Deep-learning-based Speech Enhancement (with Its Application to Assistive Oral Communication Technologies)

You-Jin Li, Roy Chou

Research Center for Information Technology Innovation

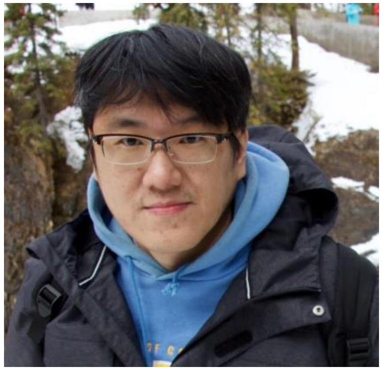
Academia Sinica

PI: Yu Tsao

Bio-ASP

中央研究院
ACADEMIA SINICA

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)

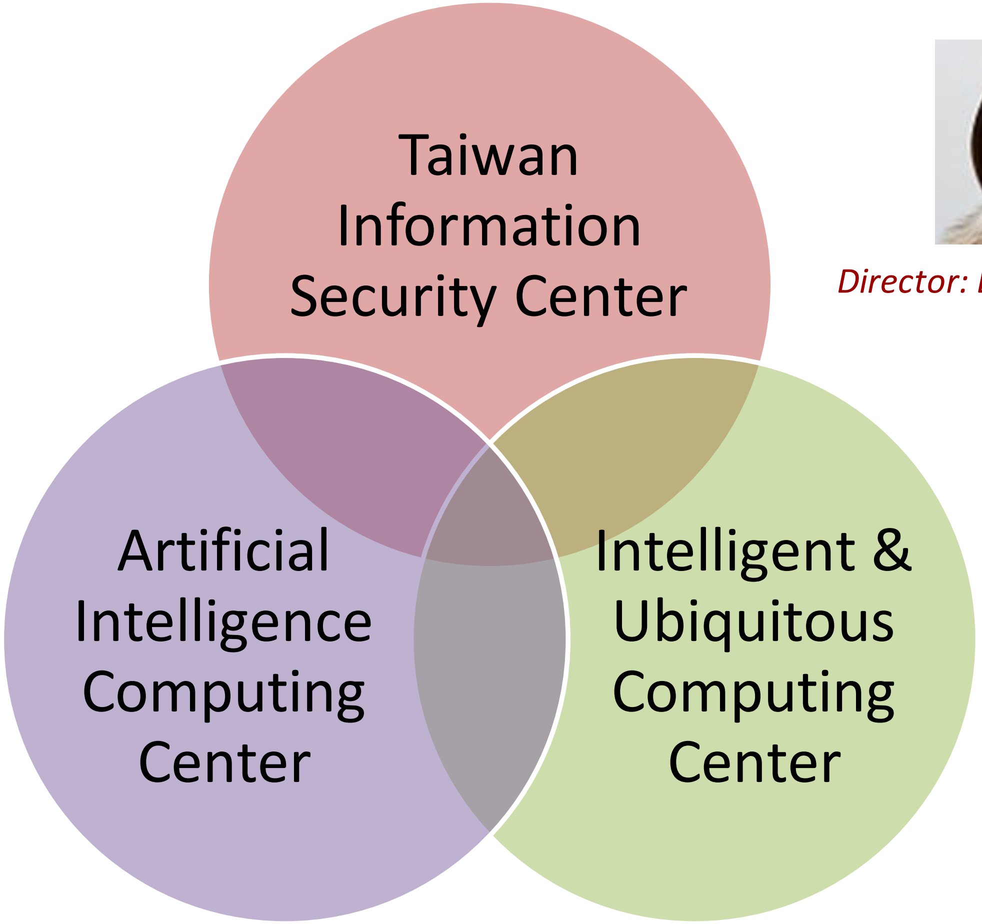
Research Fellow, Deputy Director of CITI, Academia Sinica
Biomedical Acoustic Signal Processing (Bio-ASP) Lab

Bio-ASP

– Research Interests

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

Research Center for Information Technology Innovation (CITI)



Taiwan Information Security Center

Artificial Intelligence Computing Center

Intelligent & Ubiquitous Computing Center

Multimedia (audio, speech, image, and video), mobile communication, security, and FinTech.

Director: Dr. Ai-Chun Pang

First Director: Dr. Ming-Syan Chen
NTU, Vice President

Second Director: Dr. Tei-Wei Kuo
NTU, President

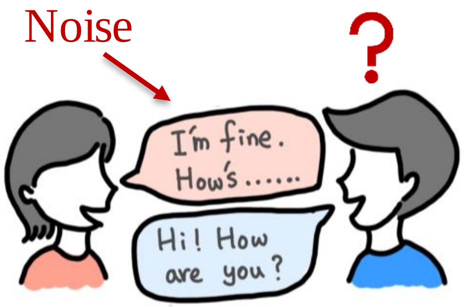
Third Director: Dr. Yennun Huang
MODA, Minister

Outline

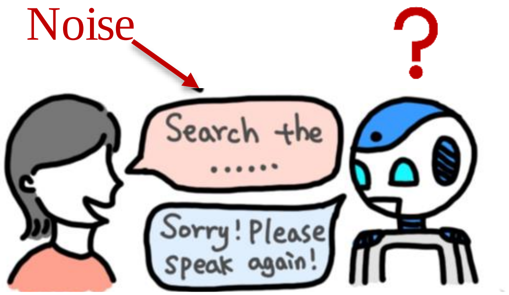
- Deep Learning (DL) based Speech Enhancement (SE)
  - **Basic DL-based SE system architecture**
  - Key factors to the DL-based SE performance
- Assistive Oral Communication Technologies
- Summary

