   - Integration of speech enhancement and visual cues
3. Summary

Speech Enhancement

- What is speech enhancement (SE)?
- Regression tasks: Noise reduction, speech denoising, speech separation, speech dereverberation, bandwidth expansion, speaker extraction.
- SE aims to convert the input speech to the enhanced one with improved quality, intelligibility, automatic speech recognition accuracies (ASR).
- SE refers to any particular regression task but requires a definite **goal (objective)**.

Traditional SE Systems



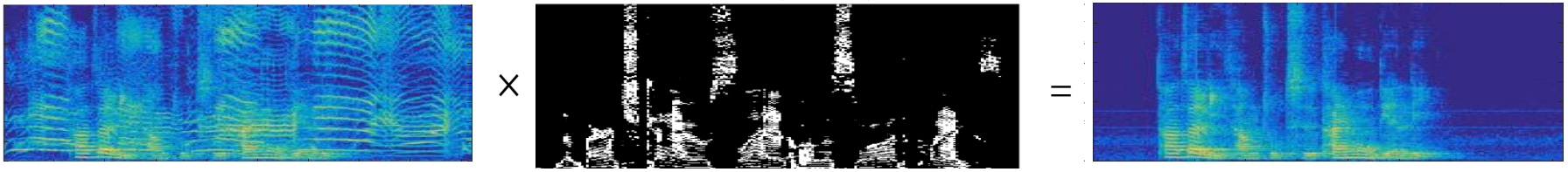


Noise tracking (estimation):
1. VAD, MCRA
2. Robust PCA
3. NMF
4. Deep learning
etc.

Gain (mask) estimation:
1. Wiener filter
2. MMSE
3. MAPA
4. MLSA
etc.

Traditional SE Systems (Gain Estimation)

Noisy × Gain (Mask) = Clean



1. Assuming speech and noise are uncorrelated
2. Time domain: $y[n] = s[n] + v[n]$
3. Freq. domain: $Y[m,l] = S[m,l] + V[m,l]$

4. Estimating a gain function: $\hat{S}_k = G_k\, Y_k$
5. Adopting the noisy phase $\hat{S} = \hat{S}_k \exp(j\theta_Y)$
6. Transforming spectral features to waveforms

Assume clean speech & noise spectra are Rayleigh and Gaussian distributions, we have

$$p(Y|X_k,\theta_X) = \frac{1}{\pi\sigma_v^2}\exp\left(-\frac{|V|^2}{\sigma_v^2}\right), p(S_k) = \frac{2S_k}{\sigma_s^2}\exp\left(-\frac{X_k^2}{\sigma_X^2}\right)$$

Assume amplitude and phase are independent,

$$p(X_k,\theta_X) = p(X_k)\cdot p(\theta_X)$$

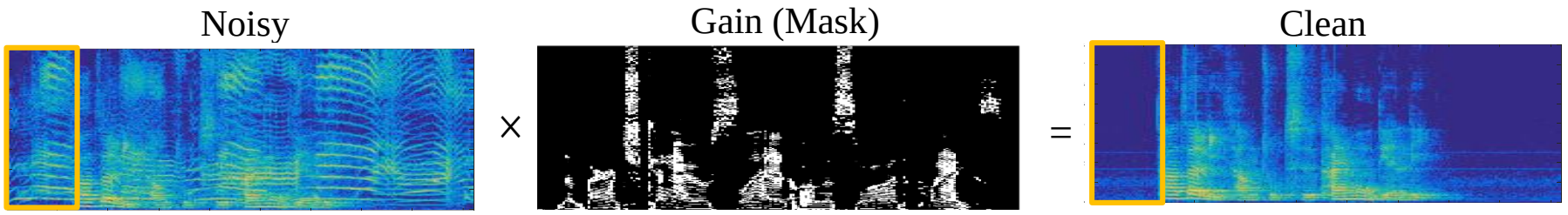
Assume $p(\theta_S)$ is a uniform distribution

$$p(\theta_S) = \frac{1}{2\pi}$$

Wiener	$\frac{\xi}{1+\xi}$
MMSE	$\Gamma\left(\frac{3}{2}\right)\frac{\sqrt{\delta}}{\gamma} exp\left(-\frac{\delta}{2}\right)\left[(1+\delta)I_0\left(\frac{\delta}{2}\right)+\delta I_1\left(\frac{\delta}{2}\right)\right]$
MAPA	$(\xi+\sqrt{\xi^2+(1+\xi)\xi/\gamma})\,/2(1+\xi)$

where *a priori* SNR ξ_k ($\xi_k = \sigma_s^2/\sigma_v^2$) and *a posteriori* SNR γ_k ($\gamma_k = Y_k^2/\sigma_v^2$),
where $\sigma_s^2 = E[|S|^2]$ and $\sigma_v^2 = E[|V|^2]$.

Traditional SE Systems (Noise Estimation/Tracking)



VAD (voice activity detection)

→ Detecting the non-speech part to obtain the noise components.

→ Generally based on pitch or energy information.

→ Colleting the first few frames to compute the noise components.

MCRA (minima controlled recursive averaging)

→ when speech is absent:

$$H_0[m, l+1]: \sigma_v^2[m, l+1] = \lambda\sigma_v^2[m, l] + (1-\lambda)|Y[m, l]|^2$$

→ when speech is present:

$$H_1[m, l+1]: \ \sigma_v^2[m, l+1] = \sigma_v^2[m, l] \ \text{ where } \ \sigma_v^2[m, l] = E[|V[m, l]|^2]$$

→ A detector determines whether each element a speech or nonspeech:

$$Y_r[m, l] \underset{H_0}{\overset{H_1}{\gtrless}} \delta$$

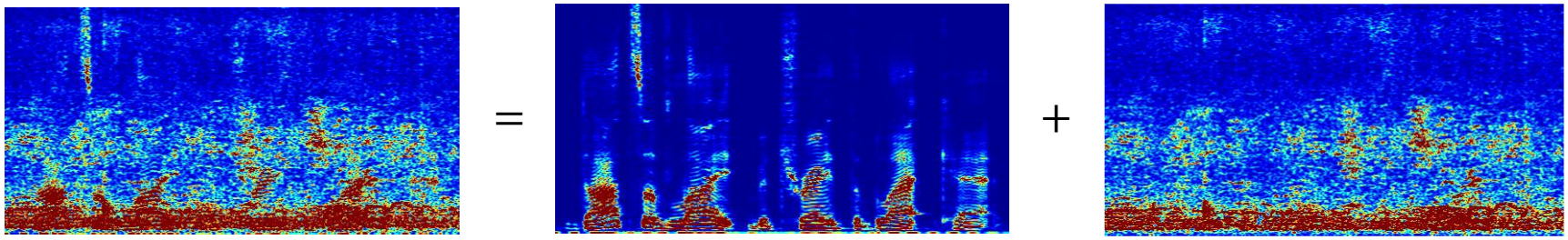
to prevent the noise information updated in a wrong way

→ Extensions: MCRA2, IMCA, EMCRA, OMLSA-MCRA

Robust Principal Component Analysis (RPCA) based SE

To separate speech and noise:

1. Different properties of speech and noise: **Robust Principal Component Analysis (RPCA)**
2. Prior knowledge about speech and noise: **Non-negative Matrix Factorization (NMF)**



Noisy speech (**Y**) = Speech (**S**) *Sparse* + Noise (**V**) *Low rank*

$$\min \{\|\mathbf{V}\|_* + C\|\mathbf{S}\|_1\} \quad \text{subject to} \quad \mathbf{Y} - \mathbf{S} - \mathbf{V} = \mathbf{0}$$

$$L(\mathbf{S}, \mathbf{V}, \mathbf{F}, \mu) = \|\mathbf{V}\|_* + C\|\mathbf{S}\|_1 + \langle \mathbf{F}, \mathbf{Y} - \mathbf{S} - \mathbf{V} \rangle + \frac{\mu}{2}\|\mathbf{Y} - \mathbf{S} - \mathbf{V}\|_{\mathbb{F}}^2$$

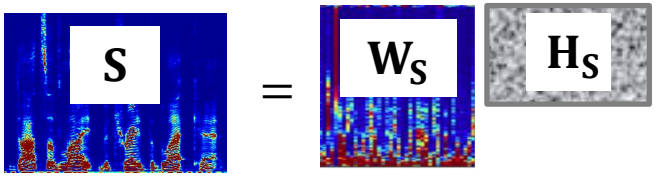
- where $\|\cdot\|_*$, $\|\cdot\|_1$, $\|\cdot\|_{\mathbb{F}}^2$ and $\langle \cdot, \cdot \rangle$, respectively, are the nuclear norm, L1 norm, Frobenius norm and inner product operators.
- **F** is an auxiliary matrix and C and μ are scalars.

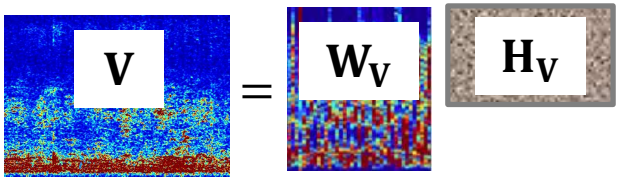
Non-negative Matrix Factorization (NMF) based SE

To separate speech and noise:

1. Different properties of speech and noise: Robust Principal Component Analysis (RPCA)
2. Prior knowledge about speech and noise: **Non-negative Matrix Factorization (NMF)**

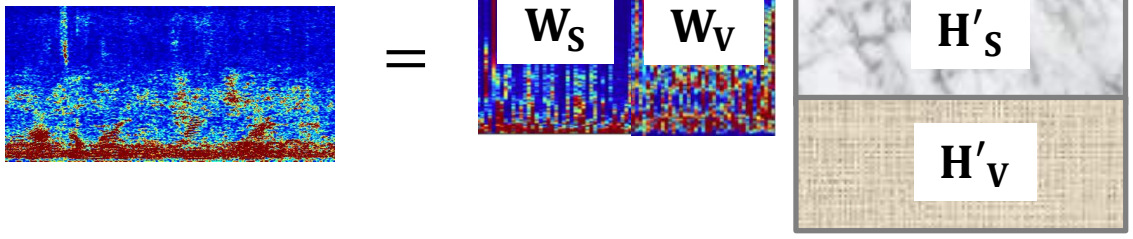
Training phase





Testing phase

Noisy speech (**Y**)



Restored speech ($\hat{\mathbf{S}}$)

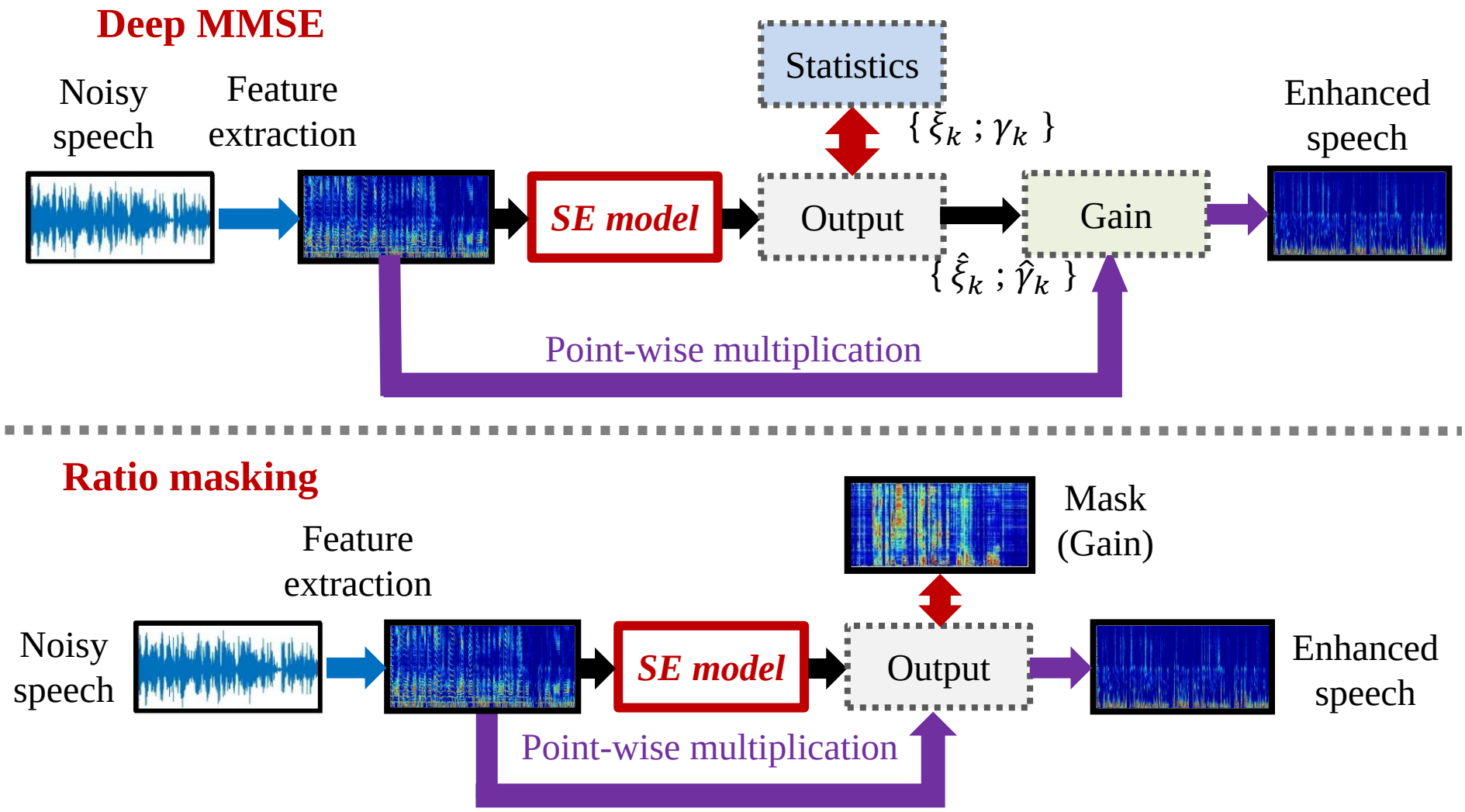
Estimated noise ($\hat{\mathbf{V}}$)

Outline

1. Background
   - Traditional speech enhancement
   - **Deep learning based speech enhancement**
   - Goal-oriented speech enhancement
2. Deep learning based speech enhancement in cochlear implants
   - Intelligent-oriented speech enhancement for CI speech perception
   - Integration of speech enhancement and visual cues
3. Summary

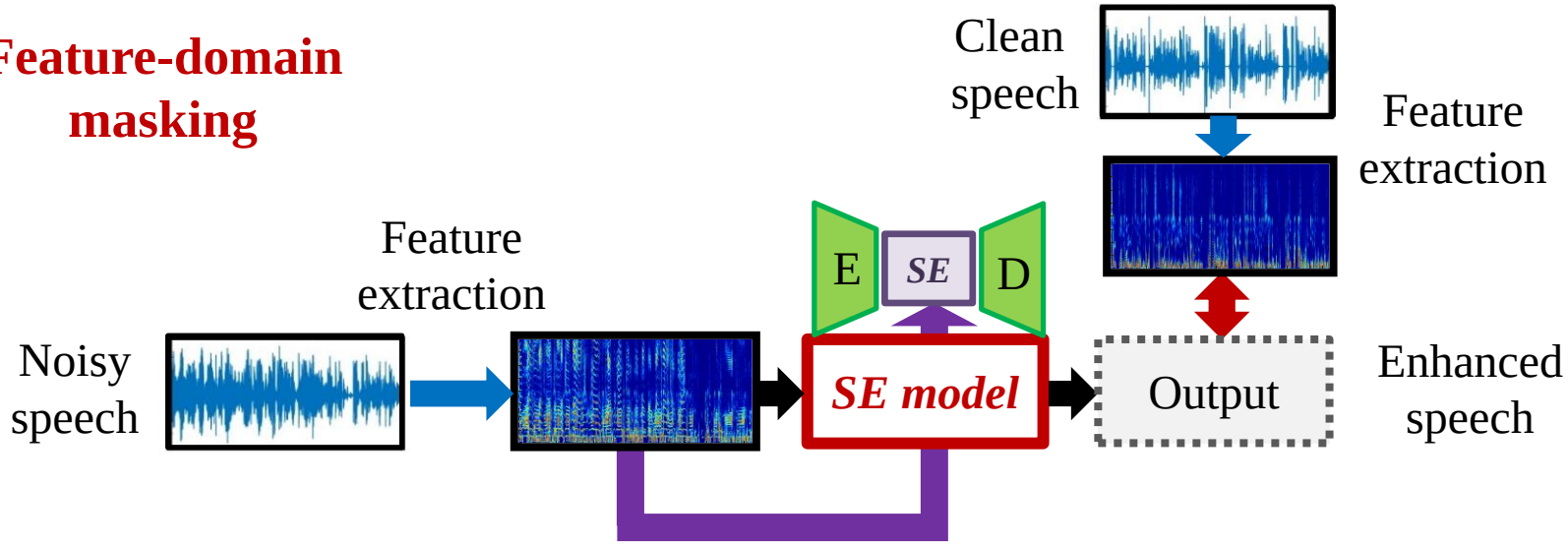
Deep Learning (DL)-based SE

- Mapping vs. masking based SE:

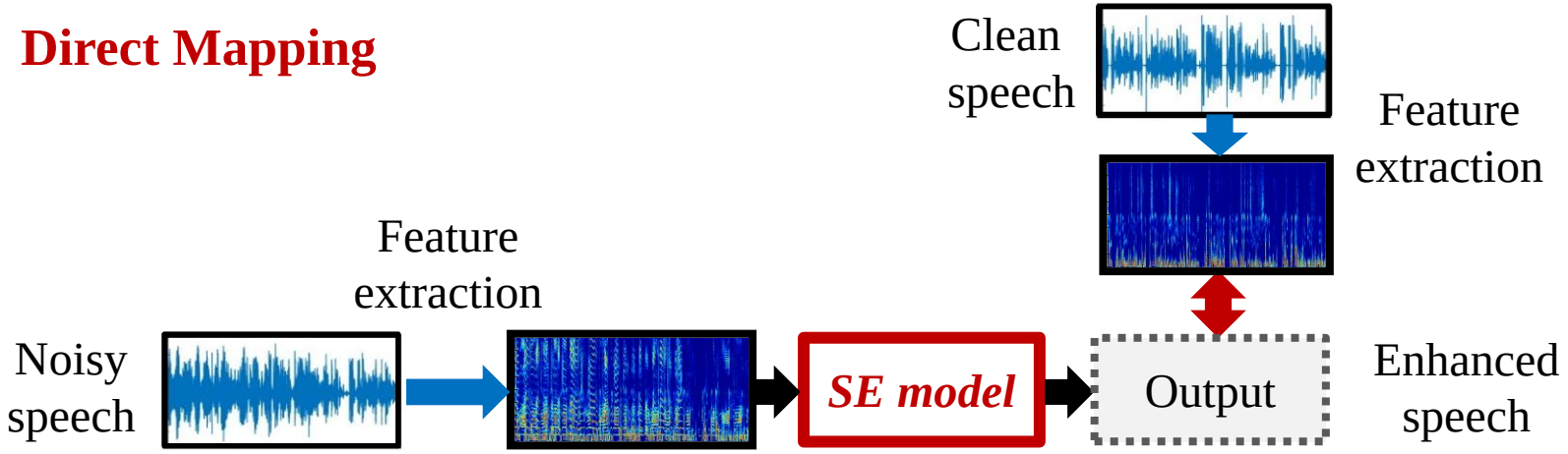


Deep Learning (DL)-based SE

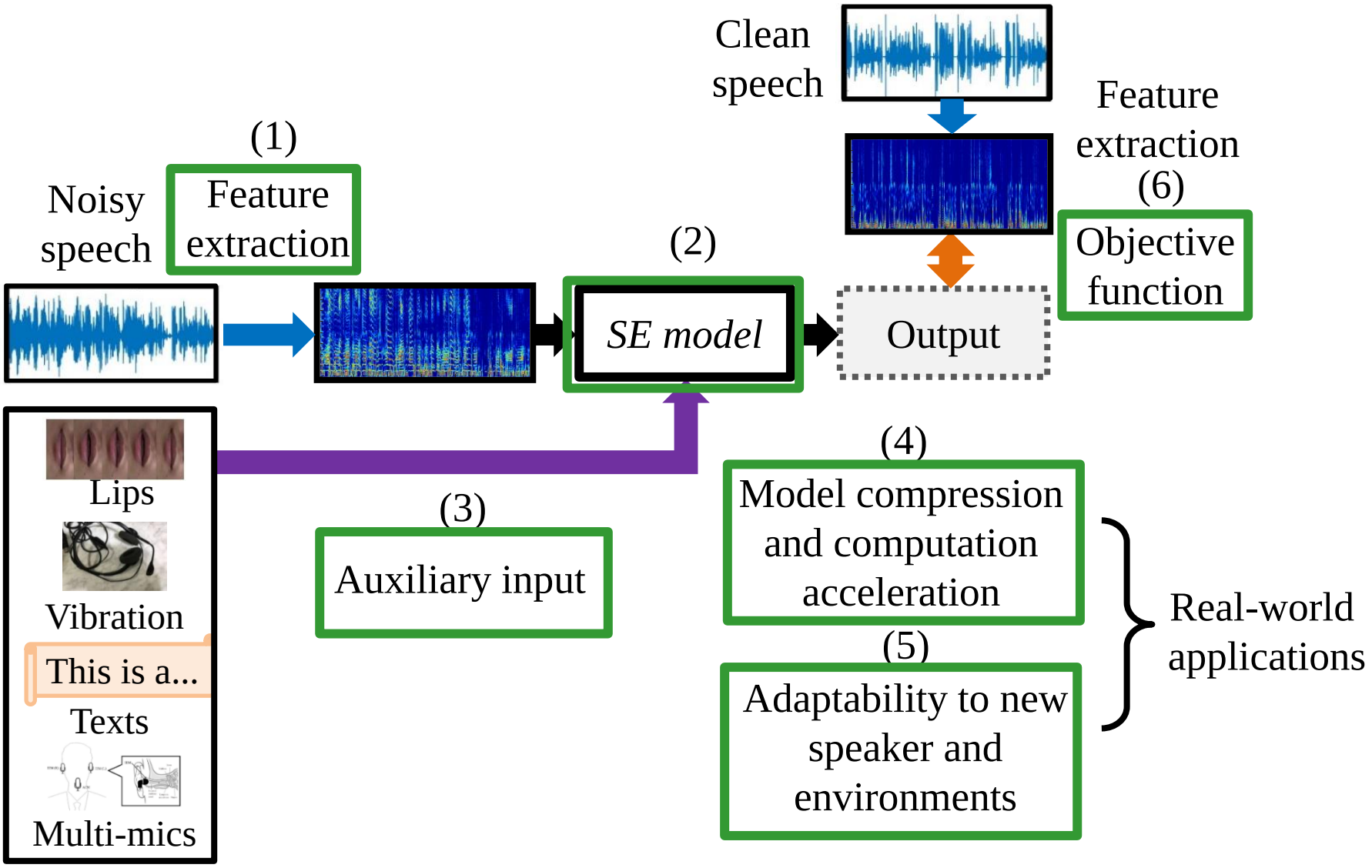
Feature-domain masking



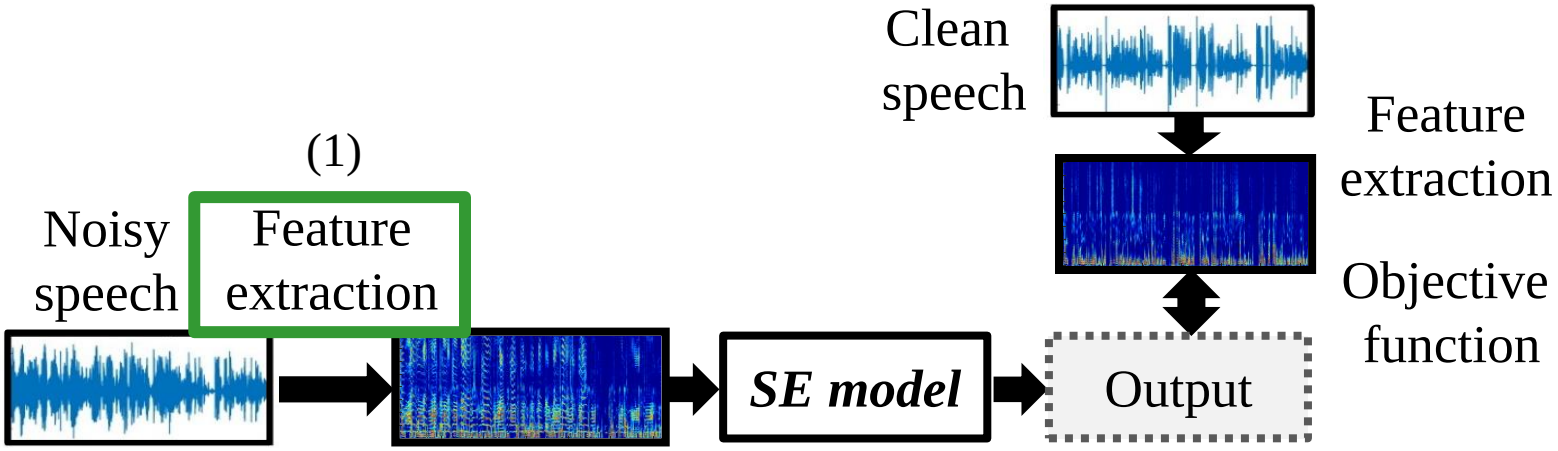
Direct Mapping



Factors of DL-based SE

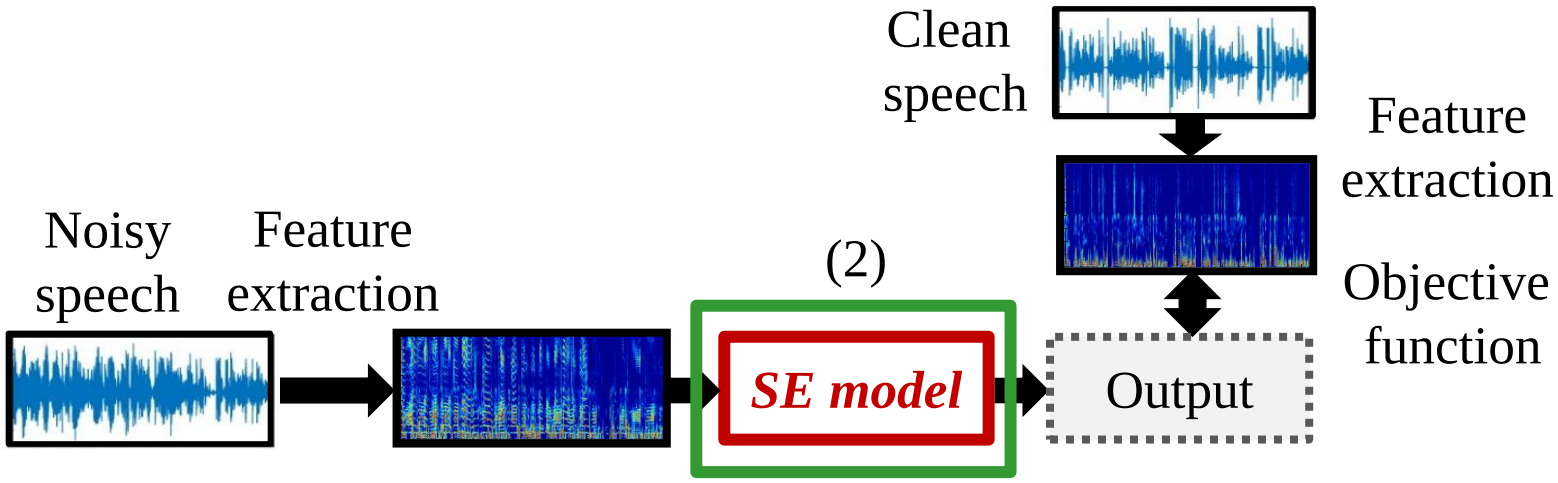


Type of Acoustic Features



- 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, Lu et al., Interspeech 2020],
- Power spectrum [Fu et al., Interspeech 2016],
- Complex spectrum [Fu et al., MLSP 2017, Tan et al., ICASSP 2019, Choi et al., arXiv 2019, Hu et al., arXiv 2020, Wang et al., TASLP 2020, Hu et al., Interspeech 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, Hsieh SPL 2020]……..

Types of DL Models



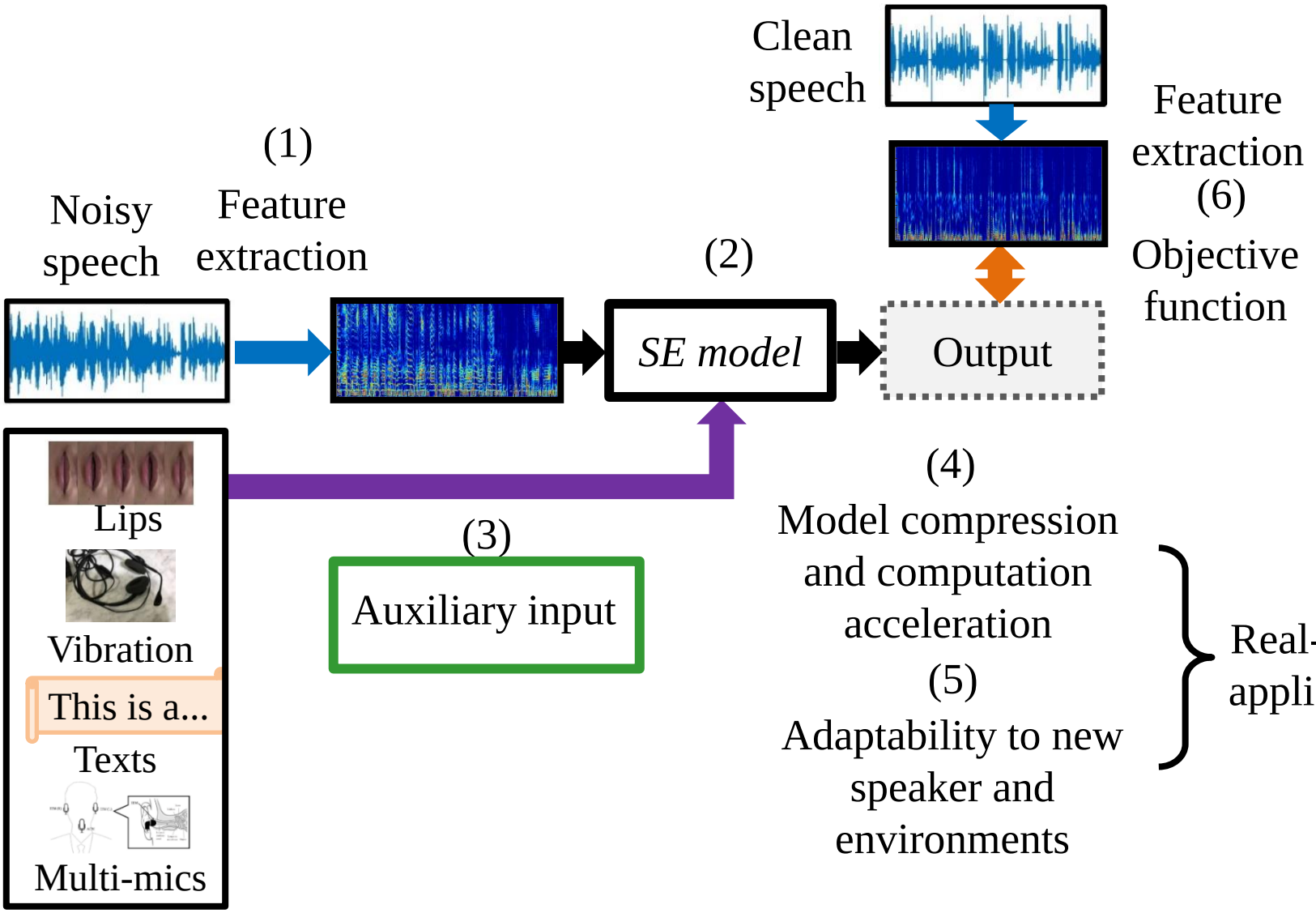
Model types:

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], Transformer [Kim et al., ICASSP 2020, Fu et al., APSIPA 2020].

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]. Ensemble learning [Le Roux, WASPAA 2013, Chazan et al., WASPAA 2017, Zhang et al., TASLP2016, Yu et al., TASLP 2020], Dual path [Pandey et l, TASLP 2020, Chao et al., ASRU 2021, Le et al., Interspeech 2021], Dual branch [Yu, et al, arXiv 2021].

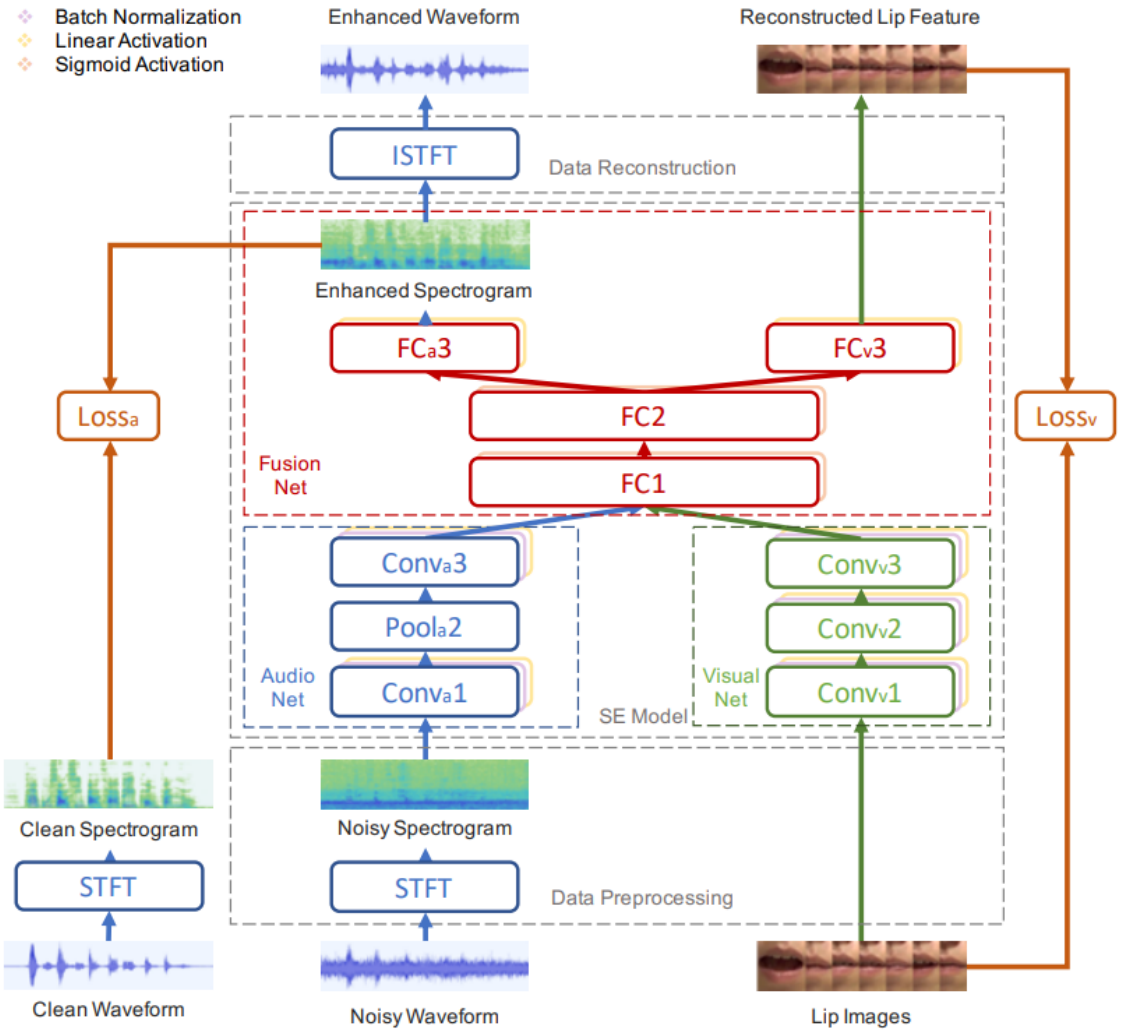
Factors of Deep Learning based SE



Multimodal SE (Visual)

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

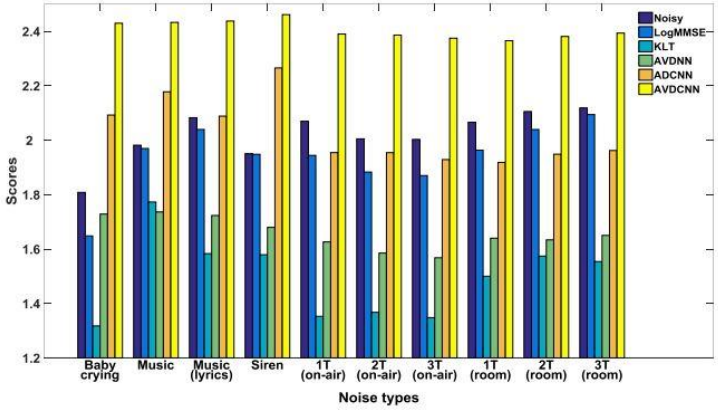
  - Additional parts
    - Lip images
    - Visual Net
    - FCv3
  - Visual target: image



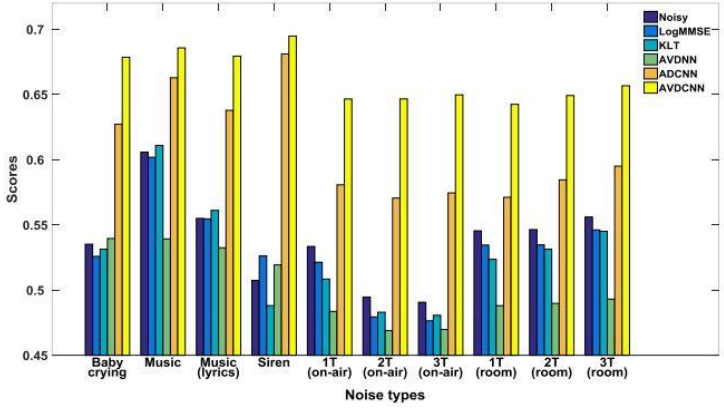
Multimodal SE (Visual)

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

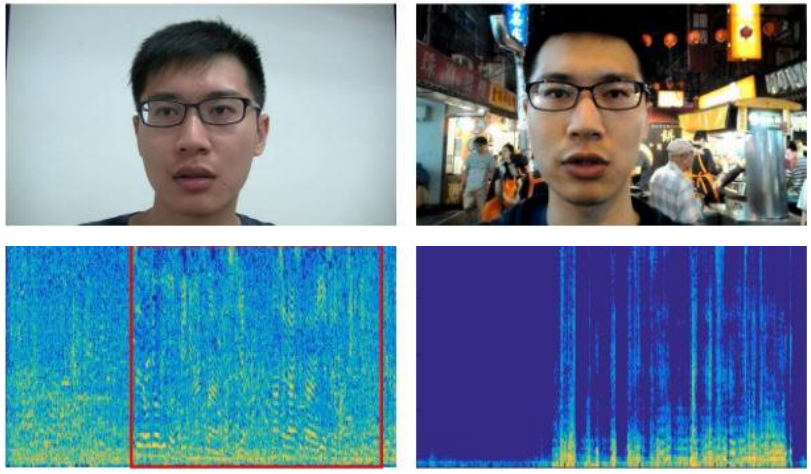
The PESQ scores



The STOI scores



Testing in the real-world conditions



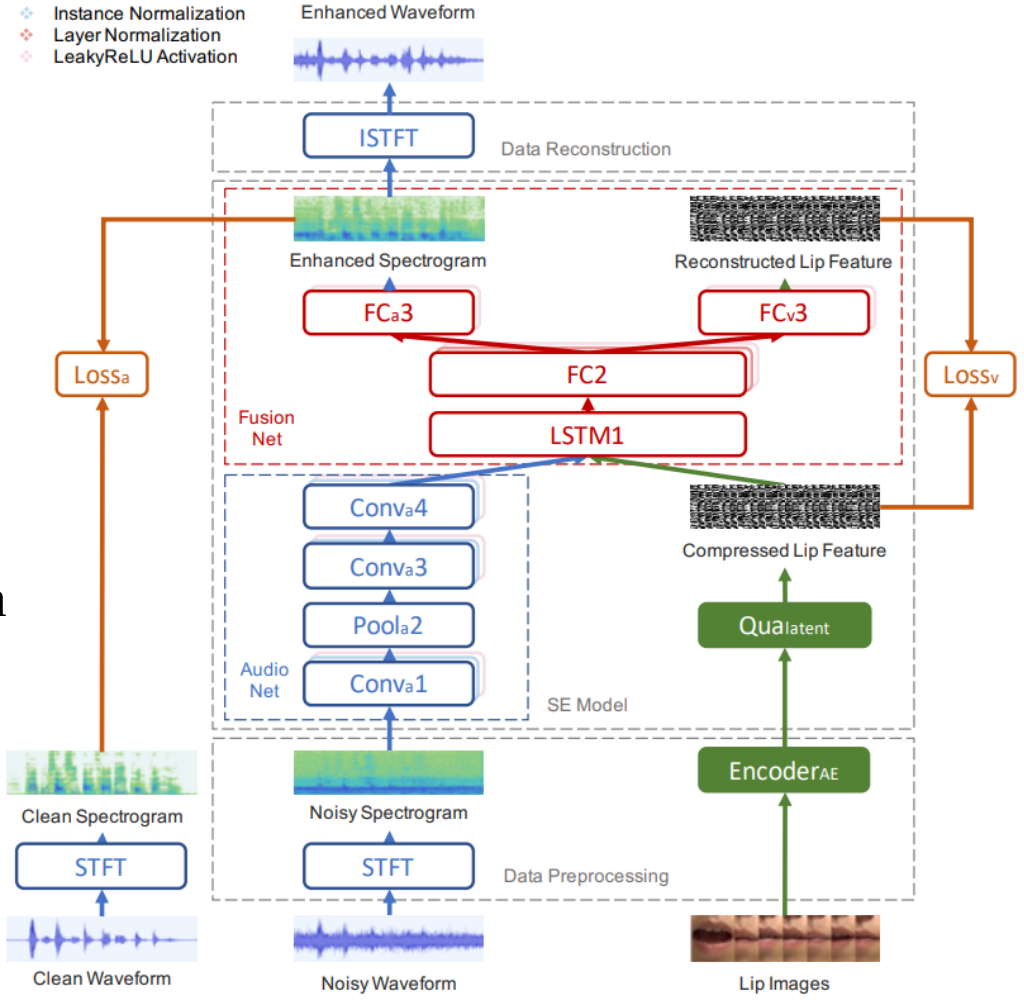
(1) Visual information improves the SE performance.

(2) The performance is robust against recording conditions as long as lips can be recorded well.

Multimodal SE (Visual)

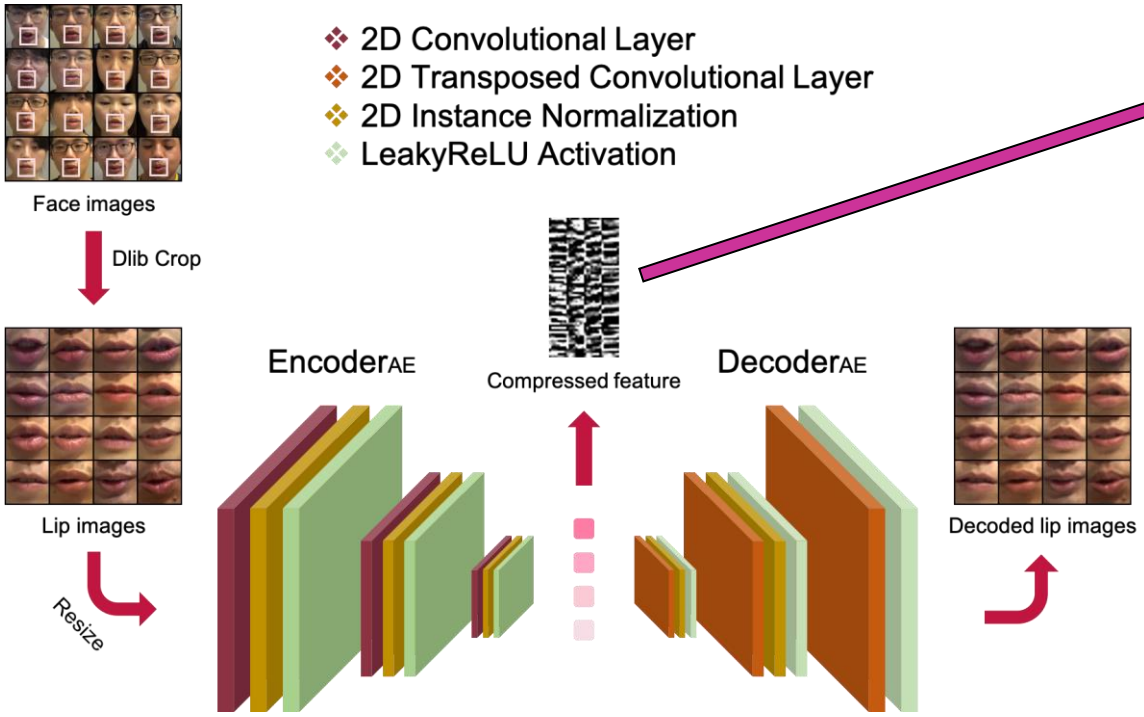
- Lite Audio-visual SE [Chuang et al., Interspeech 2020]

- 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]



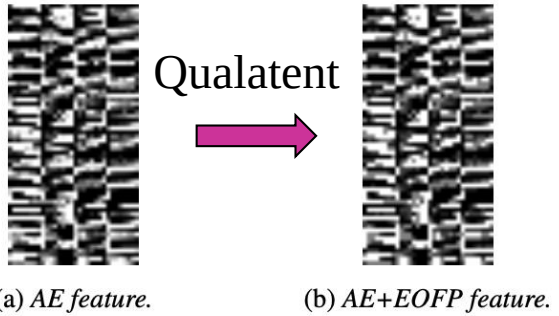


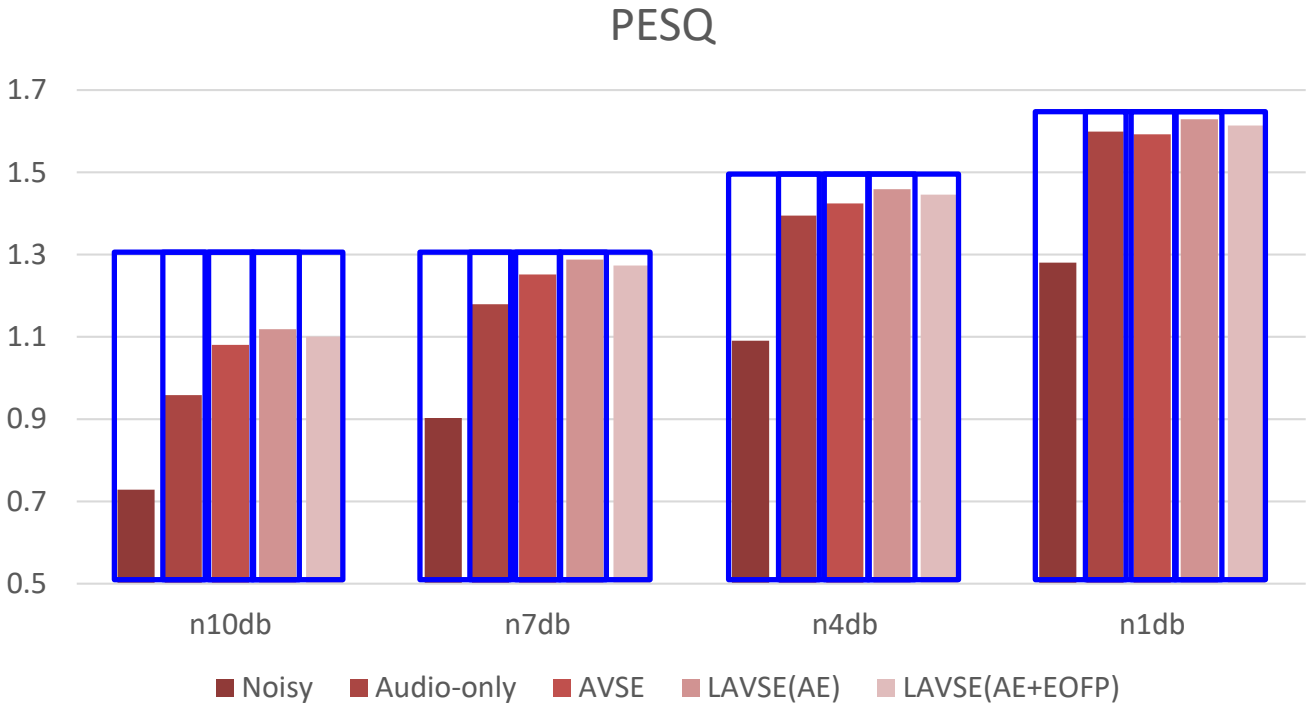
Figure 6: *Visual latent features of lips.*

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

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

Multimodal SE (Visual)

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

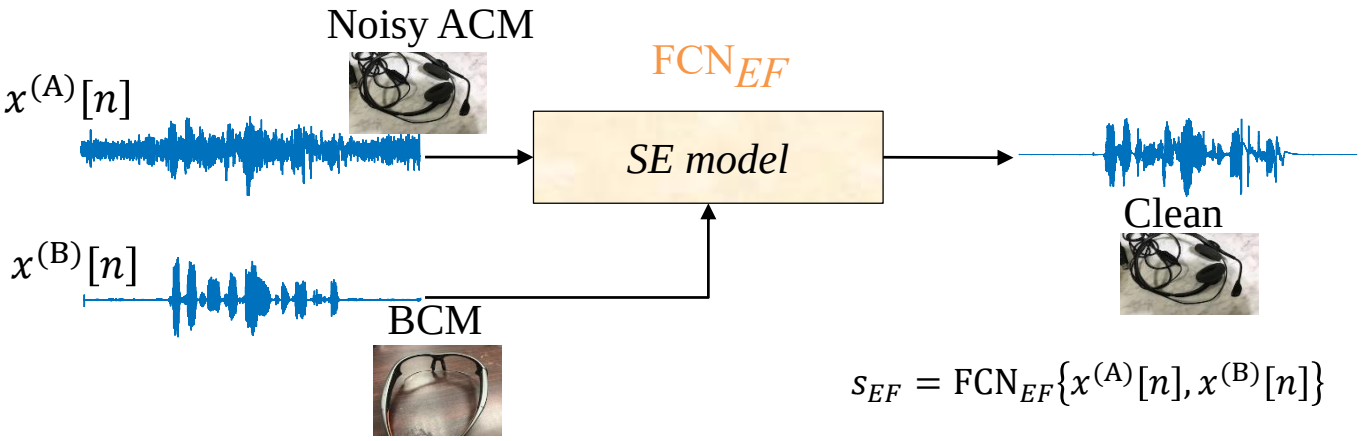


1. Lite AVSE outperforms original AVSE.
2. AVSE+EOFP slightly underperforms AVSE with a notable reduction of 48 times on the visual features.