Speech ….Easily Got Distorted

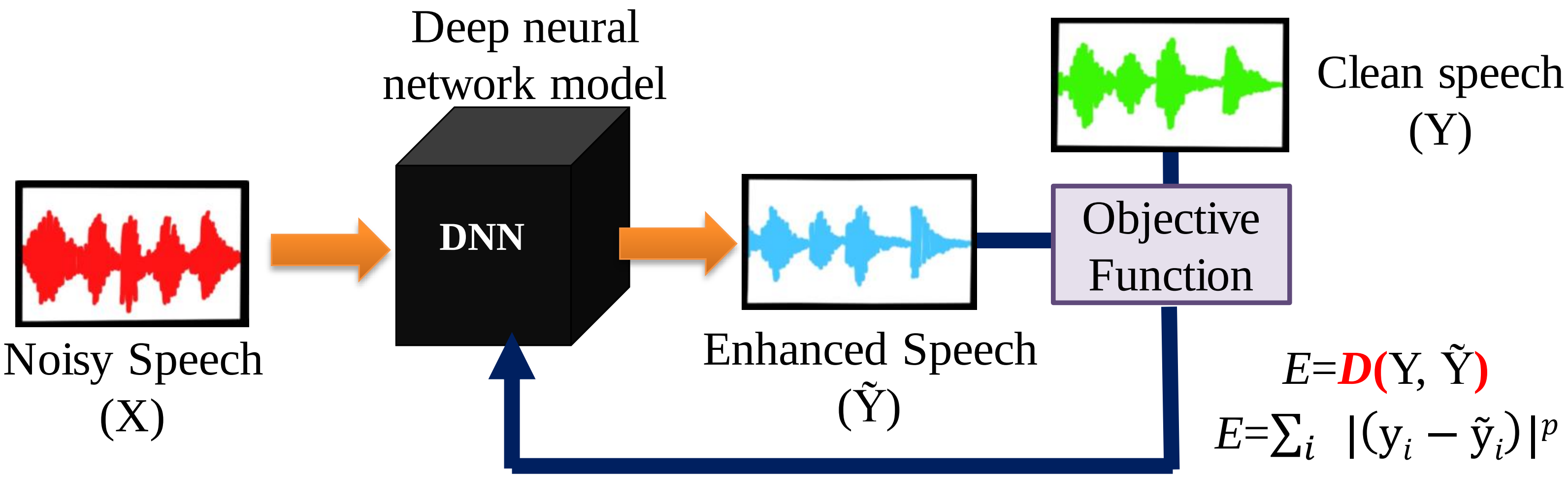
Deep Learning Based SE System



Human to Human
Speech Intelligibility
Speech Quality



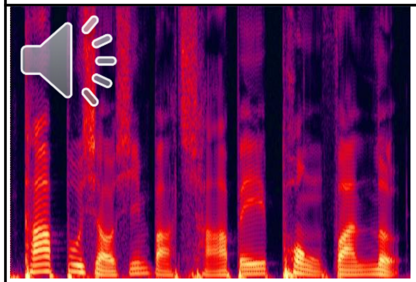
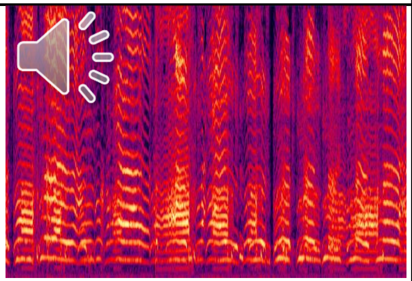
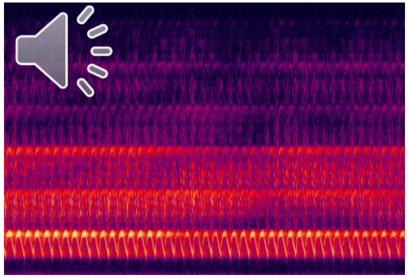
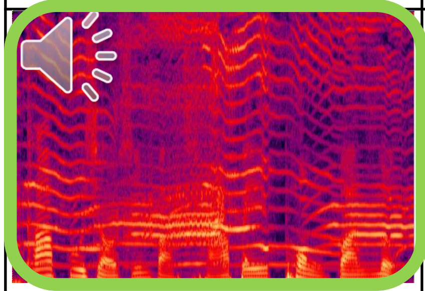
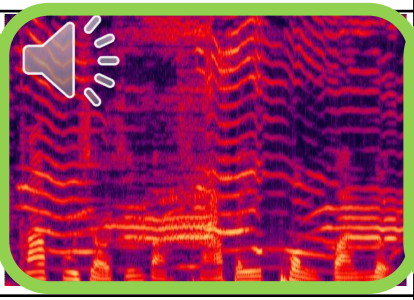
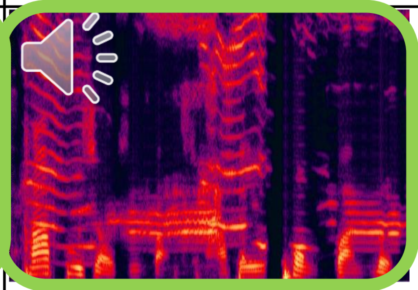
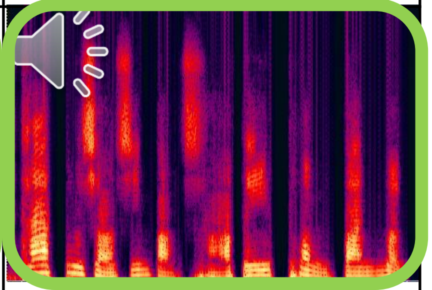
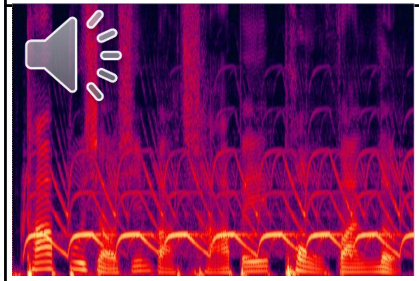
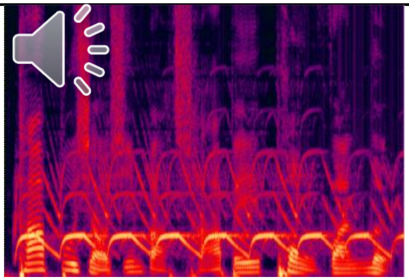
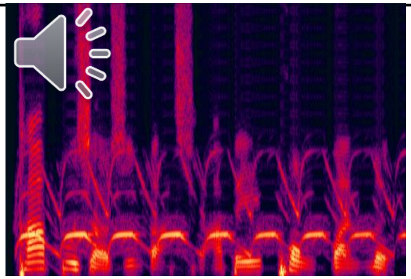
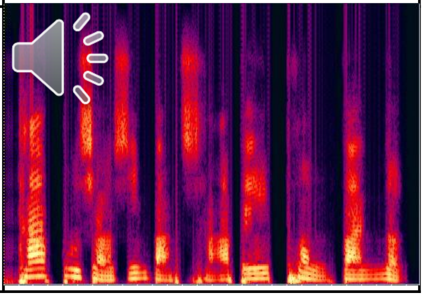
Human to Machine
Speech Intelligibility



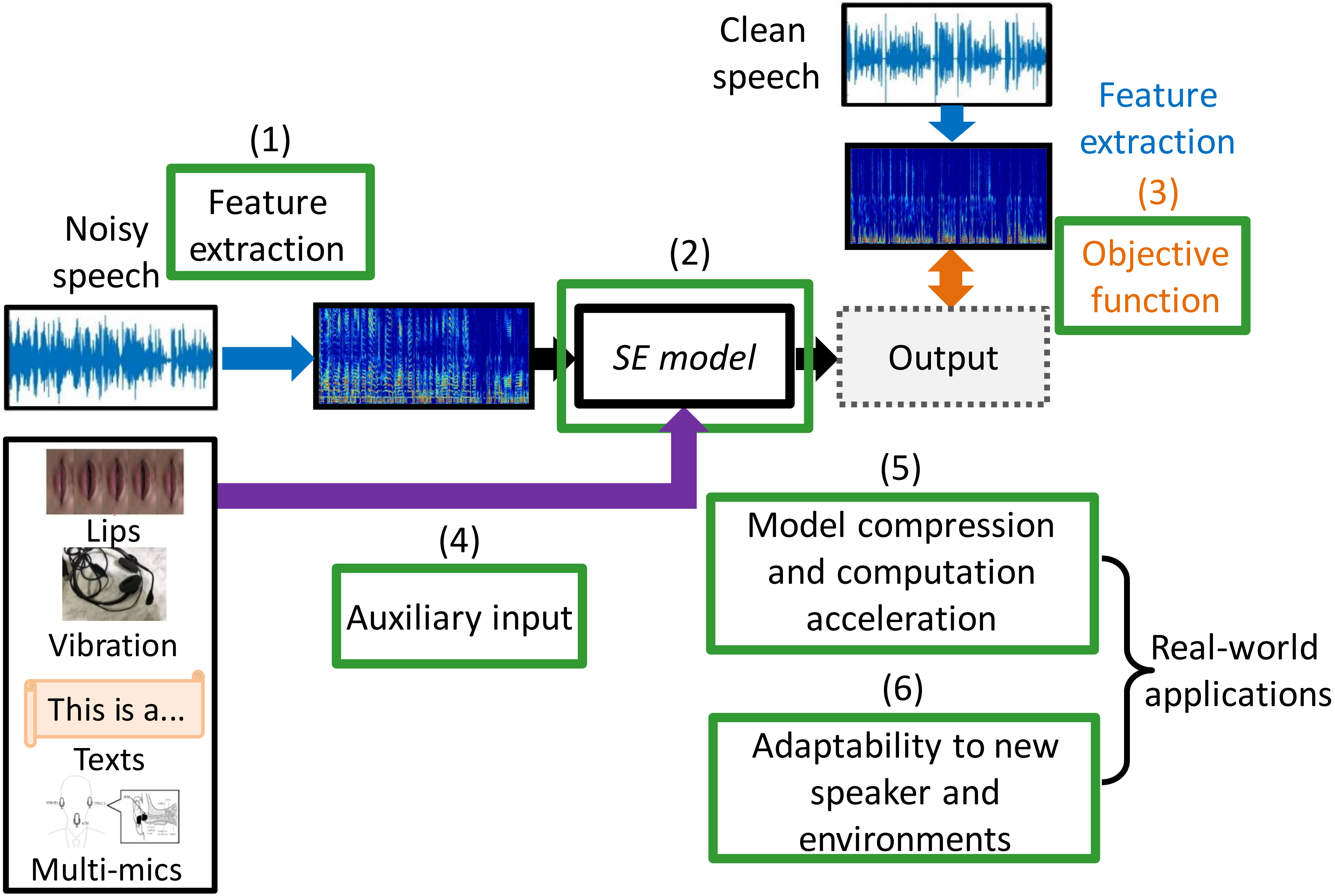
$E=\mathbf{D}(\mathrm{Y}, \tilde{\mathrm{Y}})$

$E=\sum_i |(y_i - \tilde{y}_i)|^p$

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
				

Deep Learning Based SE System



Noisy speech

(1) Feature extraction

(2) *SE model*

Output

Clean speech

Feature extraction

(3) Objective function

Lips

Vibration

This is a...

Texts

Multi-mics

(4) Auxiliary input

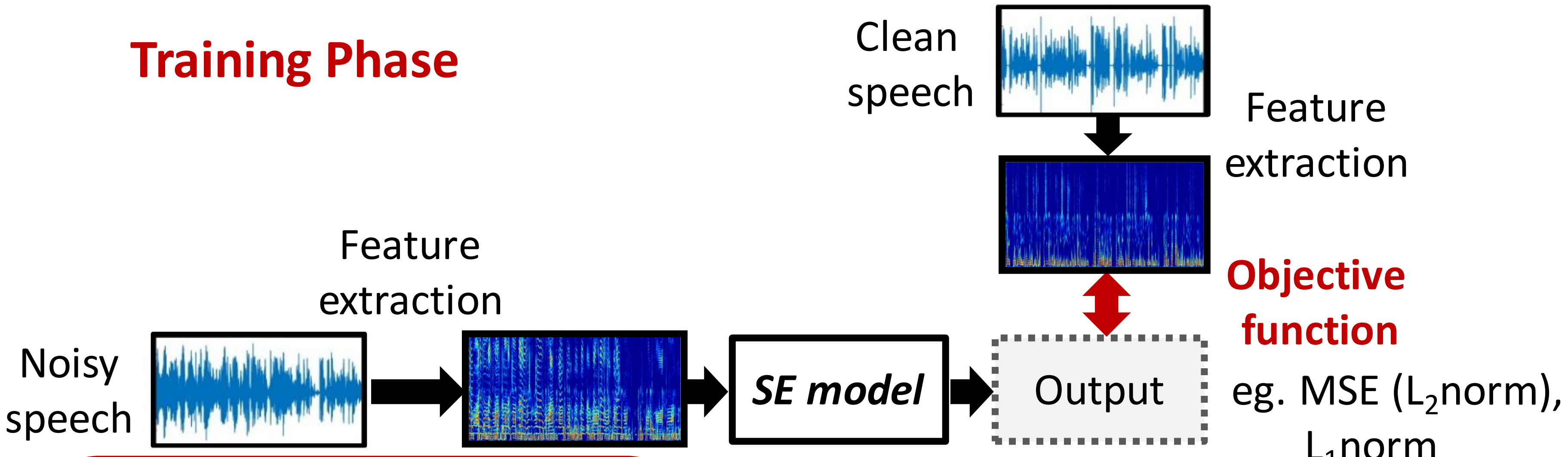
(5) Model compression and computation acceleration