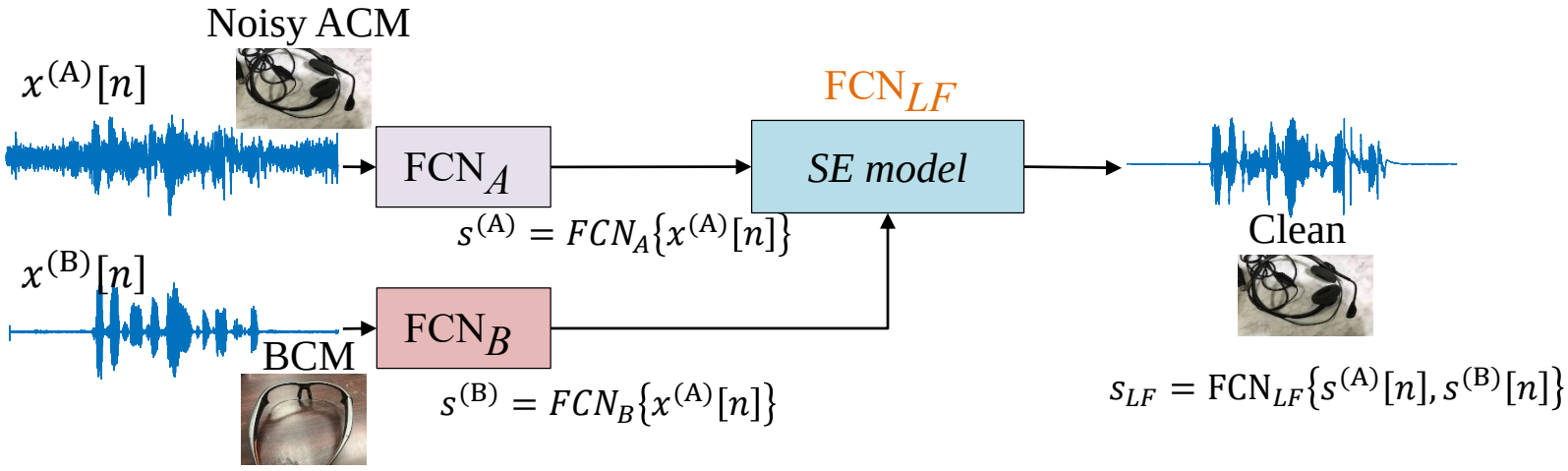
TMSV dataset: https://drive.google.com/drive/folders/1B-eJs1yYVf0qHrYOWrtxYs3a8inPHm1K

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

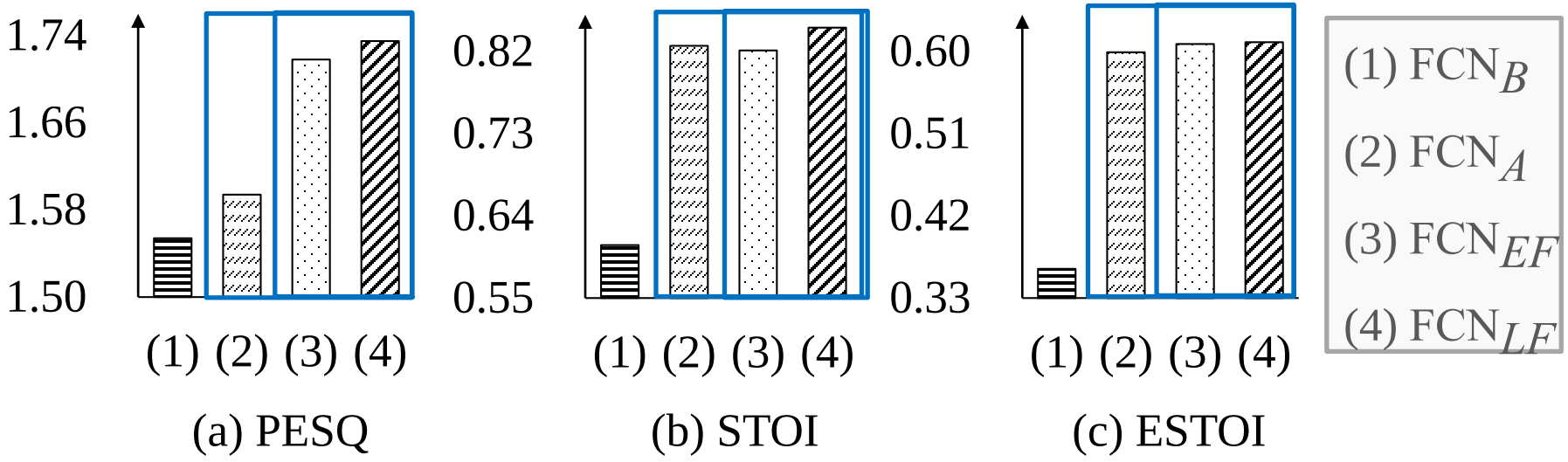


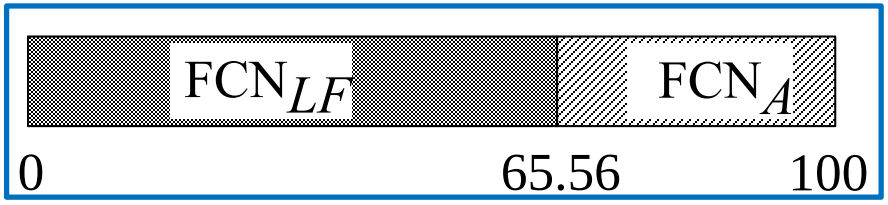
  - The input of the *Fusion* function is processed noisy and BCM signals



Multimodal SE (Bone-conducted)

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



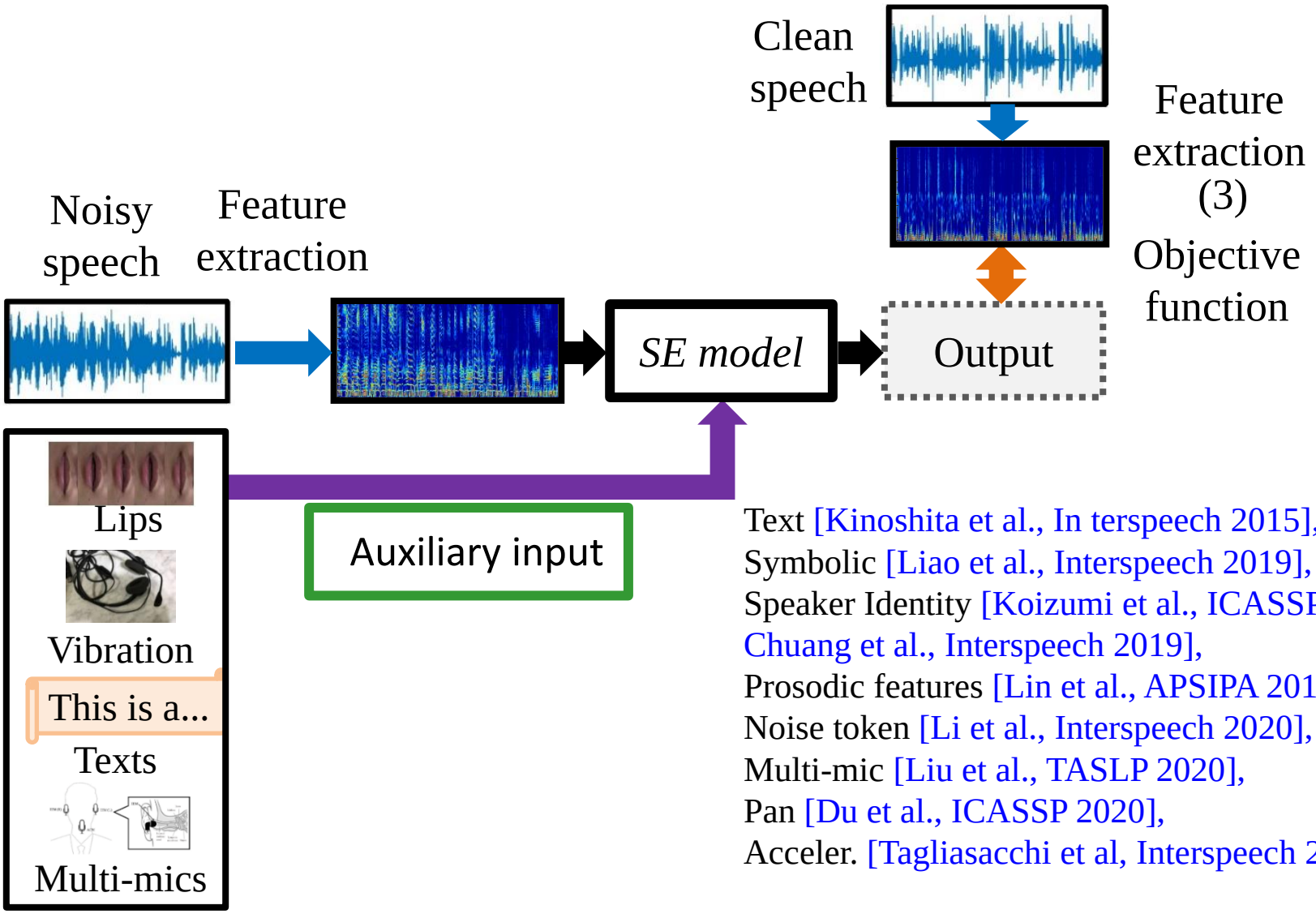


The results (in percentage, %) for the AB test that compares FCN_{LF} and FCN_A.

($p = 0.00088 < 0.01$)

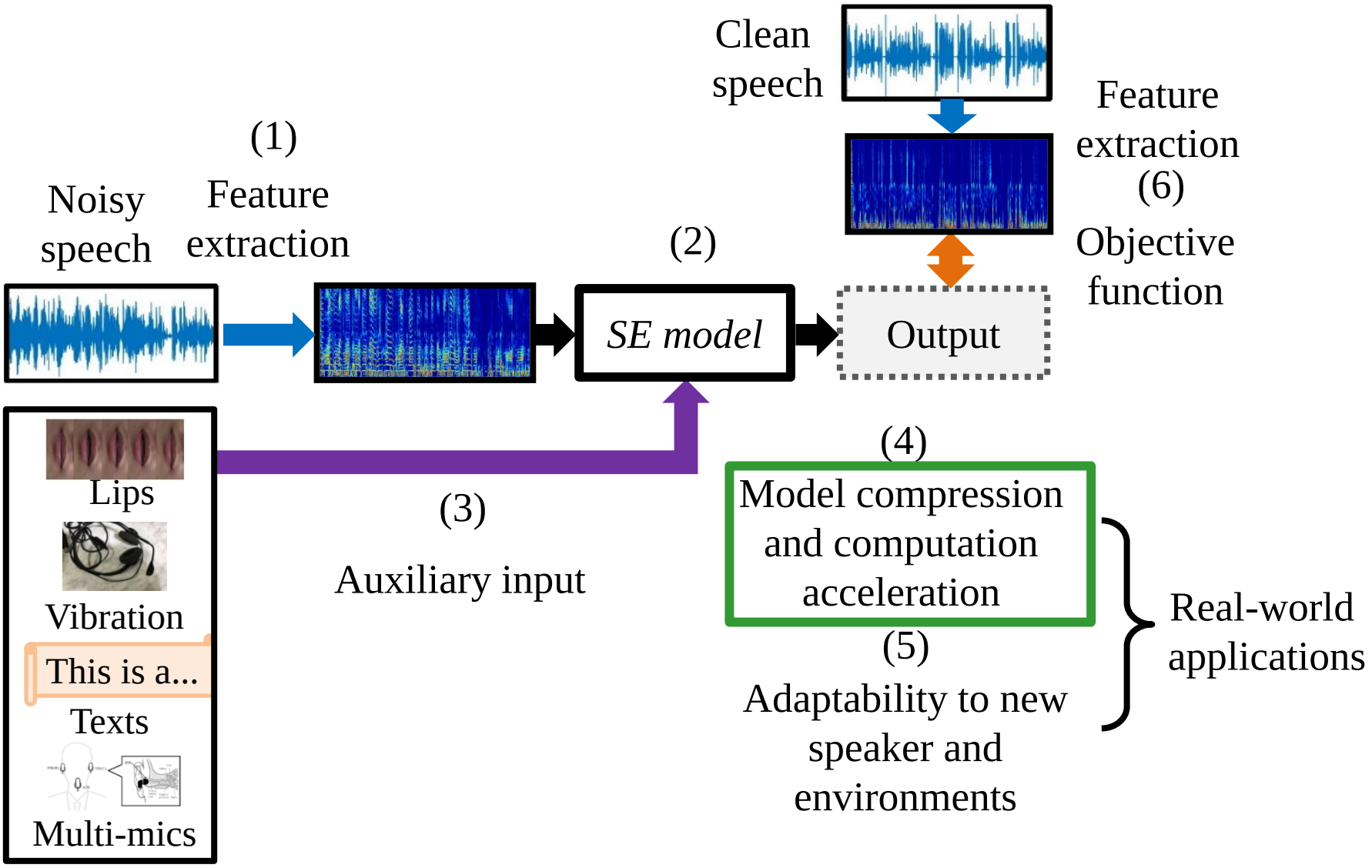
(1) BCM information improves the SE performance in terms of PESQ, STOI, ESTOI and listening tests.
(2) Late fusion outperforms early-fusion.

Factors of Deep Learning based SE

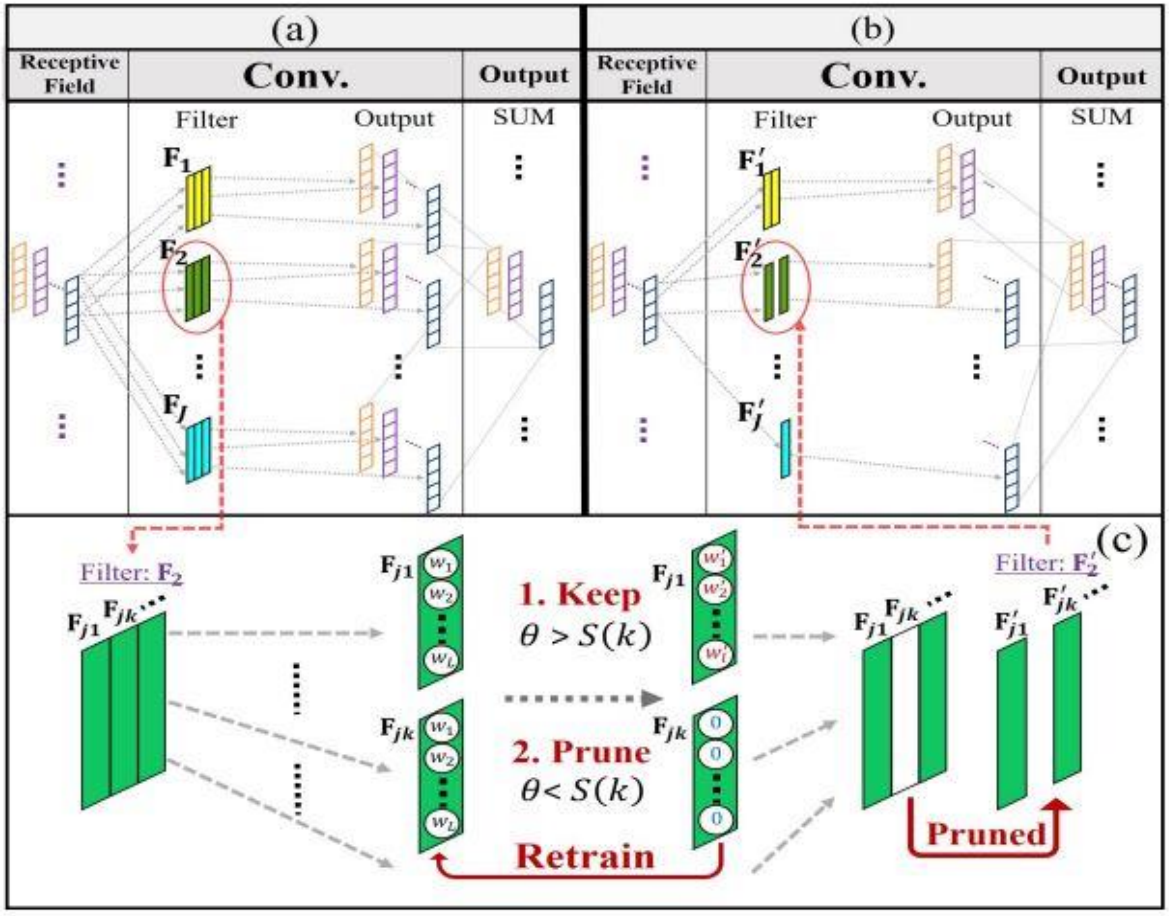


Text [Kinoshita et al., In terspeech 2015],
Symbolic [Liao et al., Interspeech 2019],
Speaker Identity [Koizumi et al., ICASSP 2020; Chuang et al., Interspeech 2019],
Prosodic features [Lin et al., APSIPA 2019],
Noise token [Li et al., Interspeech 2020],
Multi-mic [Liu et al., TASLP 2020],
Pan [Du et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

Factors of Deep Learning based SE



Model Compression (Pruning)



A computation performance optimization (CPO) approach is proposed to optimally compress the model [Wu et al., IEEE SPL 2019].

Model Compression (Pruning)

- The CPO approach 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:

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

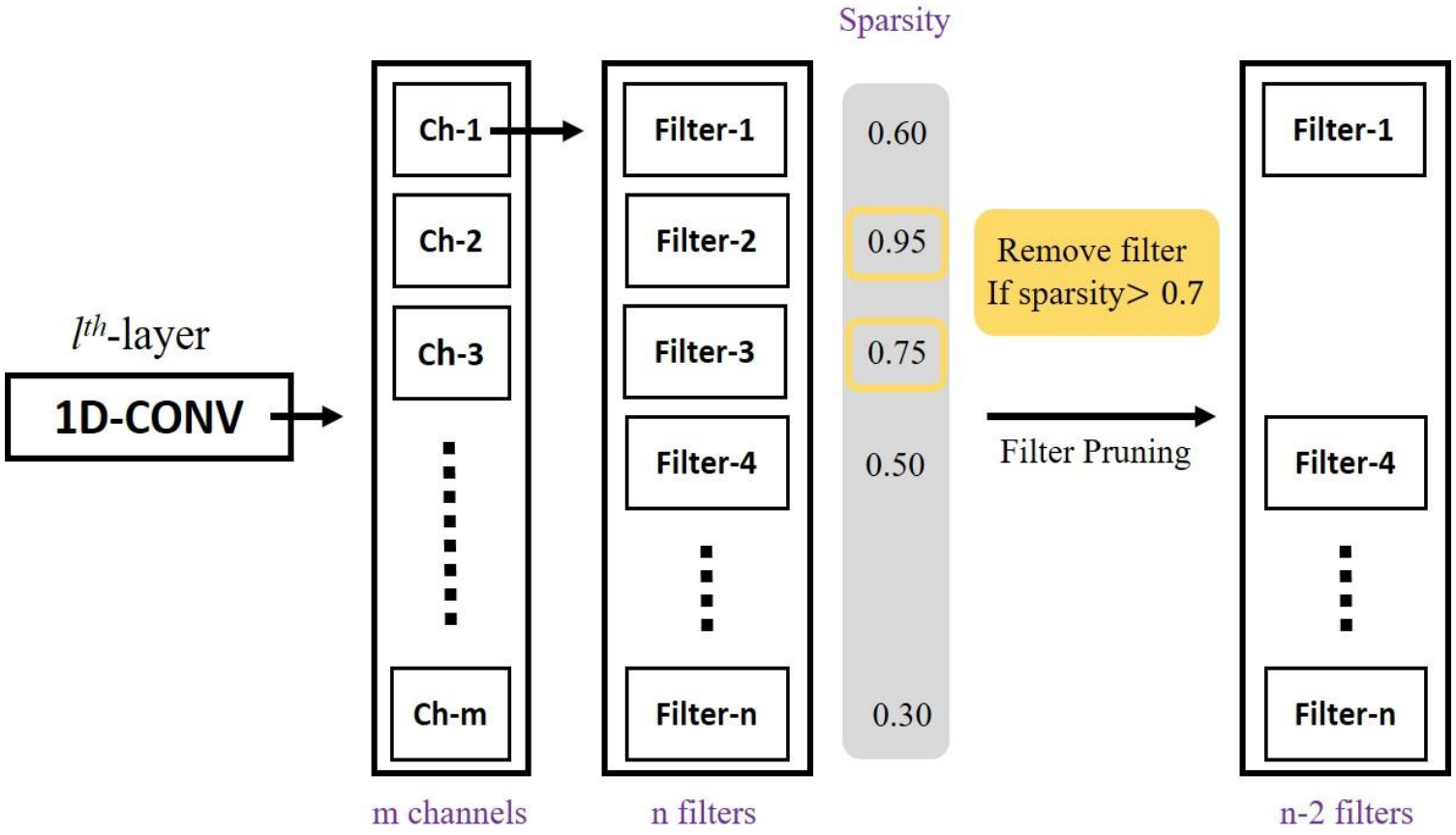
  (2) Compute the ***sparsity*** of the *n*-th filter:

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

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

Model Compression (CPO)

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



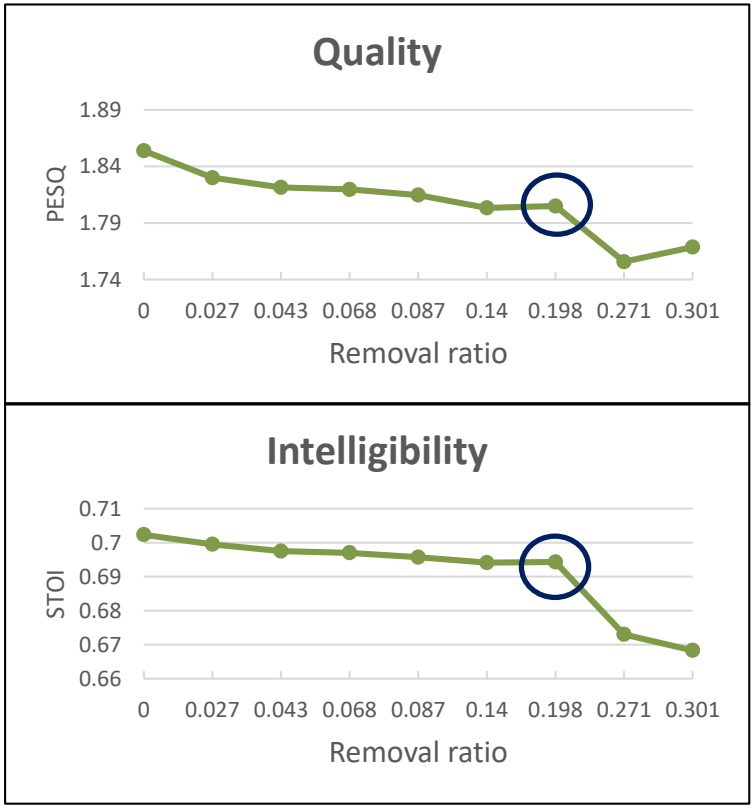
N: $filter\ number$
W: $weight\ number$

Model Compression (CPO-SE)

- The results of CPO

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

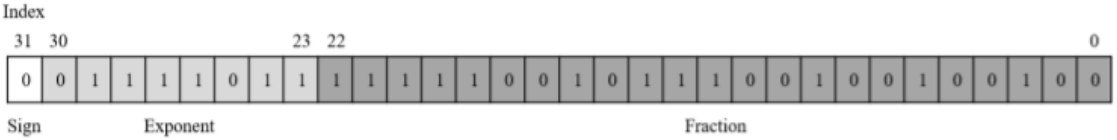
Threshold	Removal ratio	PESQ	STOI
1.0	0	1.85385	0.70231
0.95	0.027	1.83	0.6995
0.9	0.043	1.8215	0.6975
0.85	0.068	1.8197	0.697
0.8	0.087	1.8147	0.6957
0.75	0.14	1.8034	0.6941
0.7	**0.198**	1.805	0.6943
0.65	0.271	1.7558	0.673
0.6	0.301	1.7687	0.6683
Noisy		1.63713	0.66977



A notable performance drop when Threshold <0.65.
The compression ratio can be 20% as compared to the original model.