(6) Adaptability to new speaker and environments

Real-world applications

Objective Function

Training Phase



eg. MSE (L_2norm), L_1norm

Speech perception

- Quality
- Intelligibility

Automatic speech recognition

- Recognition accuracy

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.

把噪聲壓抑就好了嗎?

聞

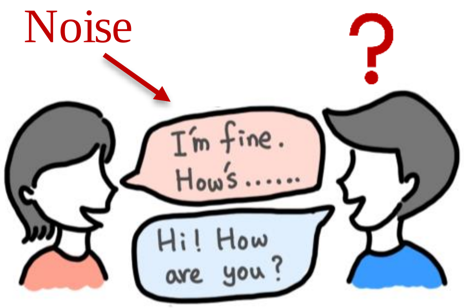
聽

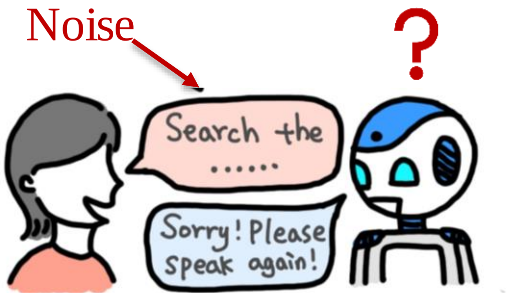
大學曰：心不在焉，聽而不聞

聲音品質(Quality)及理解度 (Intelligibility)並不相同

對口語溝通輔助系統而言，**理解度**比較重要

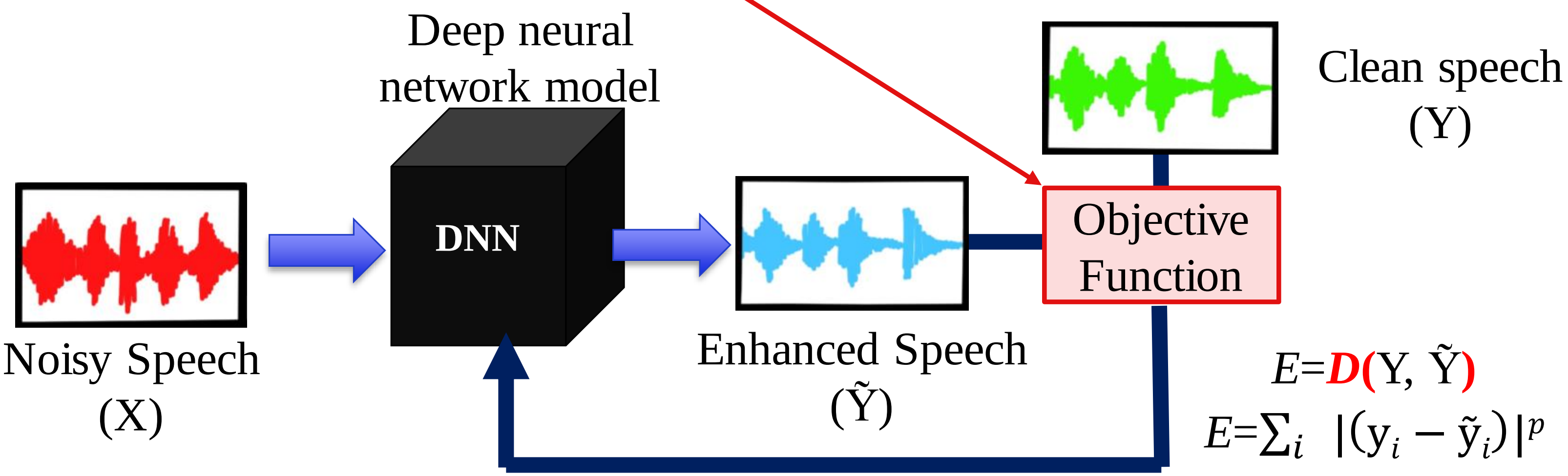
Speech enhancement model based on speech Intelligibility learning