Parameter Quantization

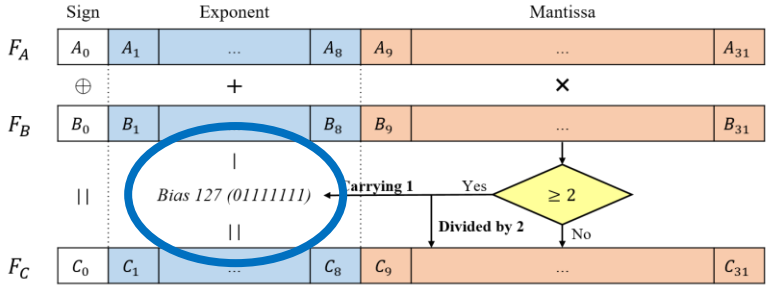
- Sign-exponent-only floating-point network [Lin et al., IEEE TASLP 2021]
  - ① **1 Sign bit**
  - ② **8 Exponent bits**
  - ③ **23 Fraction (Mantissa) bits**



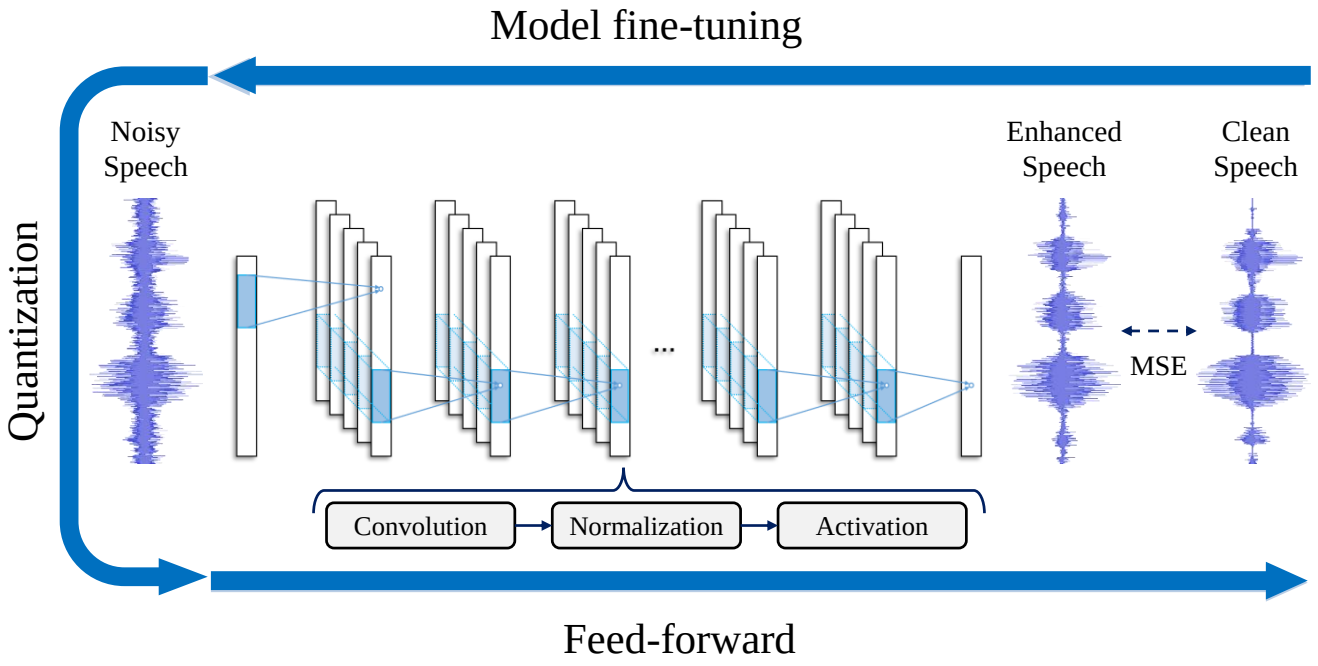
$$(value)_{10} = (-1)^{sign} \times (fraction)_{10} \times (2)^{(exponet)_{10} - bias}$$

- Circuits in neural computing
  - **Floating-point may not be an optimal** representation for neural computational on hardware devices ➔ Some **redundant** bits
  - **Multiplication** and **Division** circuits are more complicated than **Addition** and **Subtraction** circuits ➔ Inefficiency

IA-NET: Use the most efficient circuit **Integer-adder** to calculate the multiplication



Training IA-NET



- Model learns how to quantize during training:
  - ① Model fine-tuning step: Train the model based on the computed gradients
  - ② Quantization step: Quantize the parameters in the model
  - ③ Feed-forward step: Use the updated model to compute gradients

Avoid Performance Degradation

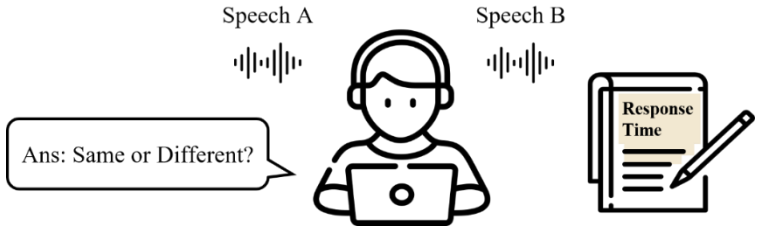
Efficiency and Performance Evaluations

- Compression and acceleration
  - **IA-Net maintains SE performance (PESQ)**
  - **Compression rate** of model size:
    - ✓ Compression ratio is **27.58%** (from 1759 to 495) of model size
  - **Speedup rate** of inference time:
    - ✓ The inference time is **1.21x faster** (110.691 to 91.308) than original model

	Noisy	**FCN-10**		**FCN-12**	
		Original	IA-NET	Original	IA-NET
SNR(dB)	PESQ	PESQ	PESQ	PESQ	PESQ
-6	1.223	1.381	1.444	1.379	1.514
0	1.622	1.843	1.877	1.843	1.920
6	2.016	2.304	2.281	2.297	2.303
12	2.439	2.729	2.700	2.715	2.691
Ave.	1.825	2.064	2.076	2.058	2.107
Time (ms)		110.691	91.308	134.035	110.709
Size (KB)		1759	495	2147	604

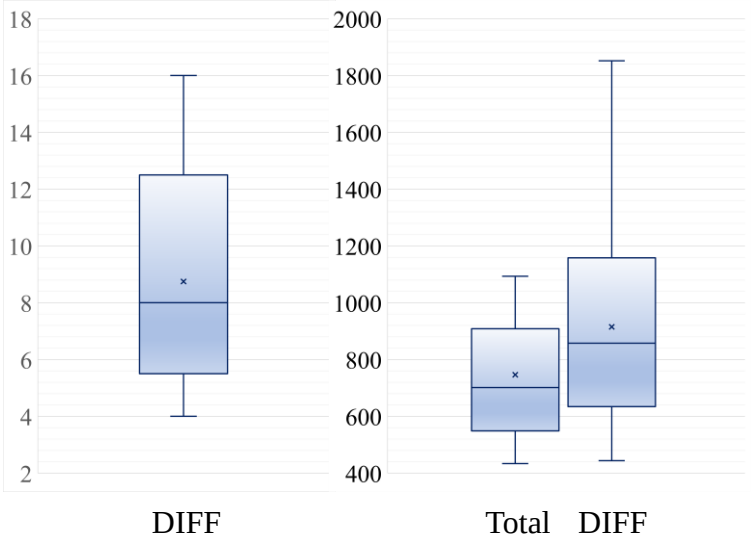
Just Noticeable Difference

- JND A/B test user study [Lin et al., IEEE TASLP 2021]
  - ➢ 20 participants
  - ➢ 100 pairs of speech signals
  - ➢ 6 bit-width scenarios
    - ✓32, 26, 20, 14, 10 and 9 bits for model parameters
    - ✓Compared against original model without quantization
    - ✓Listerner makes a decision: **SAME** or **DIFF**

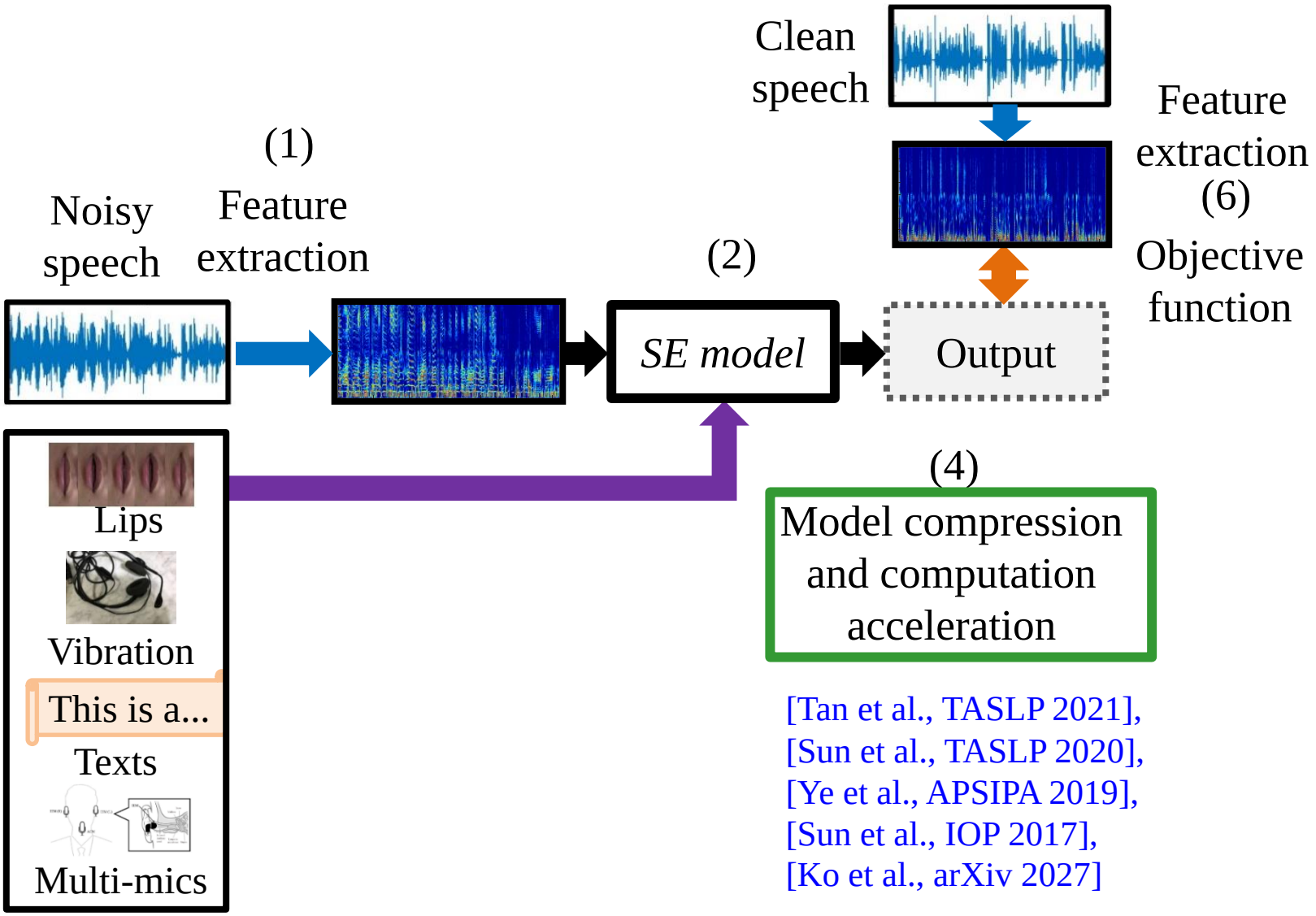


- Amount of **DIFFs** and response time
  - ➢ Average quantity of **DIFF** is only 8.75 out of the 100 pairs
  - ➢ Response time
    - ✓Total average: 746.36 ms
    - ✓DIFF average: 915.09 ms

➔ **Unable to effortlessly differentiate between A and B**

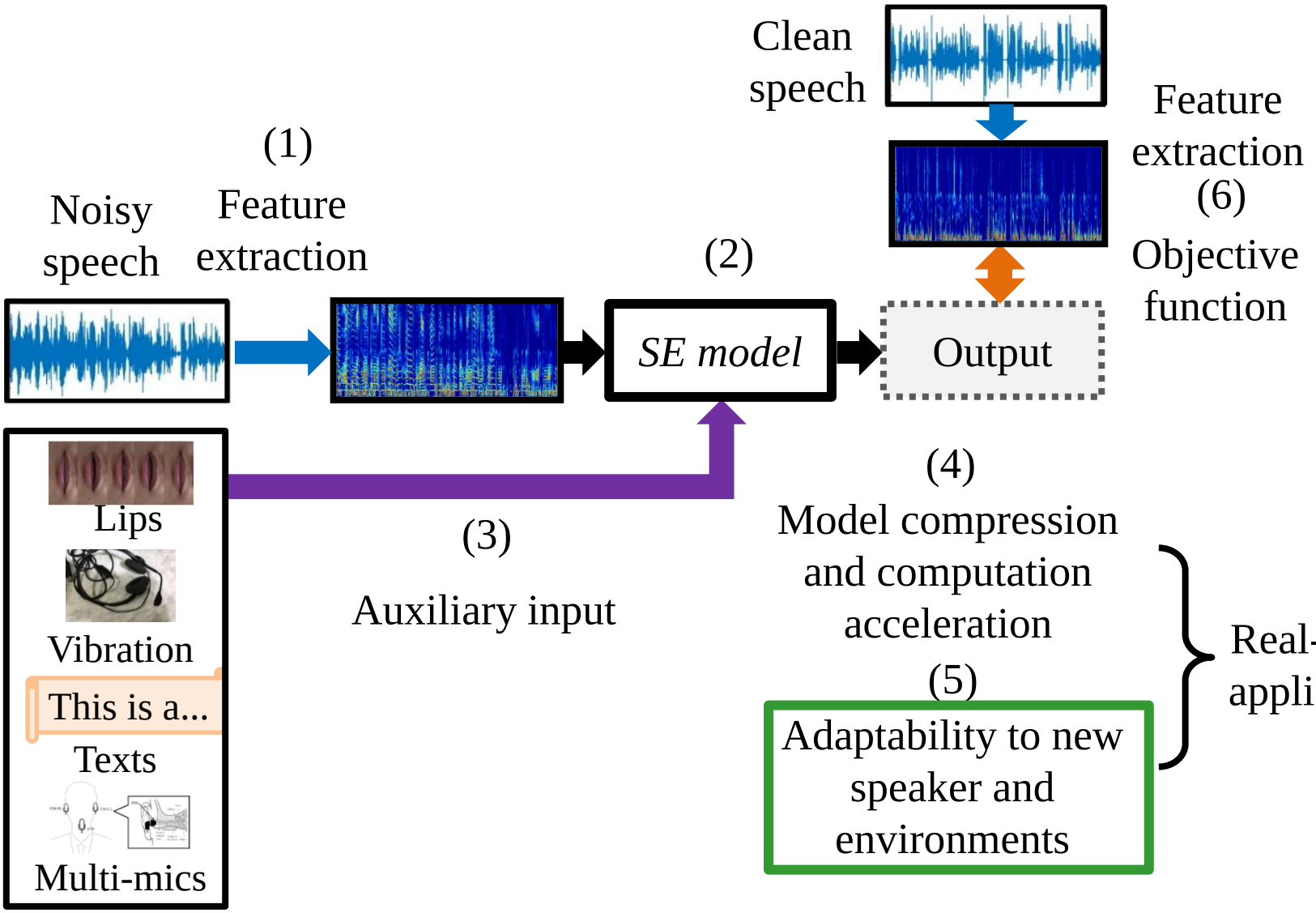


Factors of Deep Learning based SE



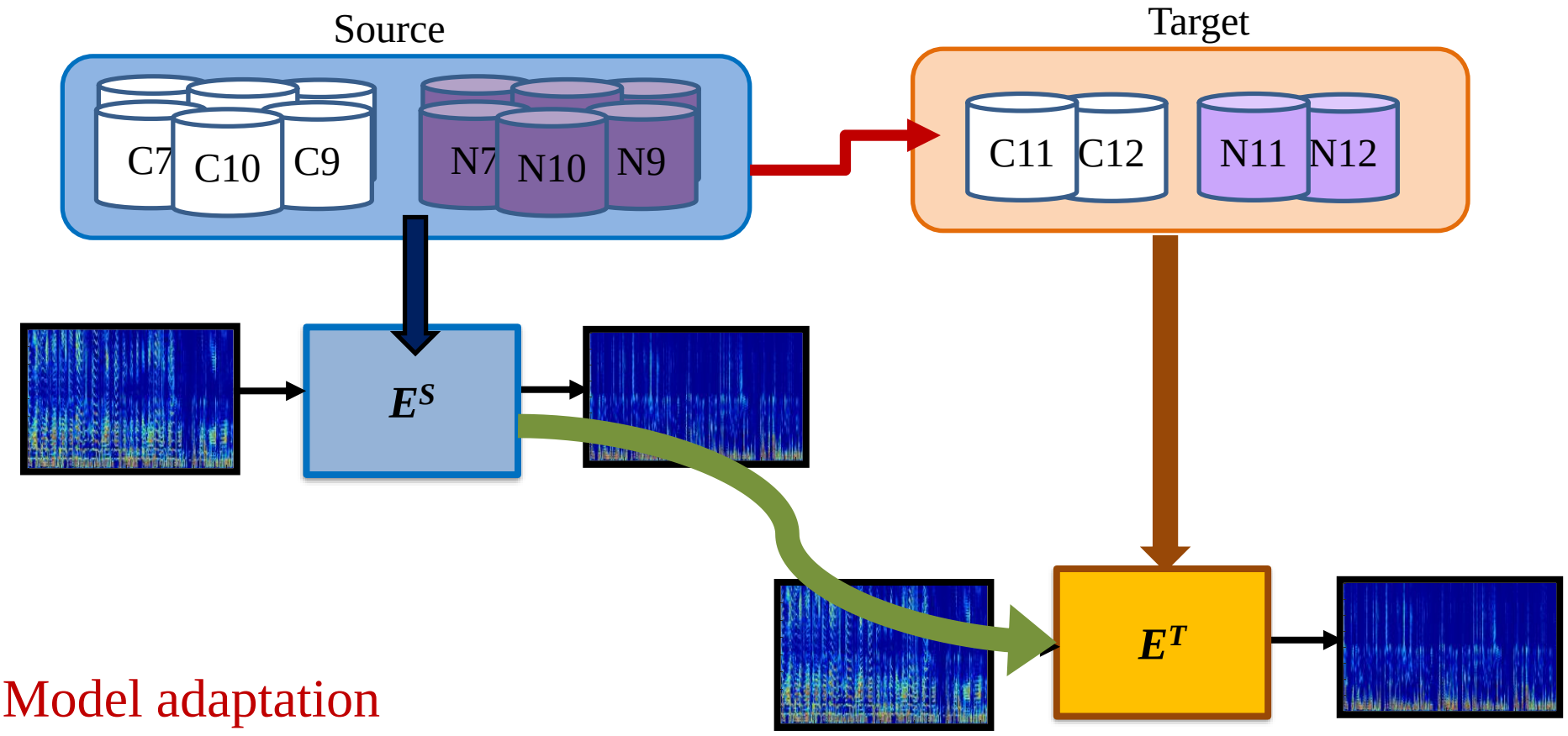
[Tan et al., TASLP 2021],
[Sun et al., TASLP 2020],
[Ye et al., APSIPA 2019],
[Sun et al., IOP 2017],
[Ko et al., arXiv 2027]

Factors of Deep Learning based SE



SE Model Adaptation

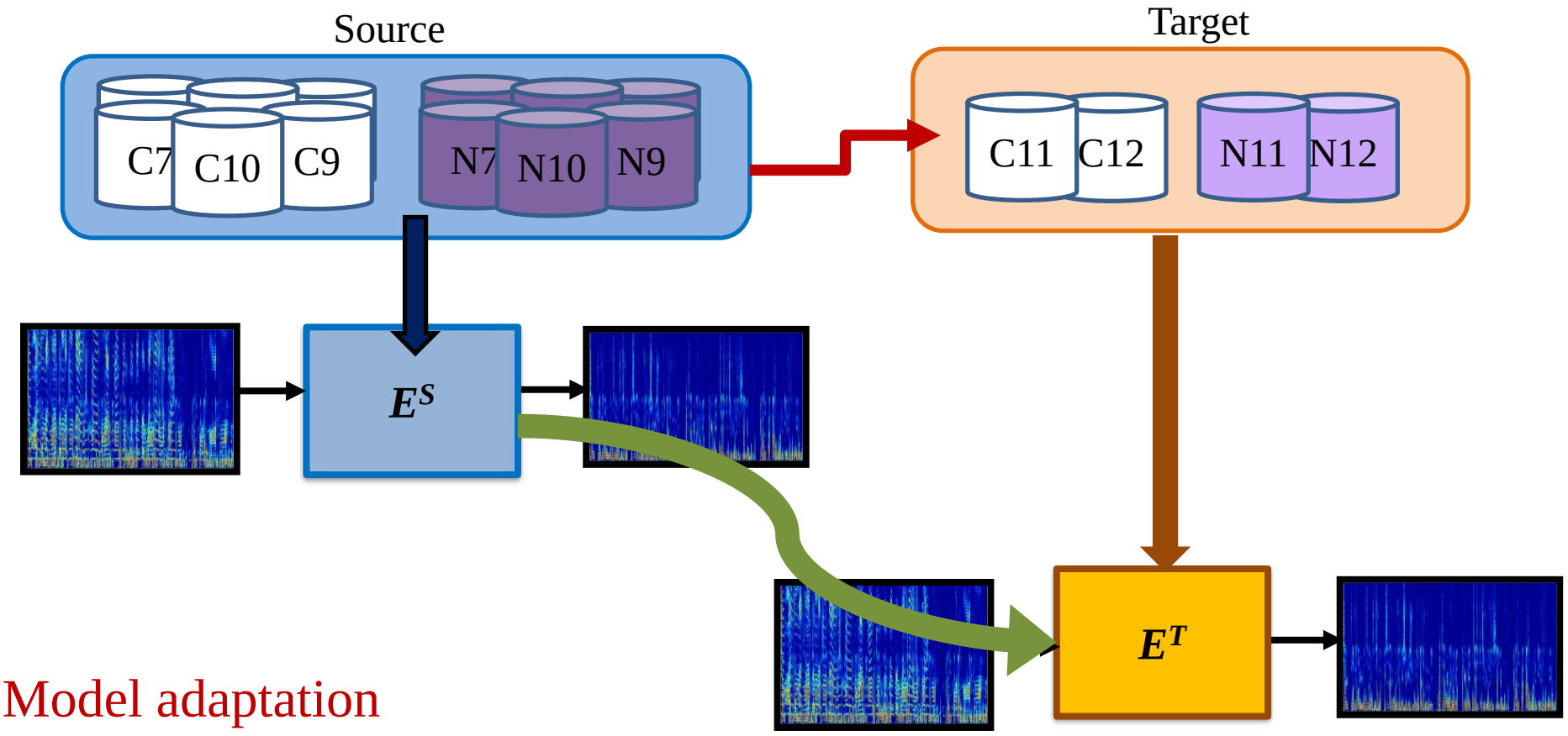
- For SE model adaptation:



Model adaptation
(1) supervised: with paired noisy/clean
(2) unsupervised: wo/ paired noisy/clean

SE Model Adaptation

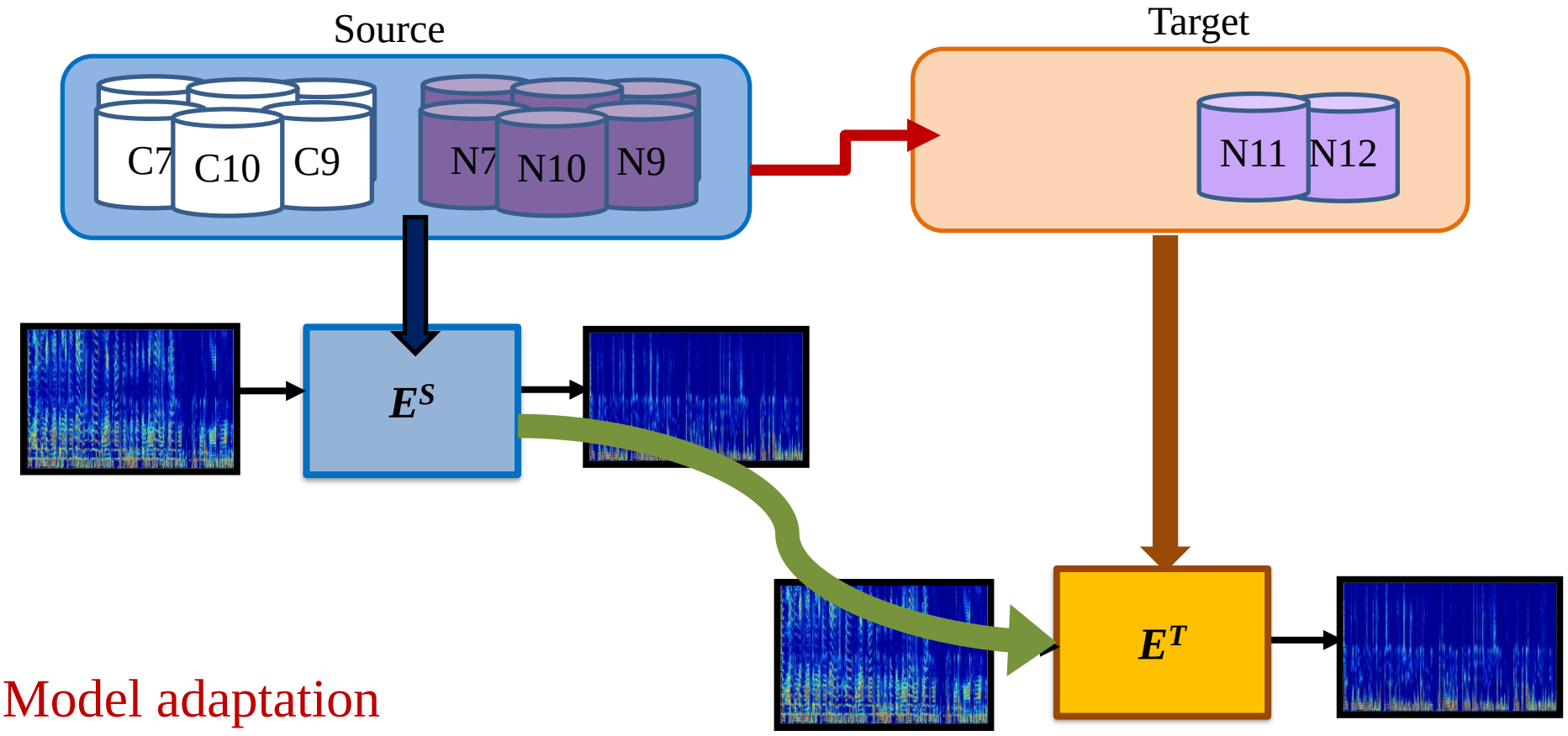
- For SE model adaptation:



Model adaptation
(1) supervised: with paired noisy/clean

SE Model Adaptation

- For SE model adaptation:



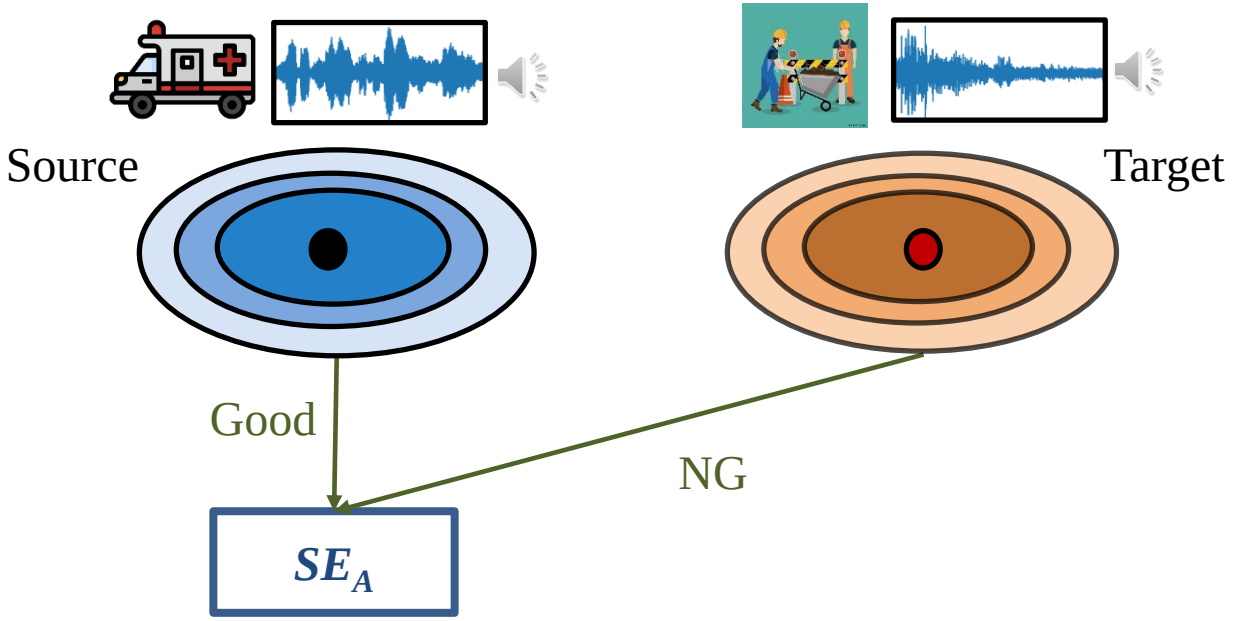
Model adaptation

(2) unsupervised: wo/ paired noisy/clean

SE Model Adaptation (Supervised)

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

✓ For supervised model adaptation:

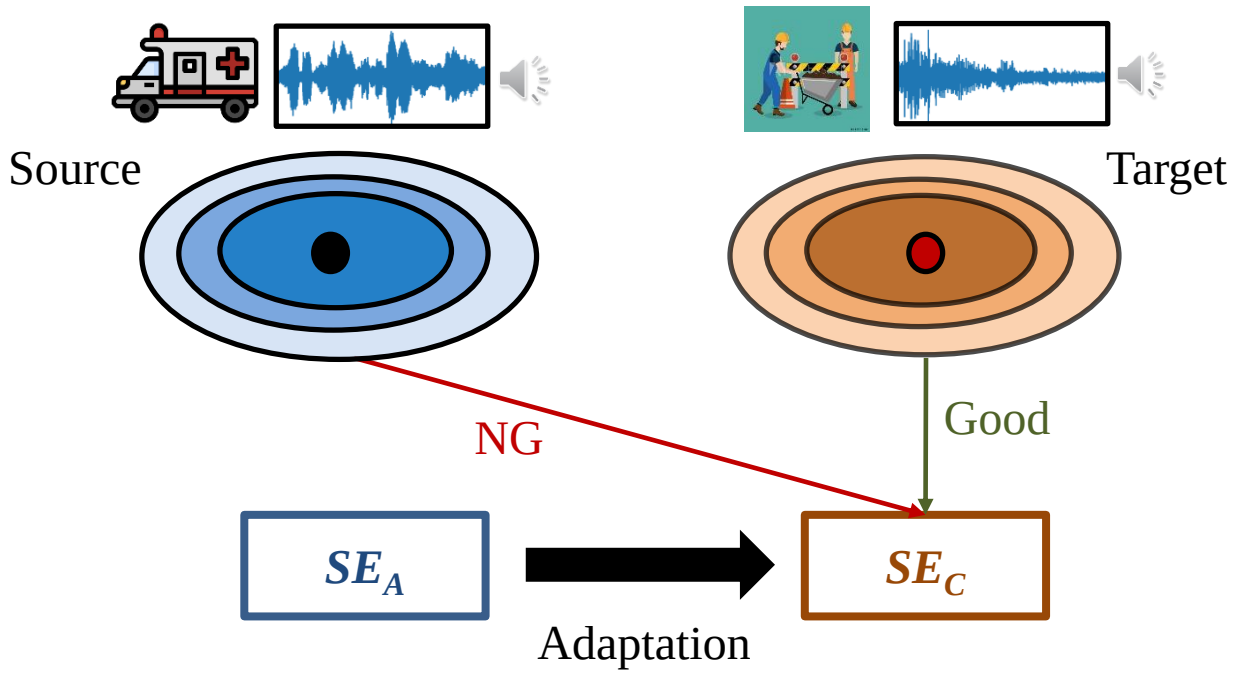


Noise/speaker mismatch may cause poor SE performance.

SE Model Adaptation (Supervised)

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

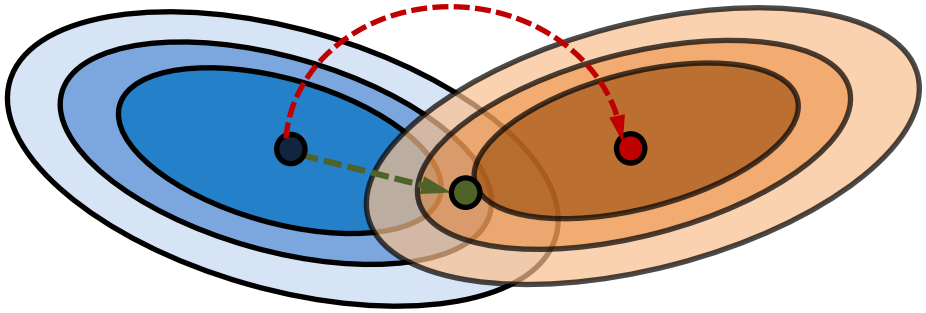
✓ For supervised model adaptation:



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

SE Model Adaptation (Supervised)

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



Rather then direct adaptation, SERIL adopts proper constraints.

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

Not available — From target data

Constraints

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]

SE Model Adaptation (Supervised)

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

Original: training set

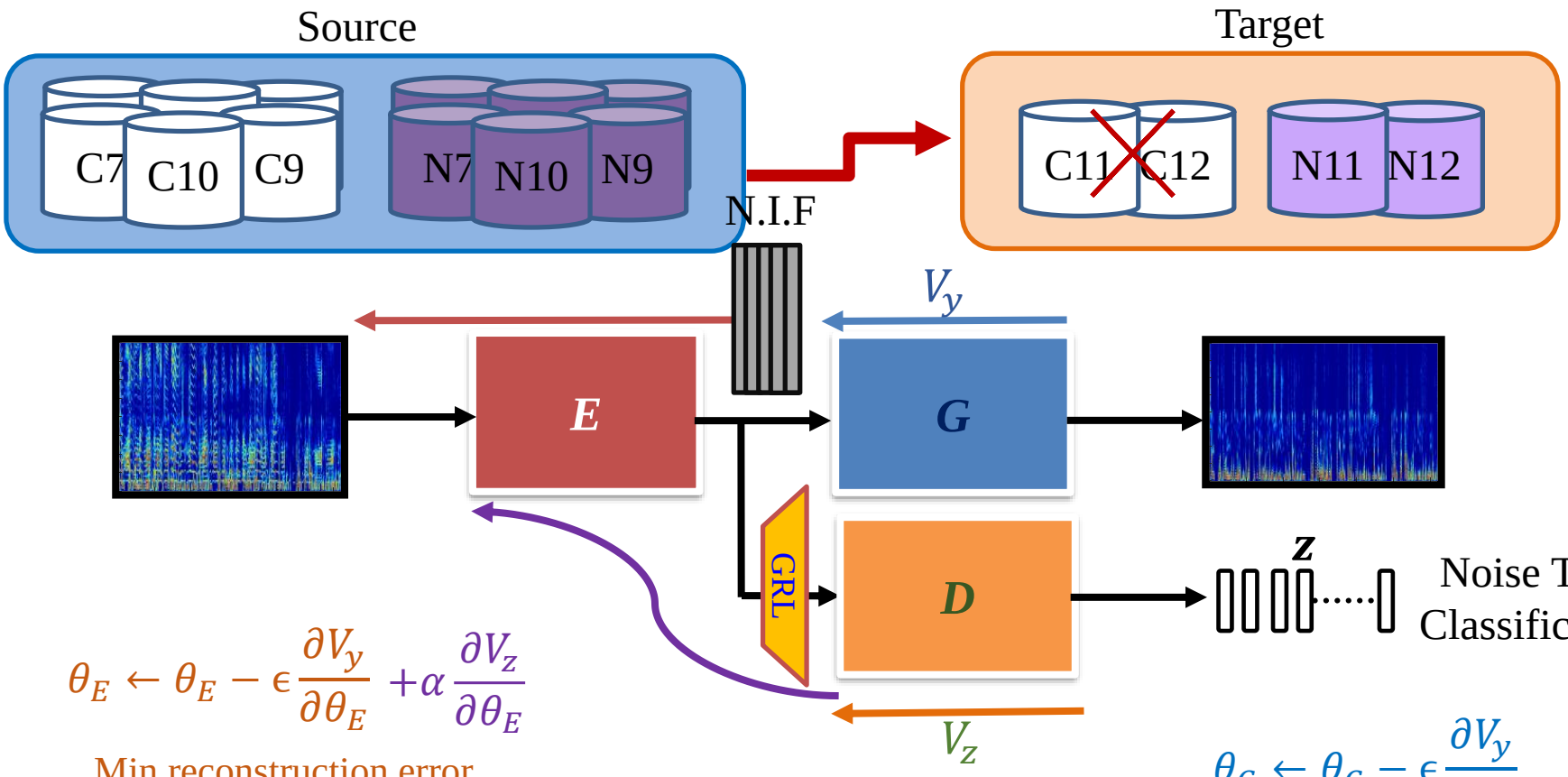
Metric	M	original	cough	door moving	foot-steps	clap
PESQ	N	2.266	2.041	1.864	1.868	1.474
	P	**2.708**	2.118	2.059	2.015	1.603
	F	2.406	2.204	2.339	2.133	**2.948**
	R	2.461	**2.375**	**2.581**	**2.381**	2.936
STOI	N	0.816	0.788	0.743	0.778	0.789
	P	**0.869**	0.798	0.779	0.799	0.801
	F	0.811	0.816	0.825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

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

(1) Original model achieves the best in the original testing set.
(2) Original model cannot perform well in new domains.
(3) Direct adaptation suffers from the catastrophic forgetting issue.
(4) SERIL consistently improves permeance for all noise types.

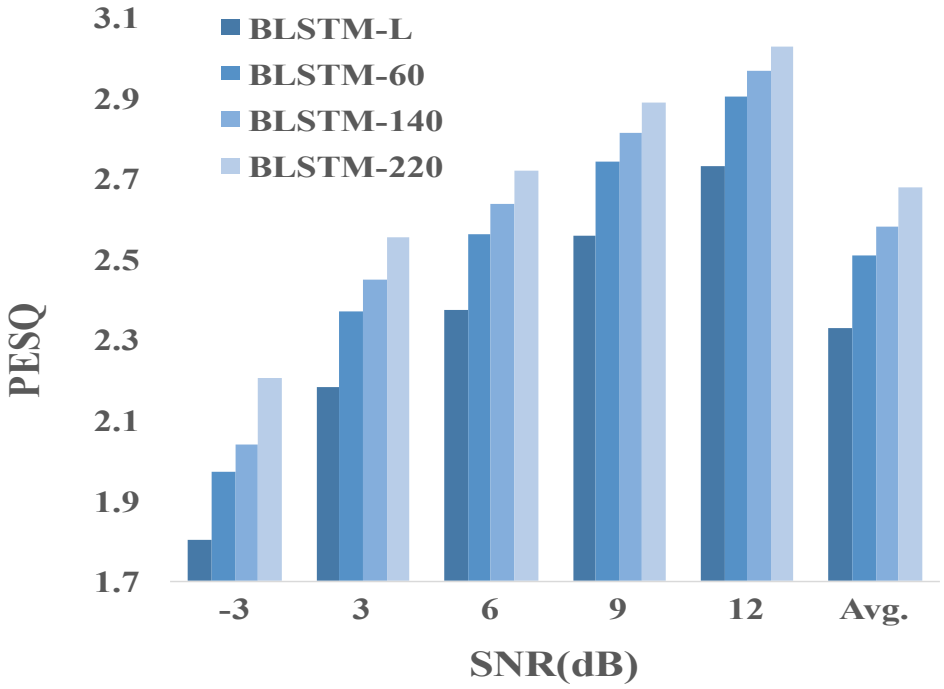
SE Model Adaptation (Unsupervised)

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



SE Model Adaptation (Unsupervised)

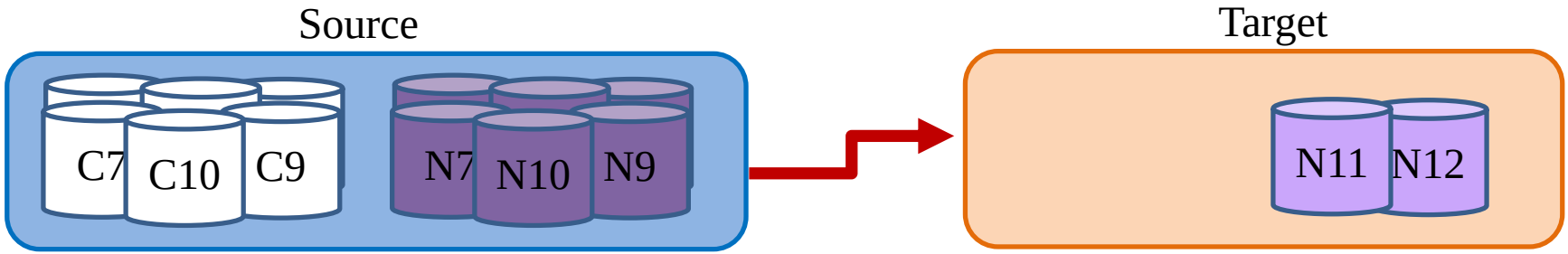
- Noise-adaptive DAT (NADAT) [Liao et al., Interspeech 2019]
  - ✓ Adapting to new noise type (Baby cry)

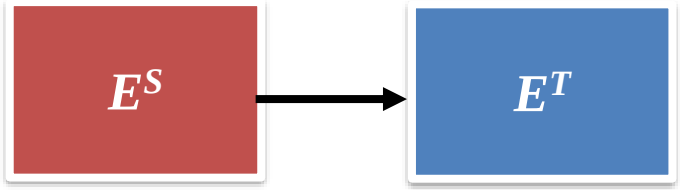


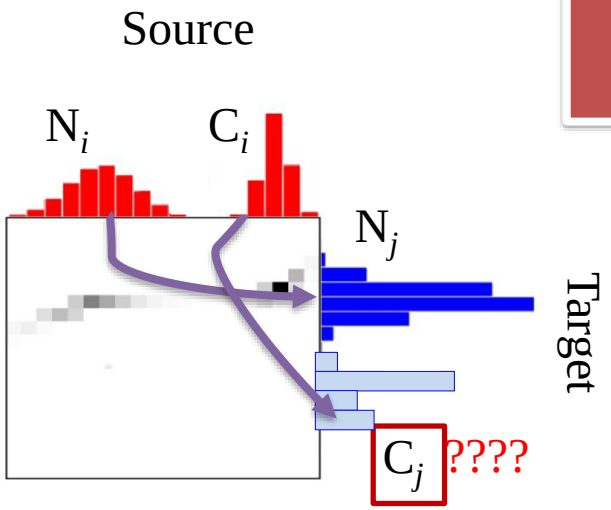
(1) DAT achieves good **unsupervised** adaptation performance (without paired noisy-clean adaptation data).
(2) More adaptation data gives higher scores.

SE Model Adaptation (Unsupervised)

- DOTN [Lin et al., NeurIPS 2021]



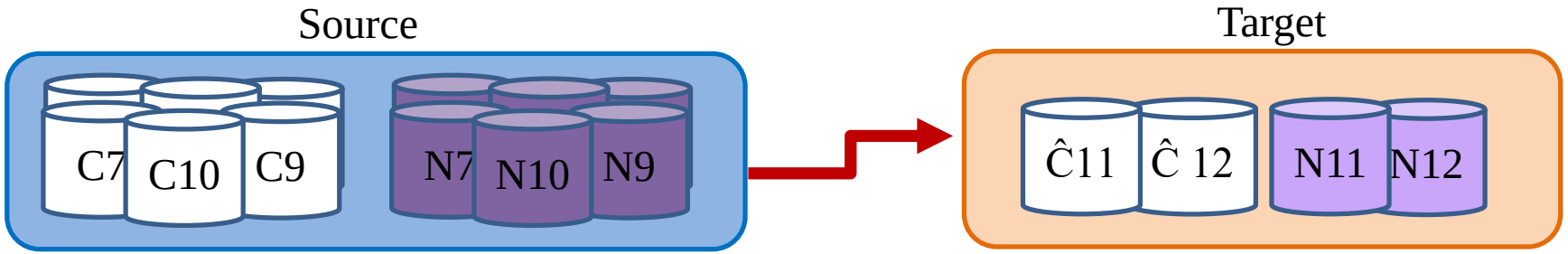


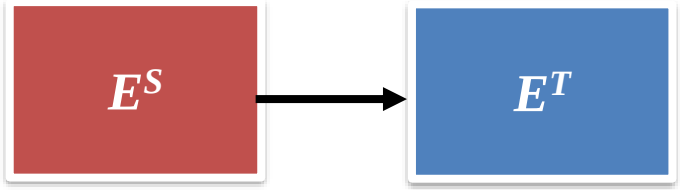


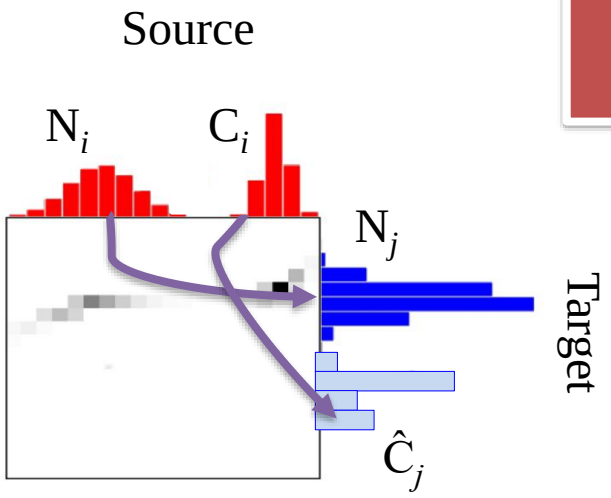
By comparing N_j (Target) and N_i (i=1…I) (Source) to determine C_j based on the OT algorithm

SE Model Adaptation (Unsupervised)

- DOTN [Lin et al., NeurIPS 2021]