Human to Human
Speech Intelligibility
Speech Quality

Human to Machine
Speech Intelligibility

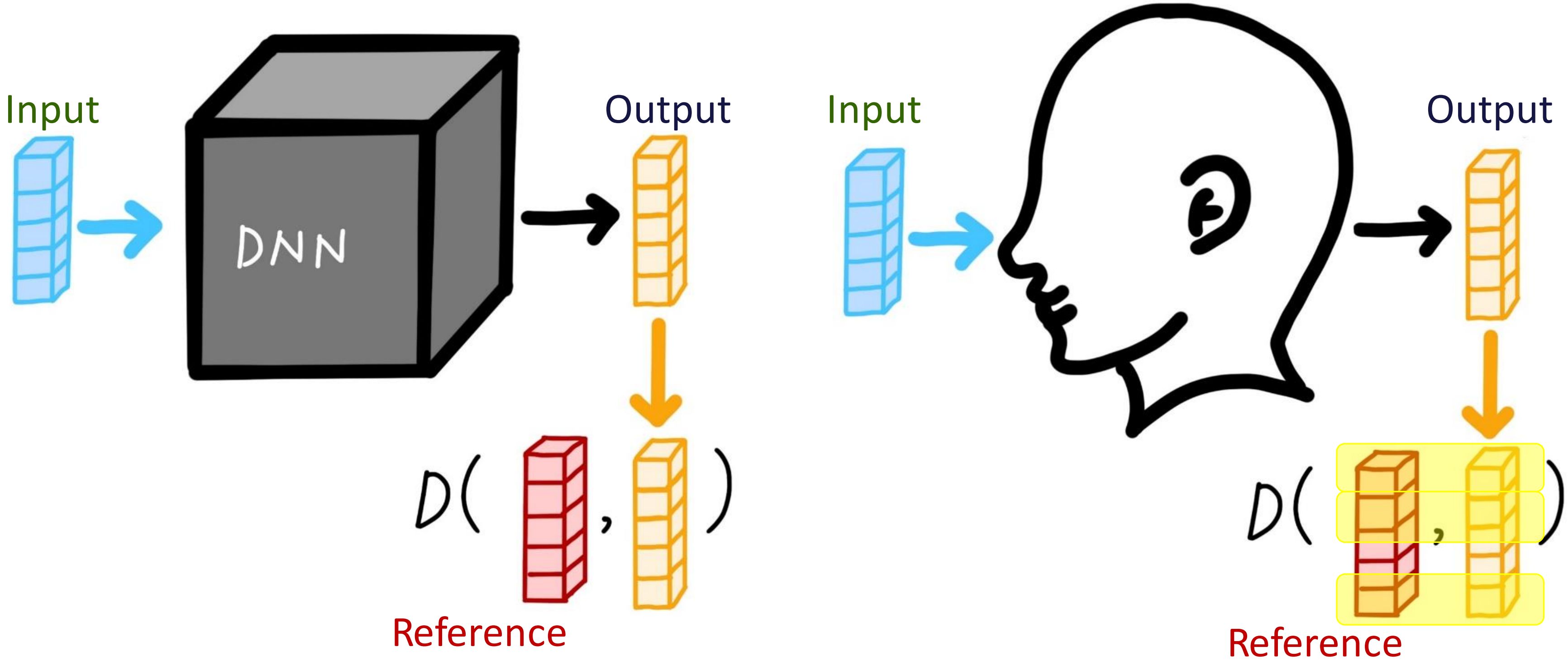


$E = \mathbf{D}(Y, \tilde{Y})$

$E = \sum_i |(y_i - \tilde{y}_i)|^p$

Objective Functions for DNN and Brain

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



Input

Output

Reference

Input

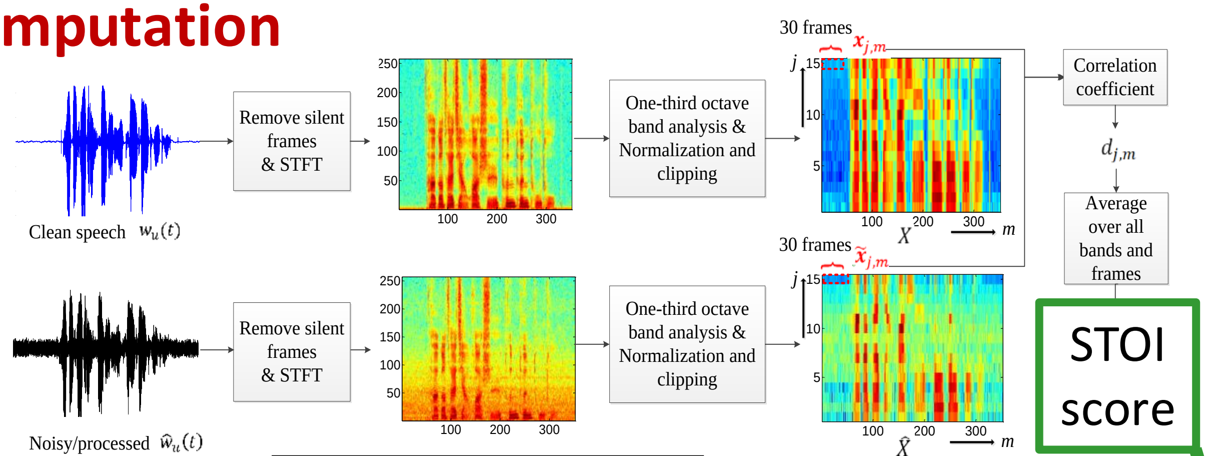
Output

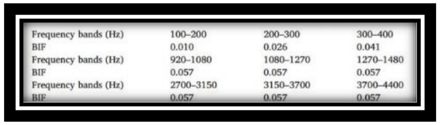
Reference

Objective Function

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

STOI Computation

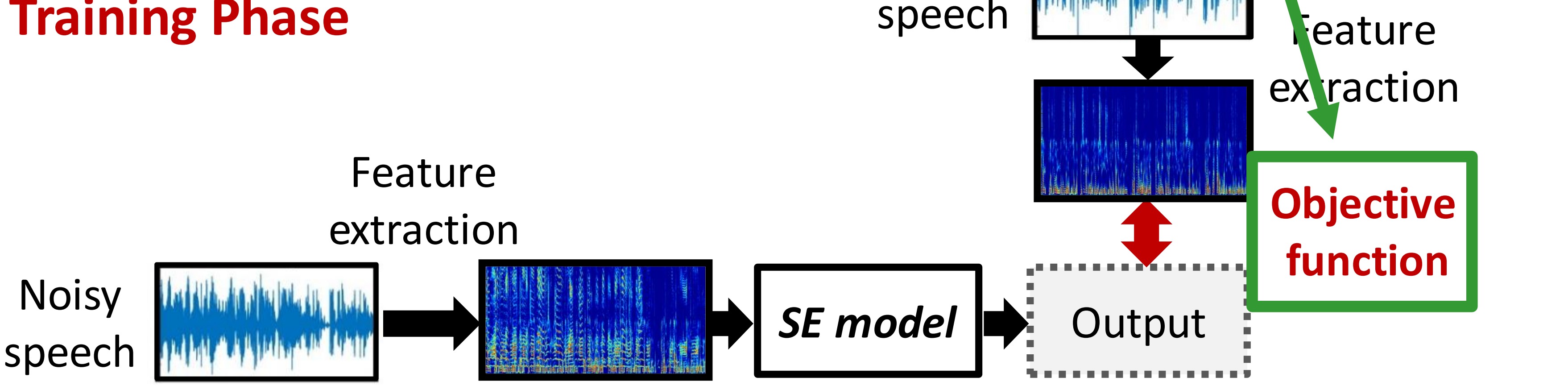

Clean speech $w_u(t)$
Remove silent frames & STFT
250
200
150
100