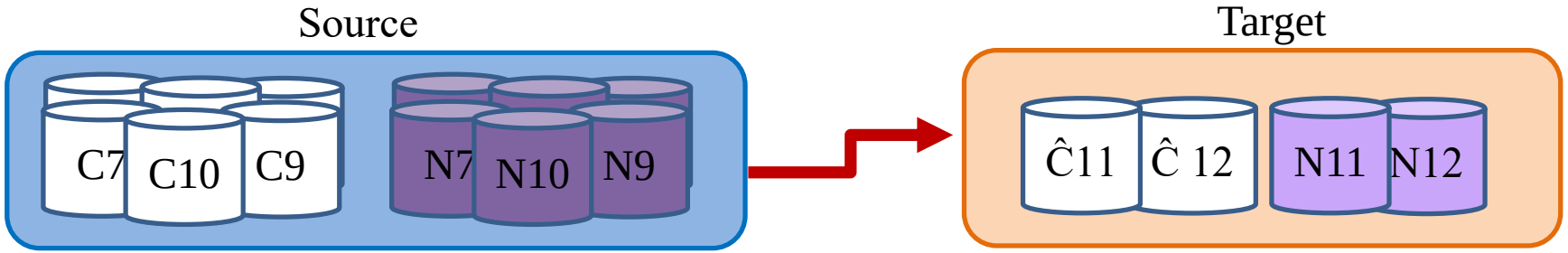


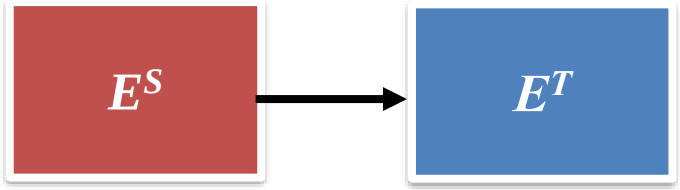


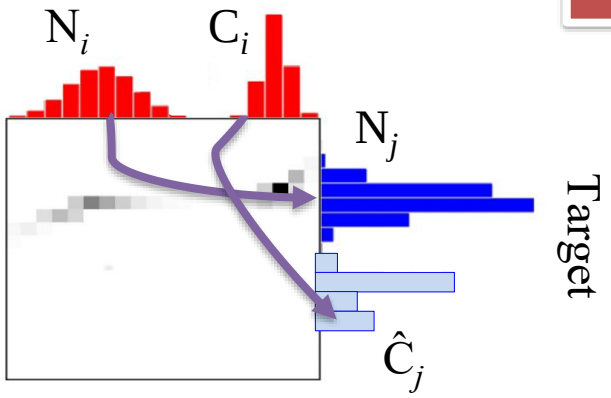
By comparing N_j (Target) and N_i (i=1…I) (Source) to determine C_j based on the OT algorithm

SE Model Adaptation (Unsupervised)

- DOTN [Lin et al., NeurIPS 2021]







By comparing N_j (Target) and N_i (i=1…I) (Source) to determine C_j based on the OT algorithm

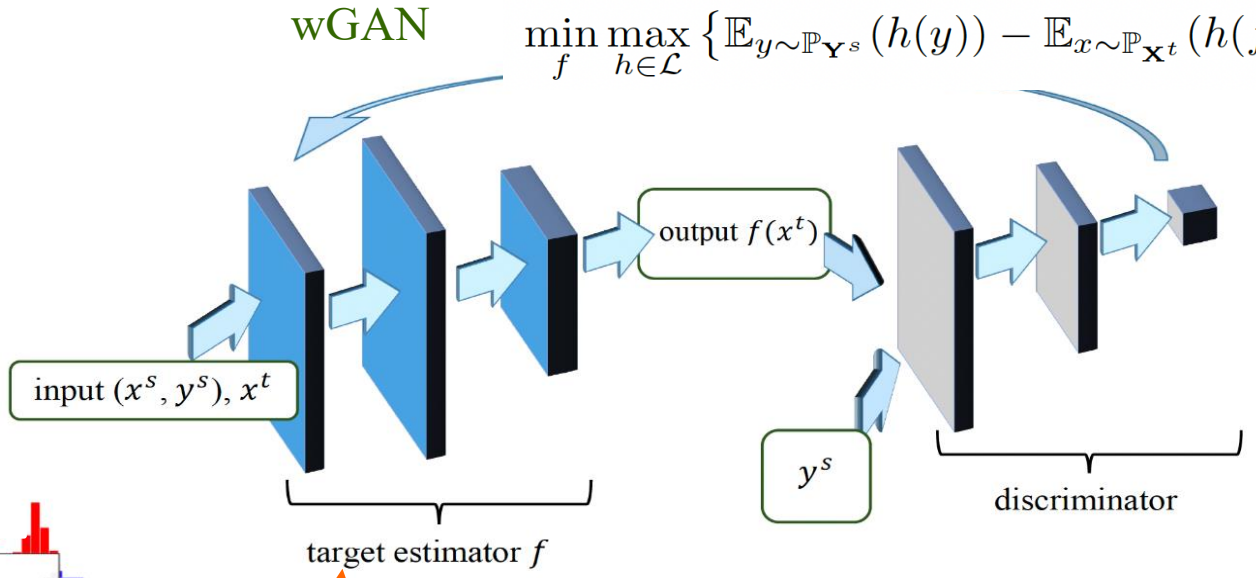
to avoid the forgetting issue

OT (reduce target & source mismatch)

$$\min_{\gamma,f} \mathcal{L} = \min_{\gamma,f} \frac{1}{N^s} \sum_i \|\mathbf{y}_i^s - f(\mathbf{x}_i^s)\|^2 + \sum_{i,j} \gamma_{ij} \left(\alpha \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \beta \|\mathbf{y}_i^s - f(\mathbf{x}_j^t)\|^2\right)$$

SE Model Adaptation (Unsupervised)

- DOTN [Lin et al., NeurIPS 2021]



$$\min_{\gamma, f} \mathcal{L} = \min_{\gamma, f} \frac{1}{N^s} \sum \|\mathbf{y}_i^s - f(\mathbf{x}_i^s)\|^2 + \sum_{i,j} \gamma_{ij} \left(\alpha \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \beta \|\mathbf{y}_i^s - f(\mathbf{x}_j^t)\|^2\right)$$

to avoid the forgetting issue

OT (reduce target & source mismatch)

SE Model Adaptation (Unsupervised)

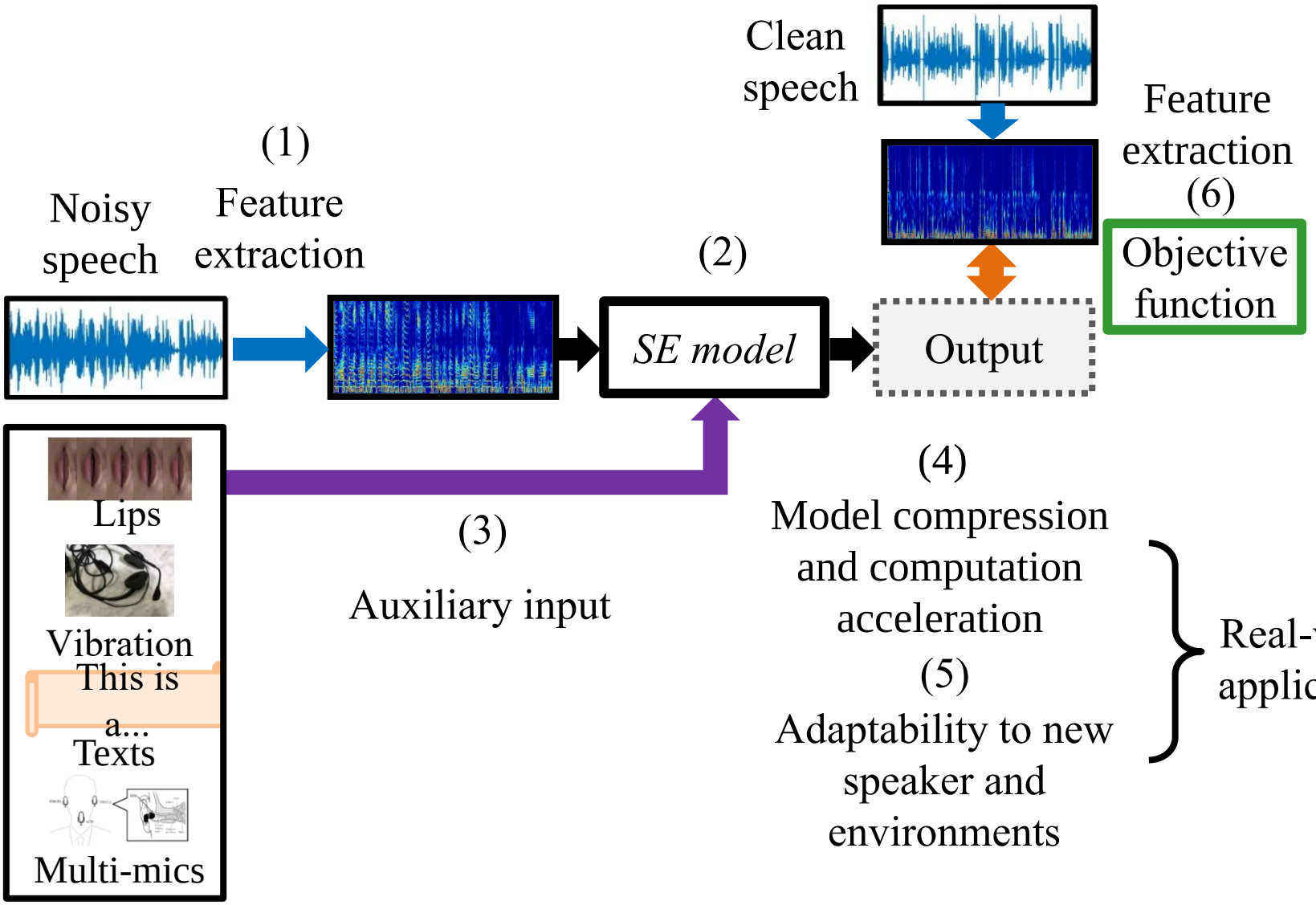
- DOTN [Lin et al., NeurIPS 2021]

	DAT [19]		MDAN [44]		DOTN		DAT [19]		MDAN [44]		DOTN	
Noise type	Helicopter						Crowd-party					
SNR(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-9	1.031	0.392	1.252	0.517	**1.455**	**0.577**	**1.483**	**0.544**	1.150	0.440	1.056	0.451
-6	1.015	0.431	1.443	0.594	**1.669**	**0.649**	**1.484**	**0.560**	1.356	0.516	1.302	0.538
-3	1.094	0.497	1.664	0.670	**1.890**	**0.716**	1.528	0.592	**1.560**	0.600	1.559	**0.621**
0	1.268	0.566	1.902	0.742	**2.104**	**0.775**	1.596	0.636	1.776	0.684	**1.816**	**0.709**
3	1.518	0.637	2.134	0.801	**2.289**	**0.822**	1.736	0.690	1.986	0.762	**2.042**	**0.782**
6	1.779	0.701	2.363	0.849	**2.497**	**0.865**	1.953	0.750	2.179	0.823	**2.236**	**0.838**
9	2.094	0.759	2.563	0.884	**2.677**	**0.895**	2.200	0.801	2.355	0.865	**2.447**	**0.885**
Avg	1.400	0.569	1.903	0.722	**2.083**	**0.757**	1.711	0.653	1.766	0.670	**1.780**	**0.690**
	DAT [19]		MDAN [44]		DOTN		DAT [19]		MDAN [44]		DOTN	
Noise type	Cafeteria						Baby-cry					
SNR(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-9	1.196	0.440	**1.248**	**0.471**	1.185	0.436	1.109	0.580	**1.116**	0.561	0.984	**0.597**
-6	1.206	0.471	1.395	**0.535**	**1.419**	0.528	**1.313**	0.630	**1.313**	0.636	1.180	**0.668**
-3	1.244	0.519	1.608	0.614	**1.631**	**0.620**	1.495	0.665	**1.500**	0.699	1.431	**0.732**
0	1.408	0.579	1.827	0.692	**1.859**	**0.699**	1.621	0.707	**1.734**	0.768	1.665	**0.789**
3	1.631	0.651	2.031	0.764	**2.075**	**0.769**	1.801	0.746	**1.909**	0.809	1.889	**0.835**
6	1.915	0.719	2.244	0.822	**2.271**	**0.826**	1.973	0.782	**2.112**	0.849	2.105	**0.871**
9	2.216	0.782	2.422	0.864	**2.458**	**0.873**	2.133	0.814	2.274	0.880	**2.277**	**0.890**
Avg	1.545	0.594	1.825	**0.680**	**1.843**	0.679	1.635	0.703	**1.708**	0.743	1.647	**0.769**

DAT (Domain Adversarial Training); MDAN (Multisource Domain Adversarial Network [CMU+IST])

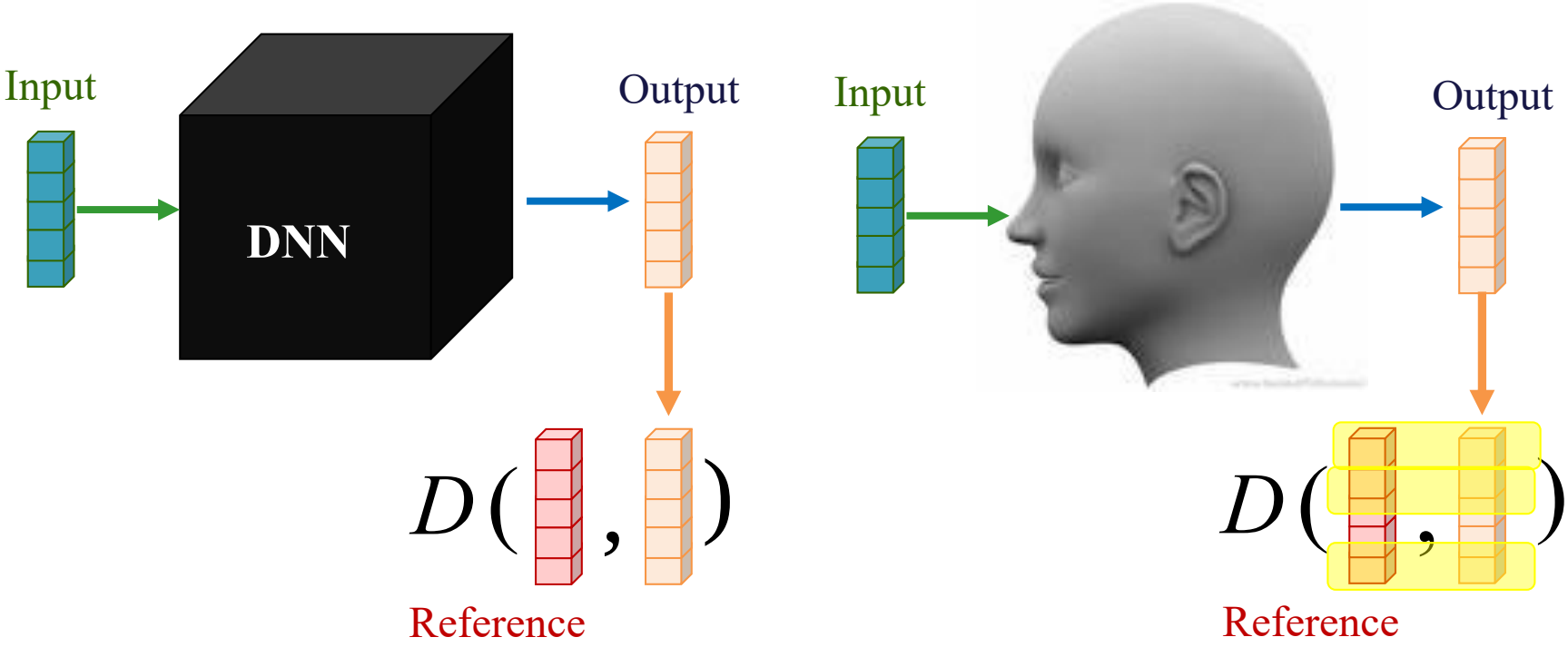
(1) DOTN is the first work that performs unsupervised SE model adaptation.
(2) DOTN yields better performance than existing unsupervised approaches.

Factors of DL-based SE



Objective Functions of DL-based Models

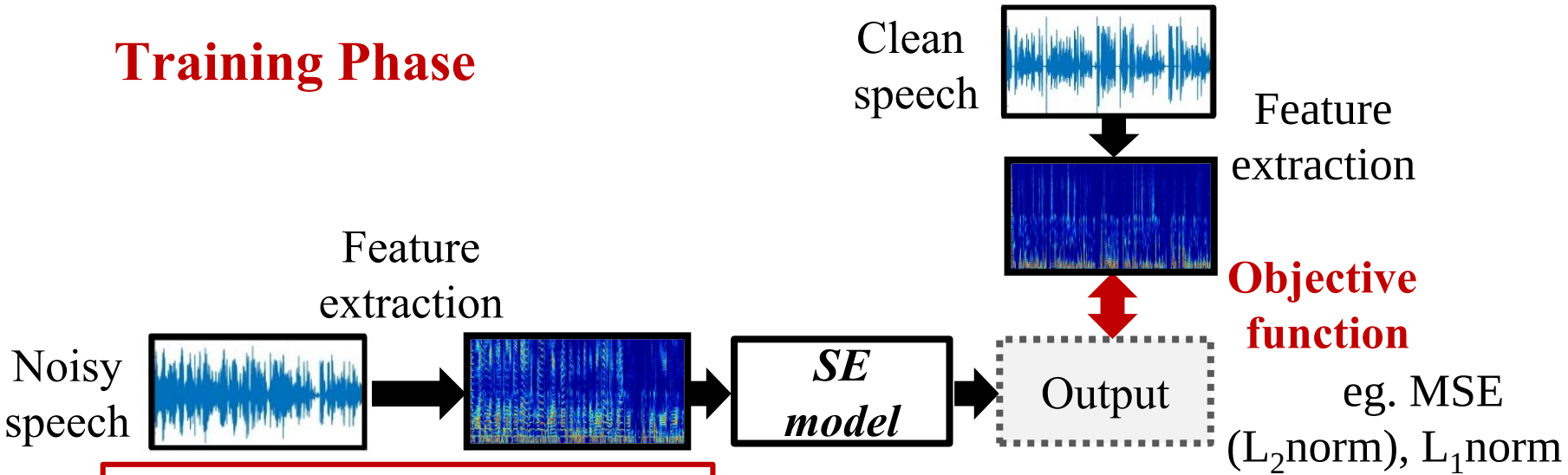
- DNN Model vs. Human Brain
  - ➢ Difficult to fully understand what is inside
  - ➢ What we can control: input, output, objective function

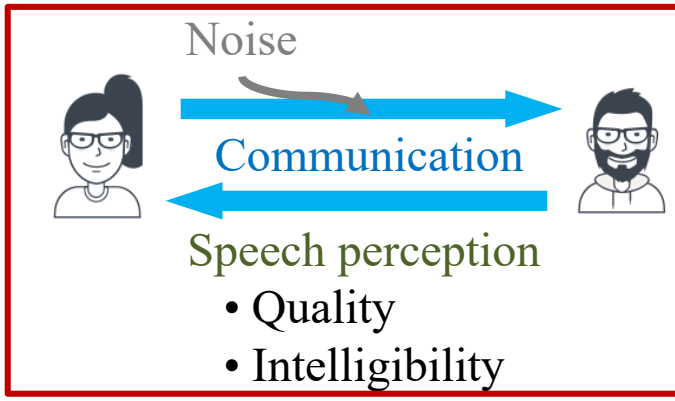


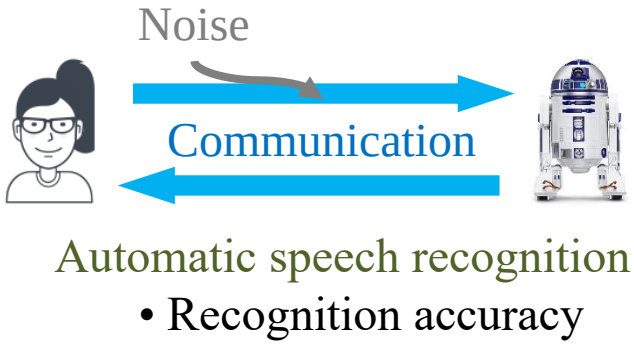
Definite goal → Metric-based objective function

Objective Function

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

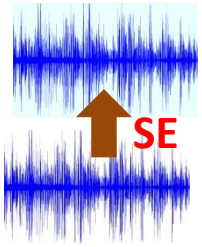
Speech perception

- Quality
- Intelligibility

Quality: 3
Intelligibility: 0.8

Objective Function

- Quality
- Intelligibility

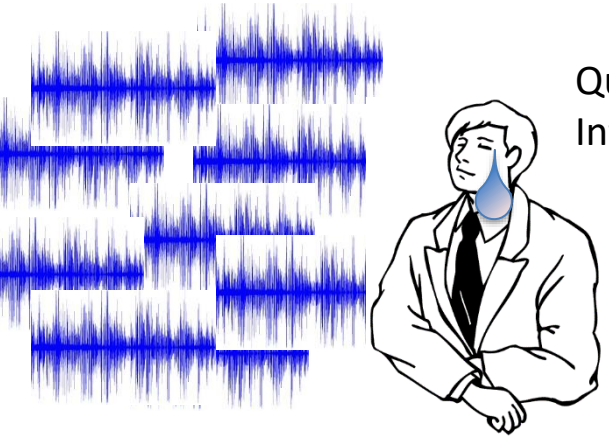


Quality: 5
Intelligibility: 1.0

Quality: 3
Intelligibility: 0.8

Objective Function

- Quality
- Intelligibility



Quality: 3
Intelligibility: 0.8

Quality:3.13
Intelligibility: 0.75
Quality:2.8
Quality:2.5
Quality:16
Intelligibility: 0.89

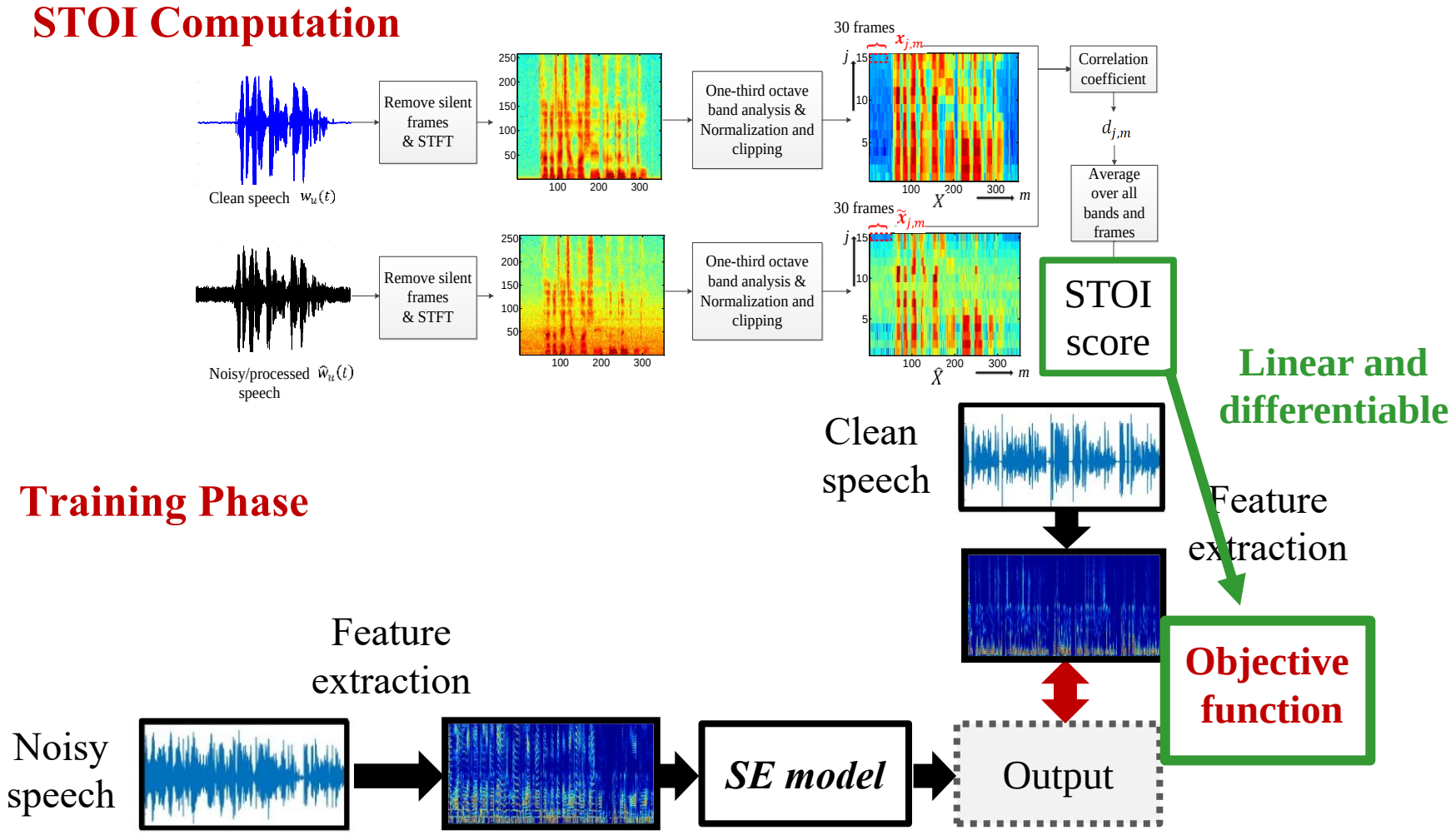
- We derived objective function based on STOI and PESQ.
- We have proposed two solutions: (1) Direct optimization on STOI[(1)];
  (2) Generative adversarial tainting (GAN) to optimize PESQ and STOI[(2)].

➢ "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

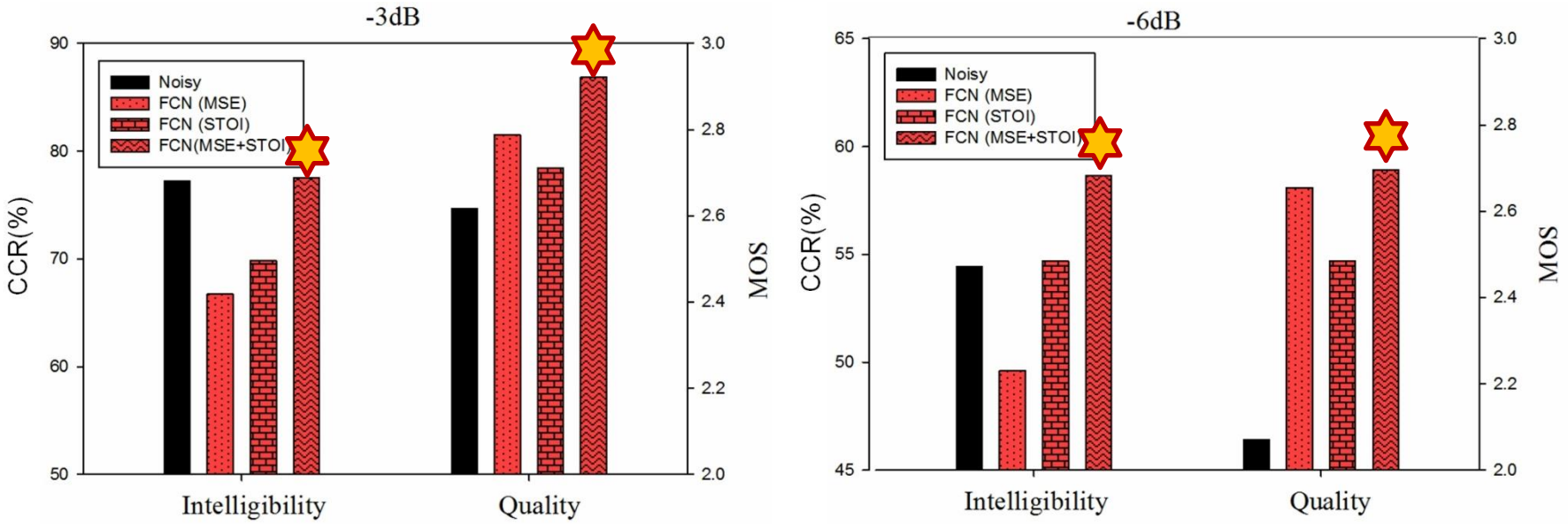
Objective Function (Intelligibility)

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



Objective Function (Intelligibility)

- 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.

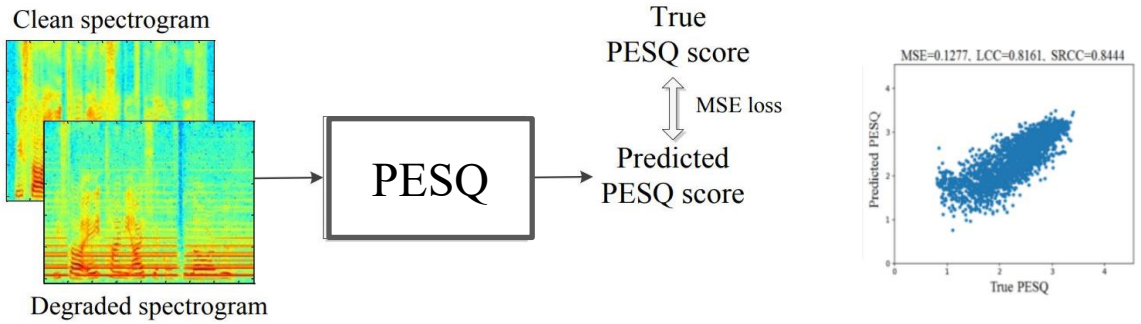
Objective Function (Quality)

- 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].

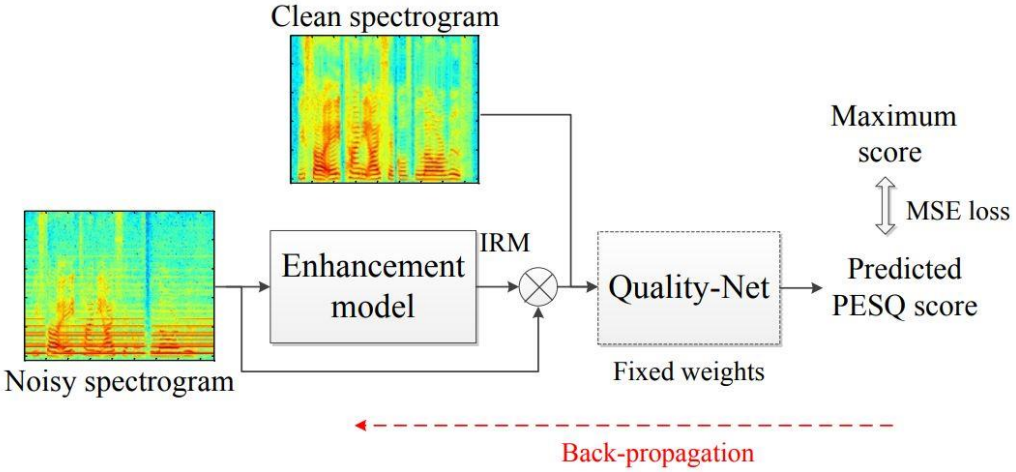
Objective Function (Quality)

- 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)

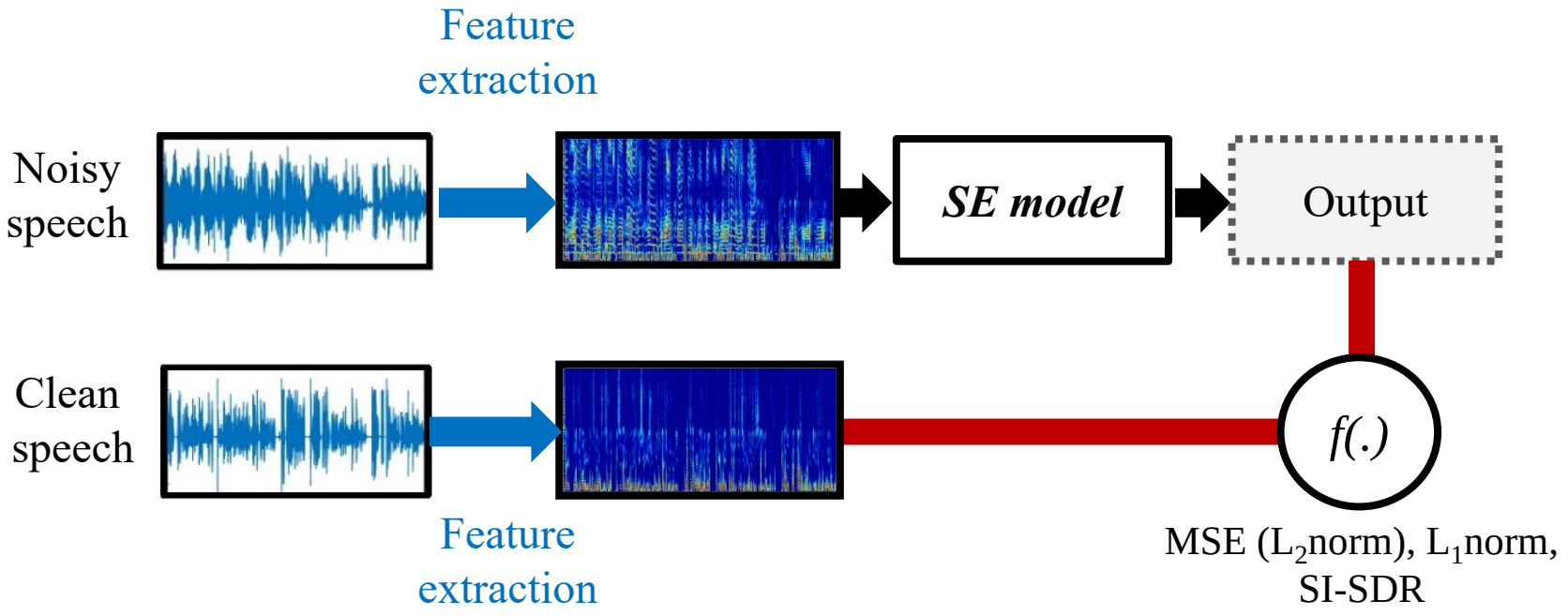


Stage 2: train the SE model based on the Quality-Net (input: paired clean and noisy speech; output: PESQ score)



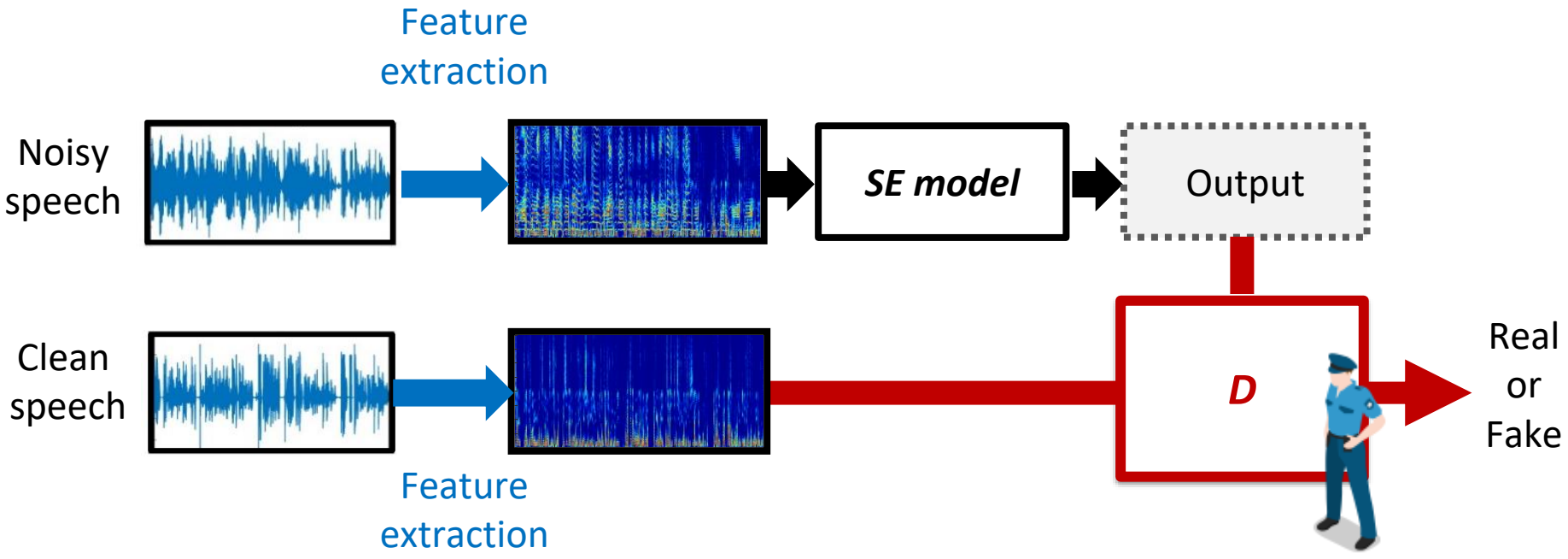
Objective Function (Quality)

- 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]



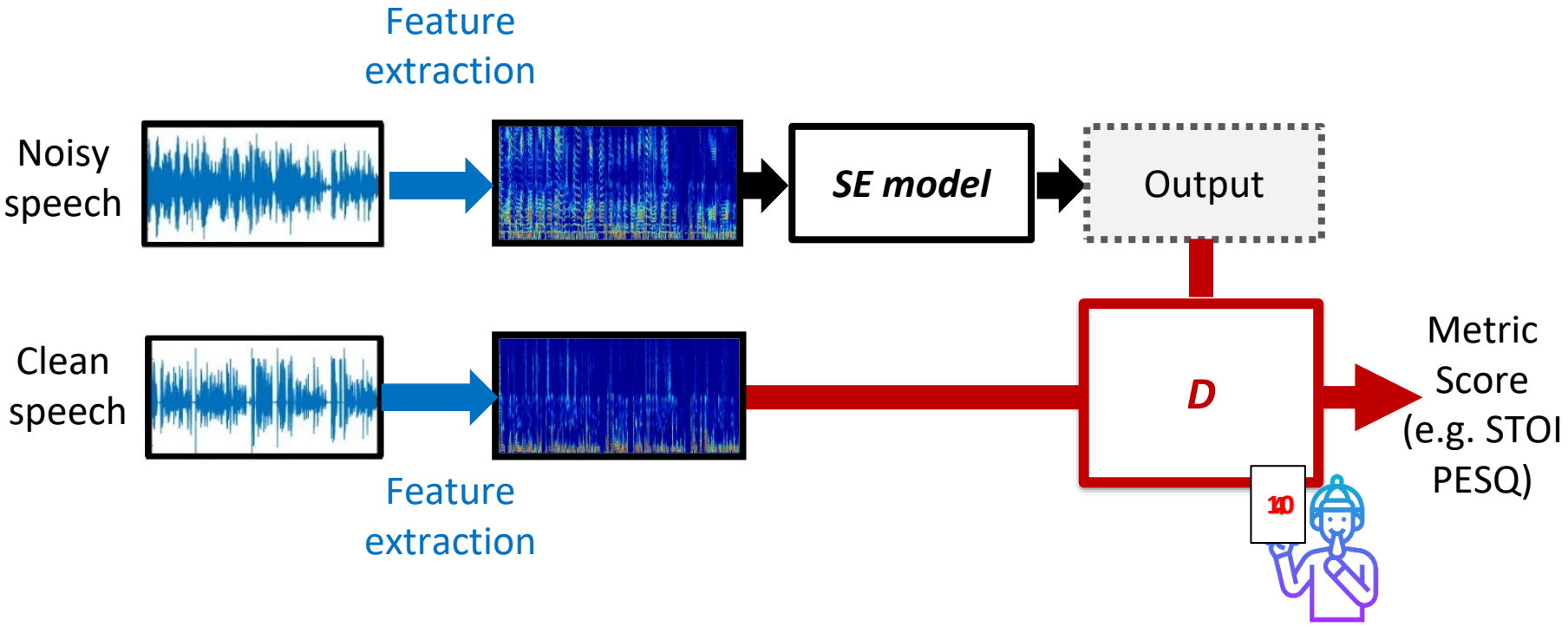
Objective Function (Quality)

- 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 (Quality)

- MetricGAN [Fu et al., ICML 2019]



Objective Function (Quality/Intelligibility)

- 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:

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

$0 \le Q'(G(x),y) < 1$ 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 (Quality/Intelligibility)

- 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:

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

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

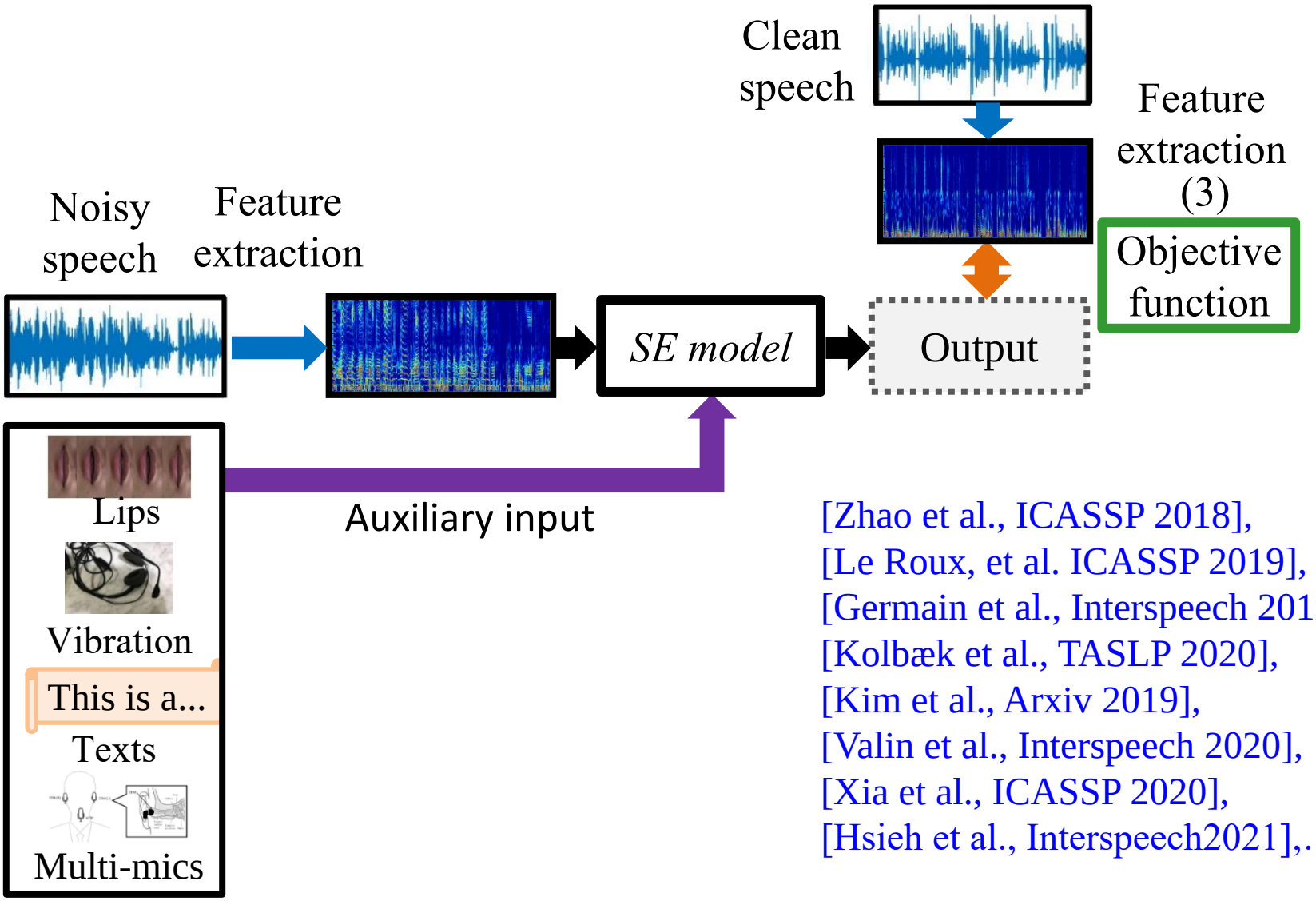
Performance comparisons on TIMIT of different methods in terms of PESQ (quality) & STOI (intelligibility)

	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.

Factors of Deep Learning based SE



[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],
[Hsieh et al., Interspeech2021],…

Outline

1. Background
   - Traditional speech enhancement
   - Deep learning based speech enhancement
   - Goal-oriented speech enhancement
2. Deep learning based speech enhancement in cochlear implants
   - **Intelligent-oriented speech enhancement for CI speech perception**
   - Integration of speech enhancement and visual cues
3. Summary

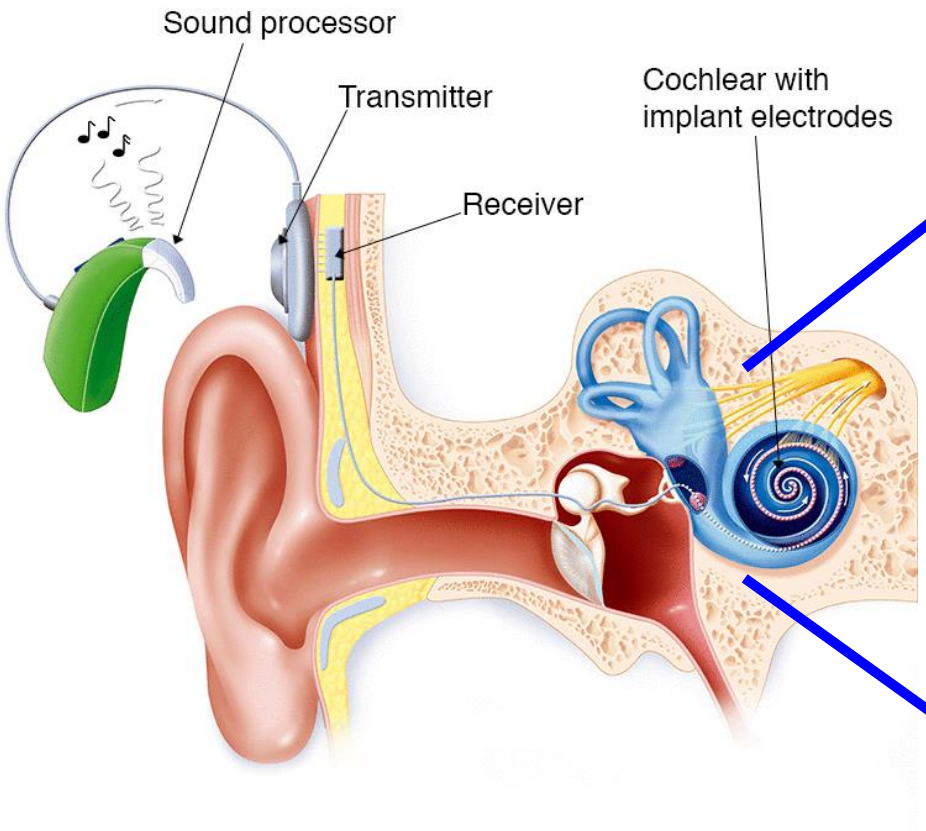
SE for Cochlear Implant

Transmitter coil

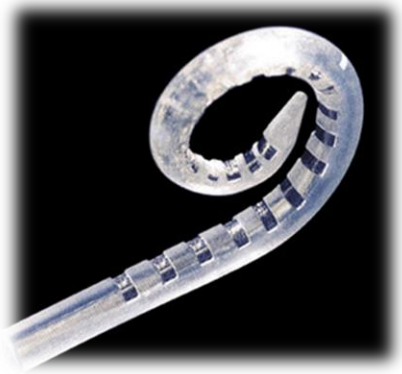
Speech processor:
1. Microphone.
2. DSP chip.
3. Battery
4. Others…

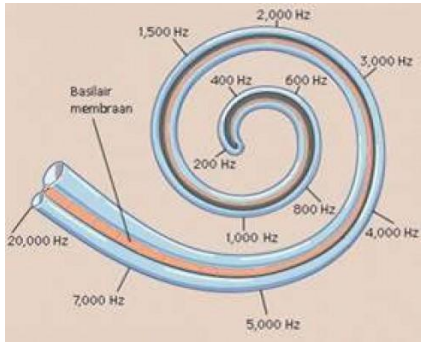
Source from:
https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cochlear-implant-surgery

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

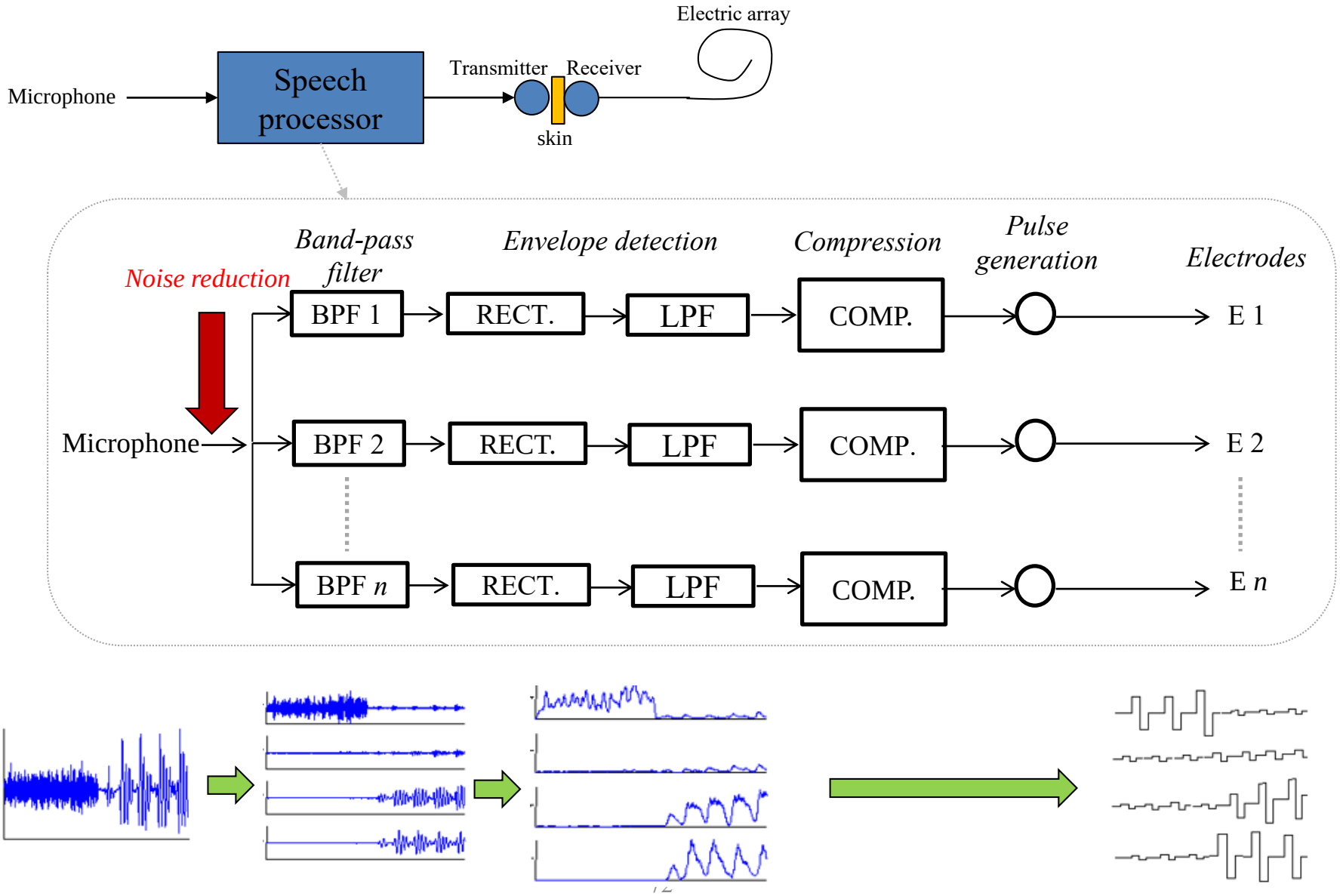
- The tremendous progress of CI technologies in the past three decades has enabled many CI users to enjoy **high level** of speech understanding **in quiet**.
- For most CI users, however, the performance of speech understanding **in noise** still remains **challenging**.