50
100
200
300
One-third octave band analysis & Normalization and clipping
30 frames
$x_{j,m}$
j
15
10
5
100
200
300
X
m
Correlation coefficient
$d_{j,m}$
Average over all bands and frames
STOI score
Noisy/processed speech $\hat{w}_u(t)$
Remove silent frames & STFT
250
200
150
100
50
100
200
300
One-third octave band analysis & Normalization and clipping
30 frames
$\tilde{x}_{j,m}$
j
15
10
5
100
200
300
$\hat{X}$
m



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

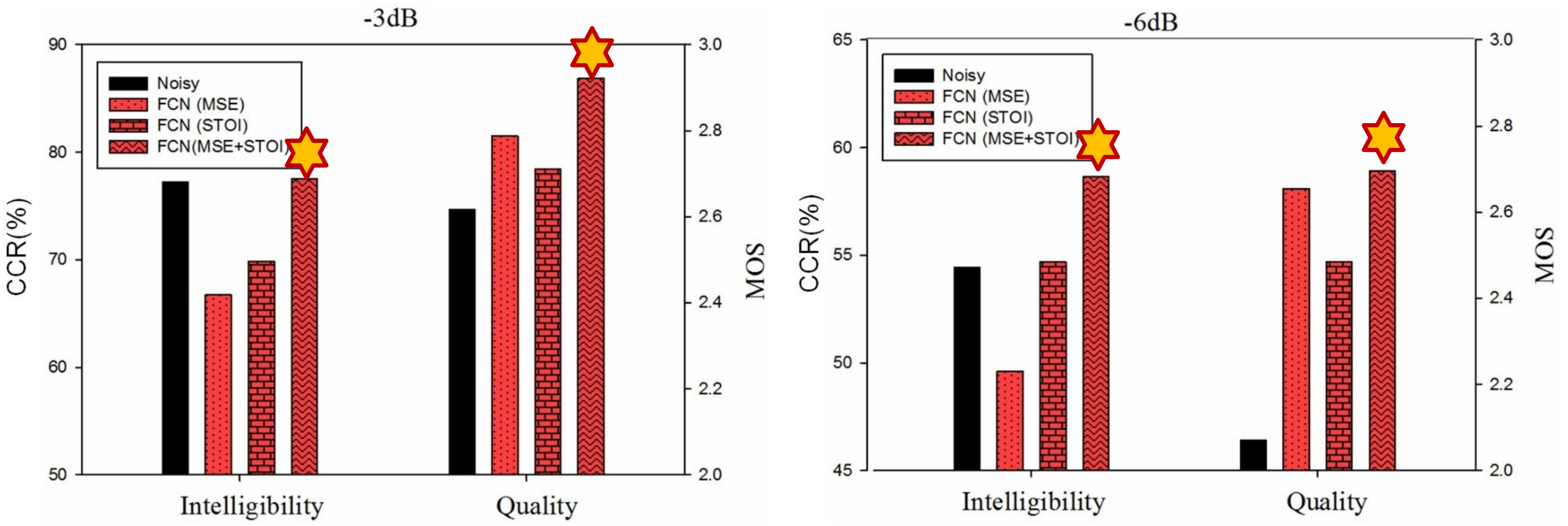
Linear and differentiable

Training Phase



針對聽覺理解度優化的語音增強系統

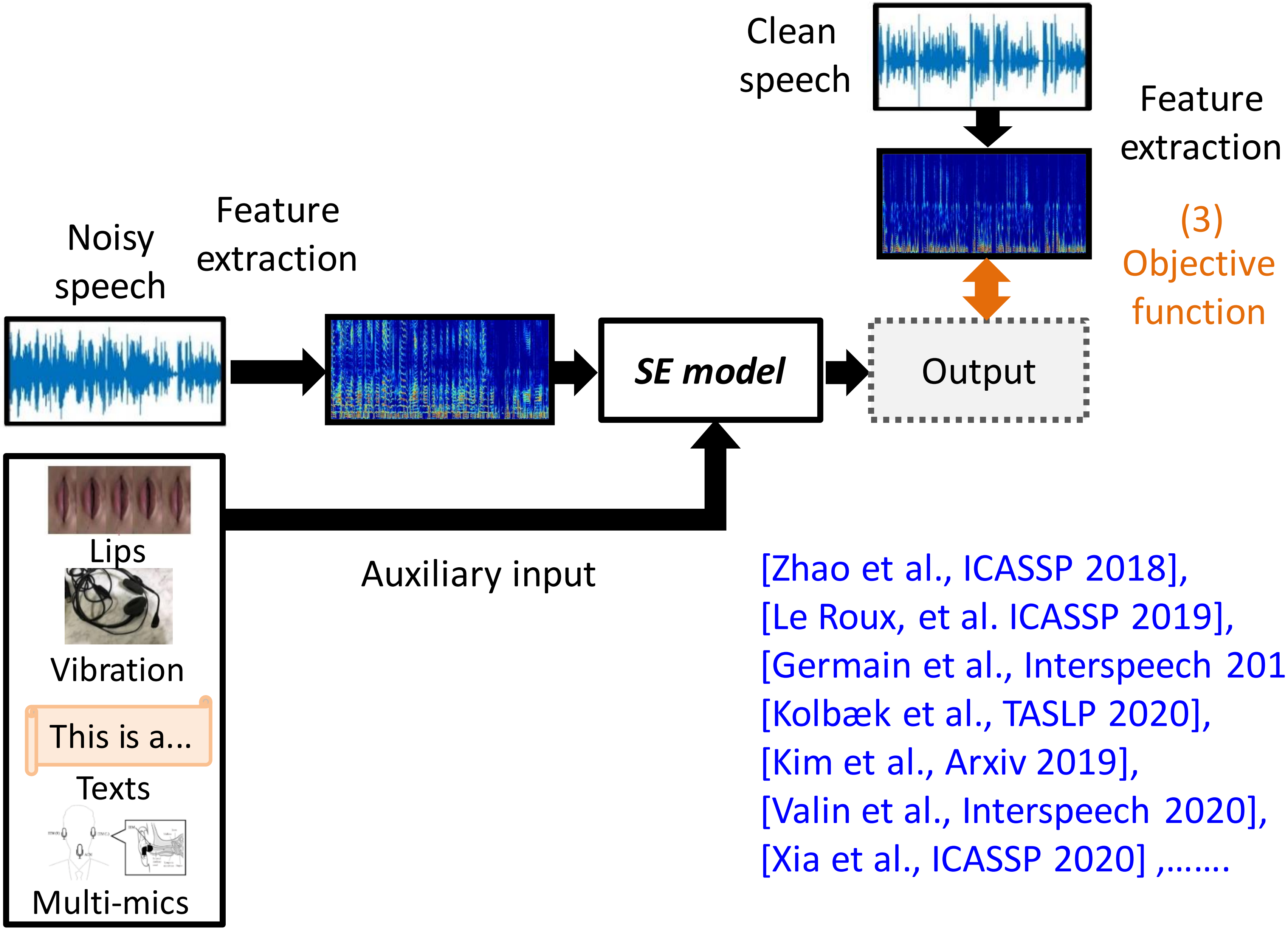
- 實驗結果 (真人聽測)