  ➢ 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.

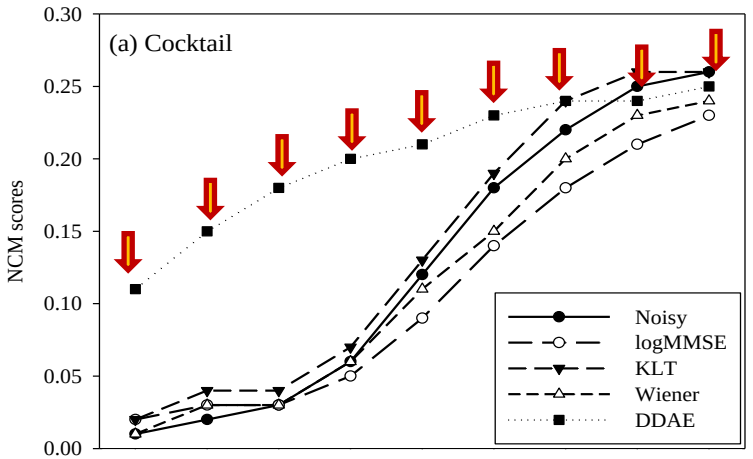
- **Deep learning** based speech enhancement (SE) for CI.



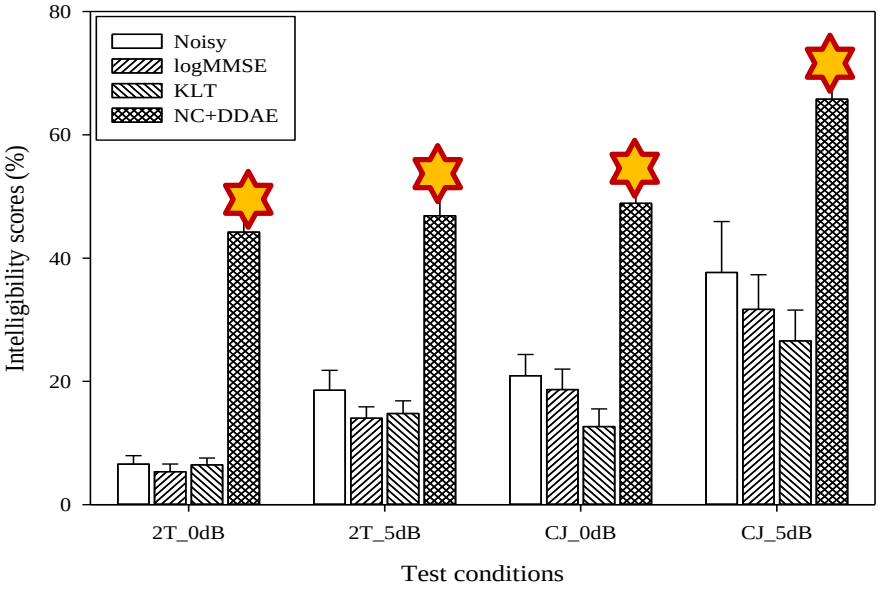
SE for Cochlear Implant



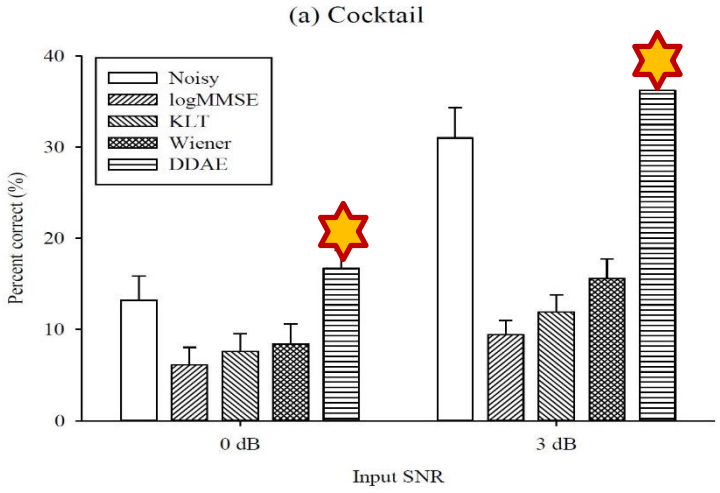
Evaluation Results (Simulations and Subject Tests)



Objective evaluation (NCM)



Clinical trial: 9 CI subjects.



Vocoder results: 10 normal hearing subjects.

(1) DL-based SE outperforms traditional SE approaches in terms of objective evaluations (NCM) and subjective listening tests (CI simulation).

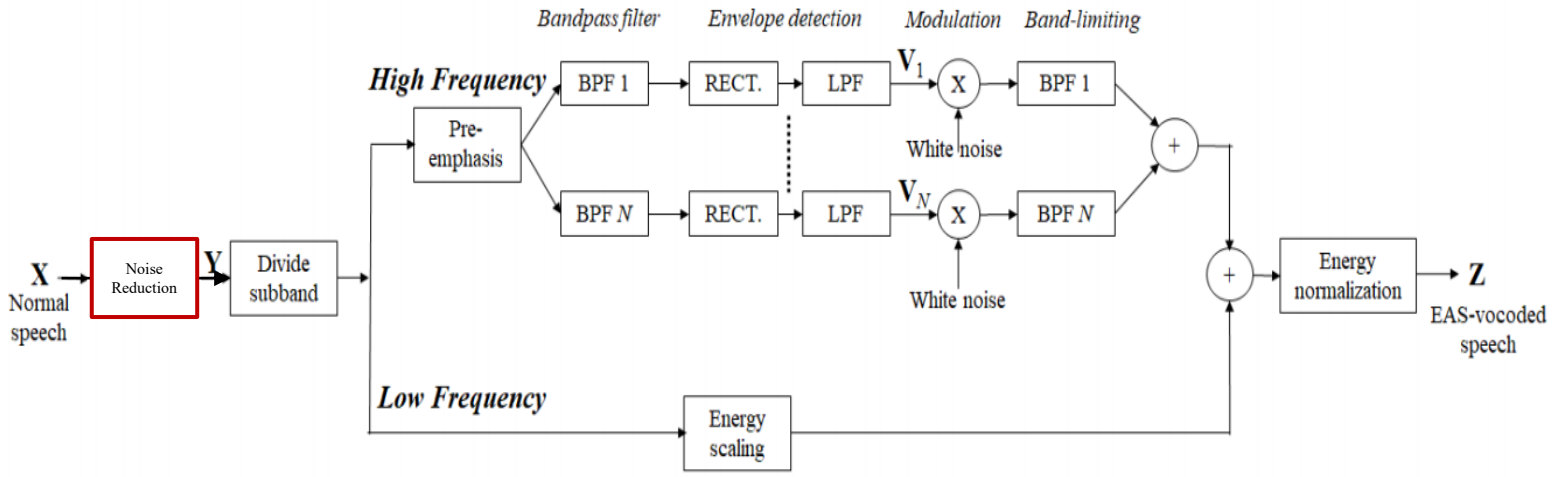
(2) DL-based SE outperforms traditional SE approaches in clinical tests.

Electric and Acoustic Stimulation (EAS)

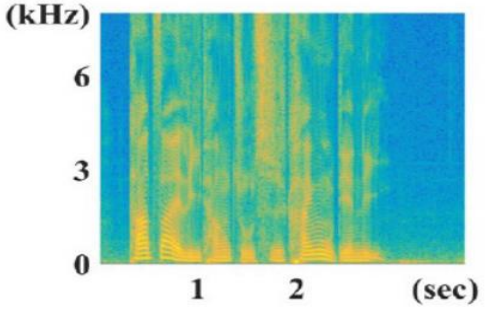
- EAS design and simulation [Wang et al., TNSRE 2020]

  ✓ In EAS, an electrode array is implanted only partially into the cochlea when recipients have residual acoustic at low frequencies

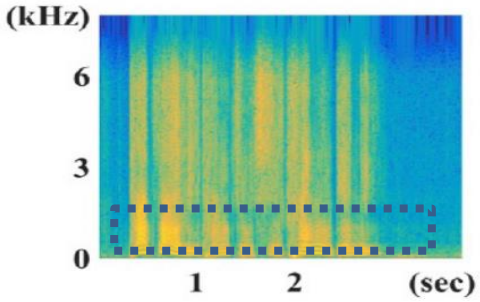
  ✓ We prepare the EAS signals by a hybrid CI and hearing aid simulator



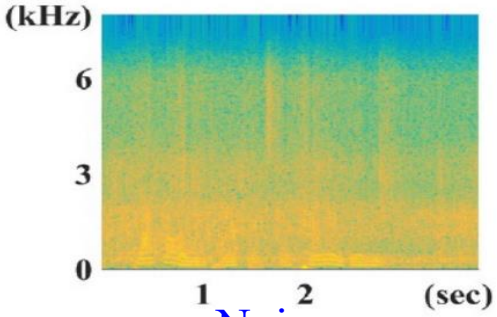
Electric and Acoustic Stimulation (EAS)



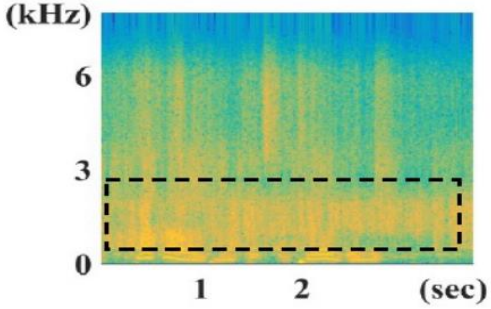
Original Speech



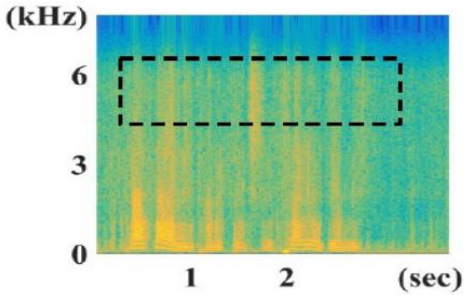
EAS-vocoded Speech



Noisy



MMSE

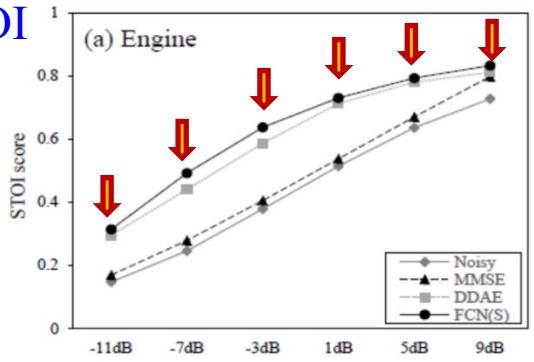


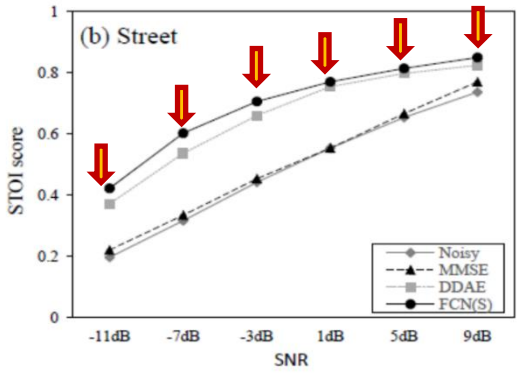
DL-based SE (FCN)

(1) Simulated EAS signals have the same patterns as those of original signals in low frequency regions.
(2) DL-based SE effectively removes noise components and restores speech structures in high frequency regions.

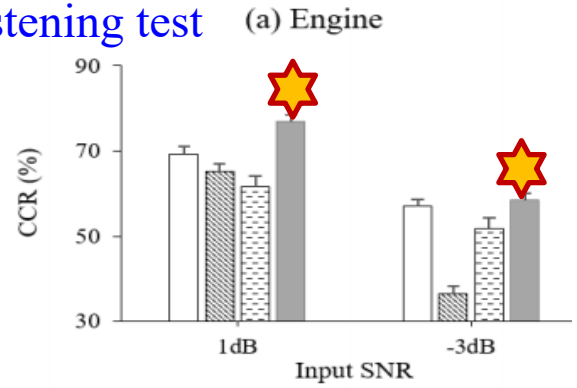
Electric and Acoustic Stimulation

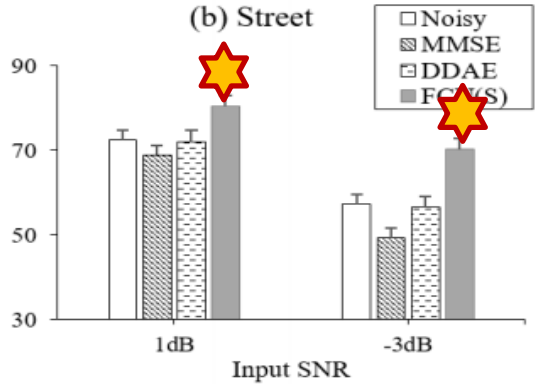
STOI





Listening test





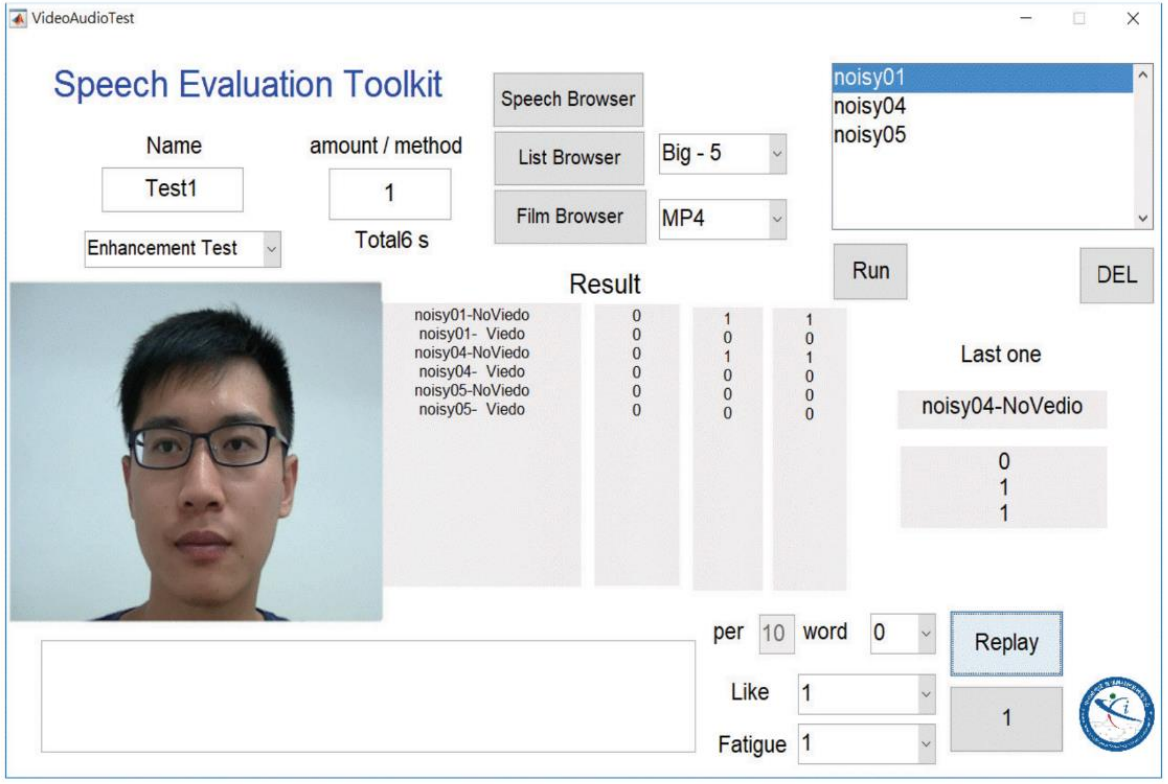
DL-based SE (FCN with STOI loss) outperforms a traditional SE approach (SE) and the DL-based SE (DDAE) approach in terms of both objective evolutions and subjective listening tests.

Outline

1. Background
   - Traditional speech enhancement
   - Deep learning based speech enhancement
   - Goal-oriented speech enhancement
2. Deep learning based speech enhancement in cochlear implants
   - Intelligent-oriented speech enhancement for CI speech perception
   - **Integration of speech enhancement and visual cues**
3. Summary

Audio-visual Integration and SE for CI

- Subjective listening tests with CI simulations [Tseng et al., TCDS 2020]
  - ✓ There are six testing modes: (1) audio only (2) audio-visual. Three types of audio signals: clean, noisy (1dB and 4 dB SNRs),and enhanced signals

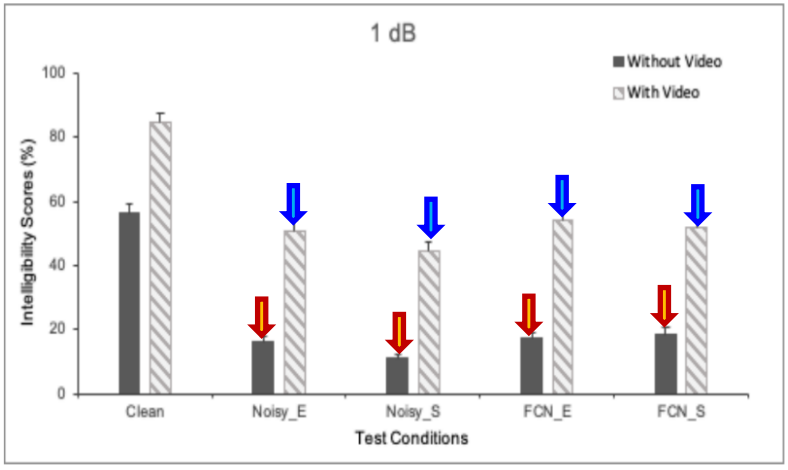


Experimental toolkit: https://github.com/JasonSWFu/VideoAudio

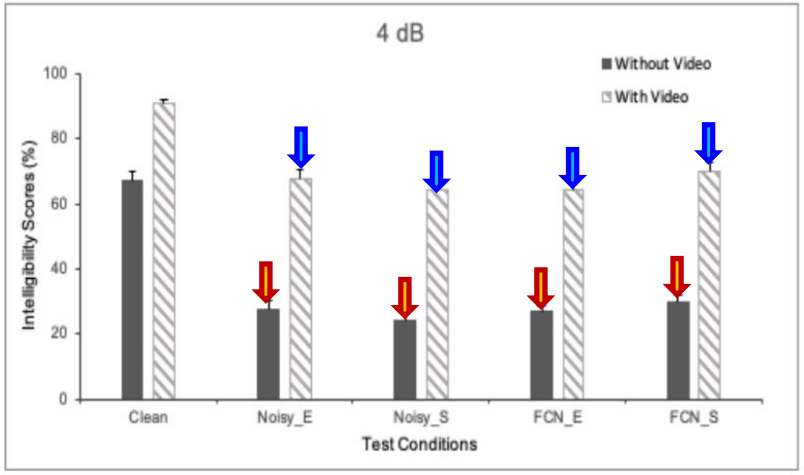
Audio-visual Integration and SE for CI

- Subjective listening tests with CI simulations [Tseng et al., TCDS 2020]

  ✓ There are six testing modes: (1) audio only (2) audio-visual. Three types of audio signals: clean, noisy (1dB and 4 dB SNRs),and enhanced signals



(a) 1 dB



(b) 4 dB

(1) DL-based SE improves speech intelligibility over different noise types (engine and street) and SNR levels (1 dB and 4 dB).
(2) Integration of visual cues consistently improves speech intelligibility over different conditions.
(3) Visual cues can be a key components in assistive hearing devices.

Summary

1. DL-based SE
   - **Notable performance improvements** have been made as compared to traditional (not DL-based) SE approaches
   - Given a target task, **metric-oriented** objective function can make DL-based SE yield optimal results
2. DL-based SE for CI and EAS applications
   - **Significant improvements** have been made by DL-based SE approaches are noted for **CI and EAS**
   - Improved performance is attained by **intelligibility-oriented** objective functions
   - Practical implementation issues to be further addressed (**model size and computation resources**)
3. **Visual cues** can bring considerable improvements for clean, noisy, and enhanced speech

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)

Adv.

The VoiceMOS Challenge

A special session proposal for Interspeech 2022

Wen-Chin Huang[1], Erica Cooper[2], Yu Tsao[3], Hsin-Min Wang[3], Tomoki Toda[1], Junichi Yamagishi[2]
[1]Nagoya University, Japan
[2]National Institute of Informatics, Japan
[3]Academia Sinica, Taiwan

Session Highlights

- Human listening tests are the gold standard for evaluating synthesized speech, and objective measures have low correlation with human ratings. With recent interest in data-driven approaches for mean opinion score (MOS) prediction using machine learning, a challenge for this task to encourage research in this area is timely and important.
- We recently collected a large-scale dataset of MOS ratings for a large variety of text-to-speech and voice conversion systems spanning many years. This challenge releases this data to the public for the first time.
- This challenge also has an out-of-domain track to test generalization ability of submitted systems and to encourage research on supervised and unsupervised adaptation approaches for the MOS prediction task.
- The organizers have experience running challenges such as the Voice Conversion Challenge, ASVspoof, and the Voice Privacy Challenge, and we expect this challenge to attract widespread interest from researchers in speech synthesis, voice conversion, and speech enhancement.
- A challenge would provide a centralized and standardized evaluation for comparing and benchmarking different approaches for the MOS prediction task.

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.
- Defossez, Alexandre, Gabriel Synnaeve, and Yossi Adi. "Real time speech enhancement in the waveform domain." arXiv preprint arXiv:2006.12847, 2020.
- Hu, Yanxin, et al. "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement." arXiv preprint arXiv:2008.00264, 2020.
- Fu, Szu-Wei, et al. "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement." arXiv preprint arXiv:2104.03538せ 2021.

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.

DL-based SE for CI/EAS

- 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, volume 64, number 7, pages 1568 - 1578, July 2017.
- Goehring, Tobias, et al. "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users." Hearing research 344 (2017): 183-194.
- 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, volume 39(4), number 4, pages 795-809, July 2018.
- N. Y.-H. Wang, H.-L. S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Improving the Intelligibility of Speech for Simulated Electric and Acoustic Stimulation Using Fully Convolutional Neural Networks," IEEE Transactions on Neural Systems & Rehabilitation Engineering, volume 29, pages 184-195, December 2020.
- 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," IEEE Transactions on Cognitive and Developmental Systems 2020.

Thank You Very Much for Your Attention