在 (a) −3 dB and (b) −6 dB SNR 下字辨識率以及聲音品質聽測分數

(1) 相同的DNN，在不同目標函數下，有不一樣的效能。
(2) 基於 (STOI +MSE) 函數的語音增強模型有最佳的效能。
(3) 此研究成果獲得2021 IEEE Signal Processing Society (SPS) Young Author Best Paper Award (台灣第二次獲此殊榮)。

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] ,…….

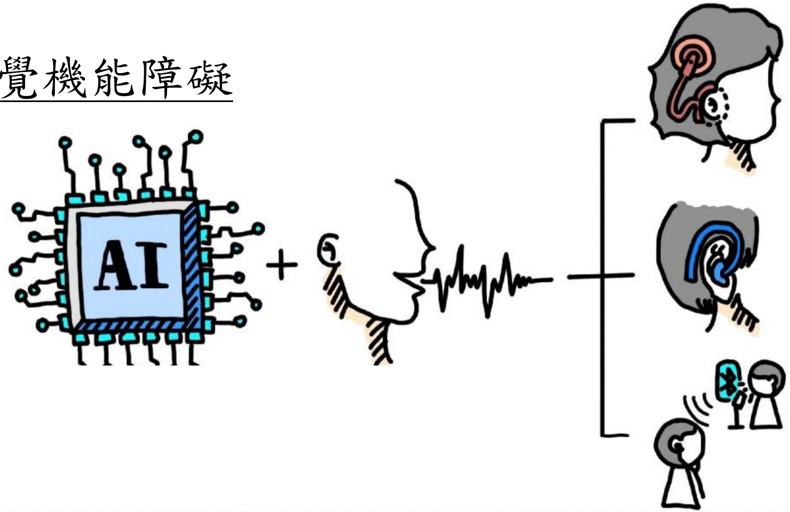
Outline

- Deep Learning (DL) based Speech Enhancement (SE)
  - Basic DL-based SE system architecture
  - Key factors to the DL-based SE performance
- **Assistive Oral Communication Technologies**
- Summary

基於AI的口語溝通輔助系統

- 聽 AI

聽覺機能障礙



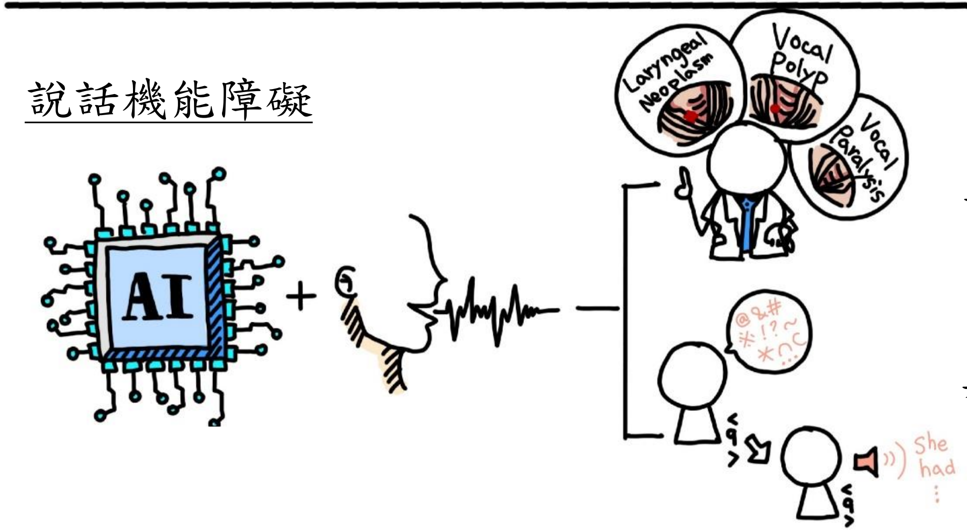
人工電子耳

助聽器

輔聽器

- 說 AI

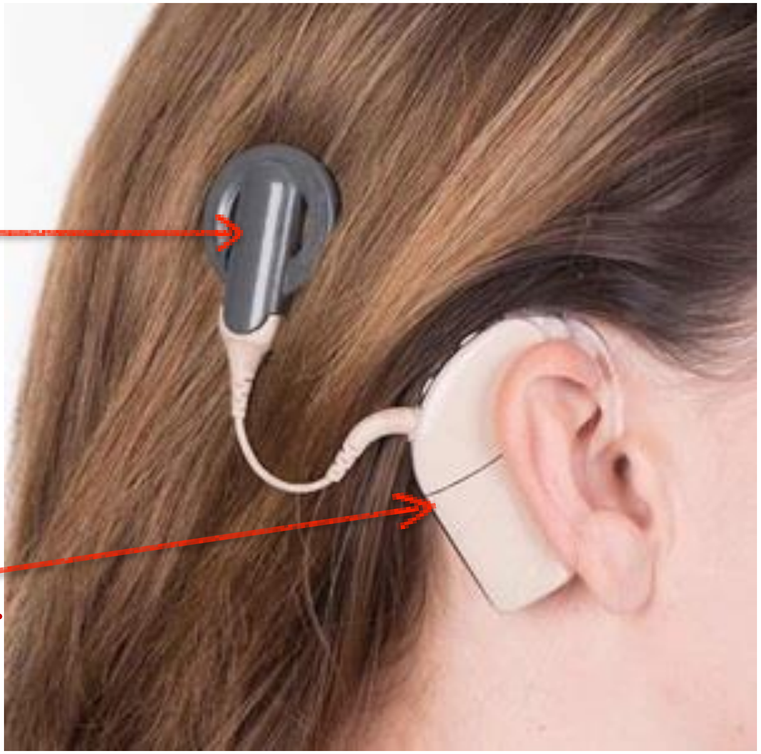
說話機能障礙



異常語音偵測

異常語音增強

人工電子耳



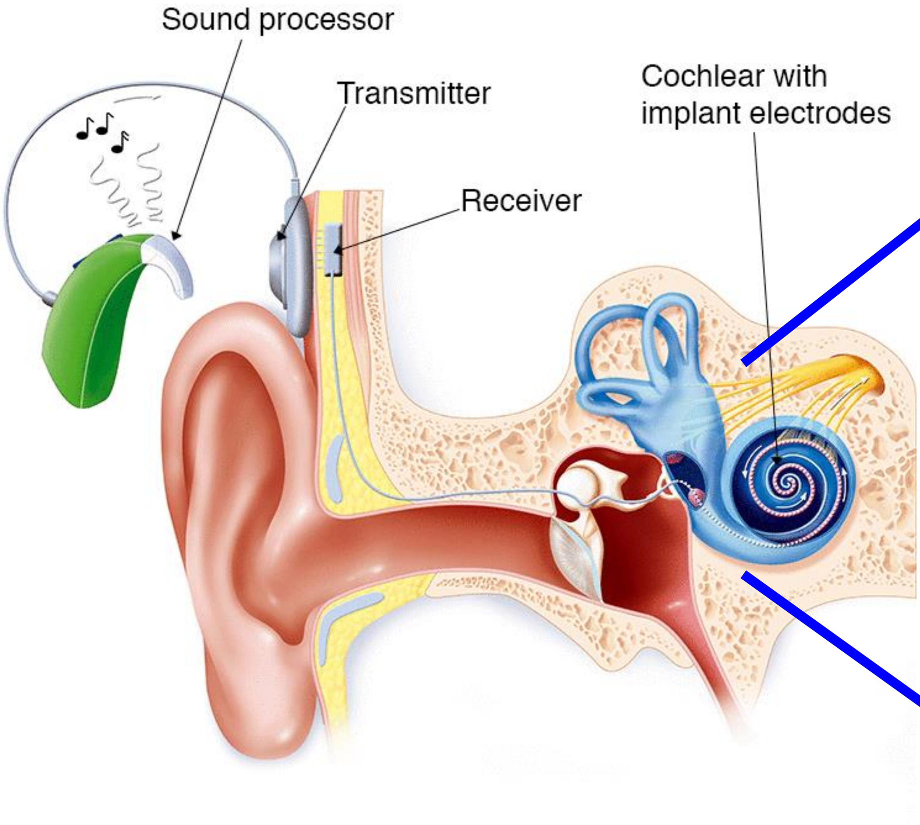
RF 發射線圈

語音處理綺

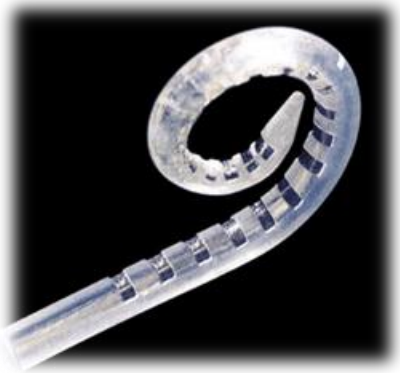
1. 麥克風
2. DSP 處理器.
3. 電池
4. …

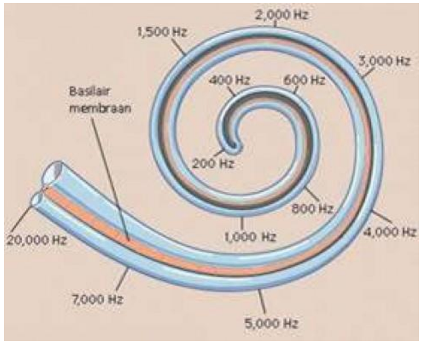
資料來源:
https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cochlear-implant-surgery

人工電子耳



植入電極





NOBEL PRIZE WINNER

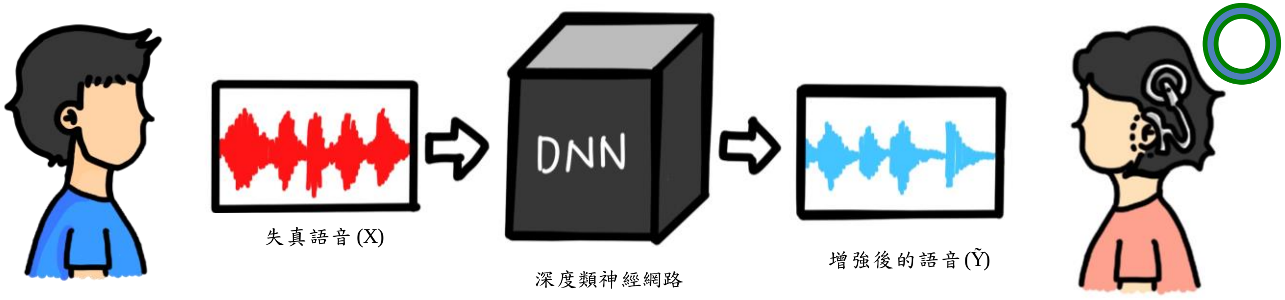
Traveling wave theory (Nobel Prize 1961)

資料來源:
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

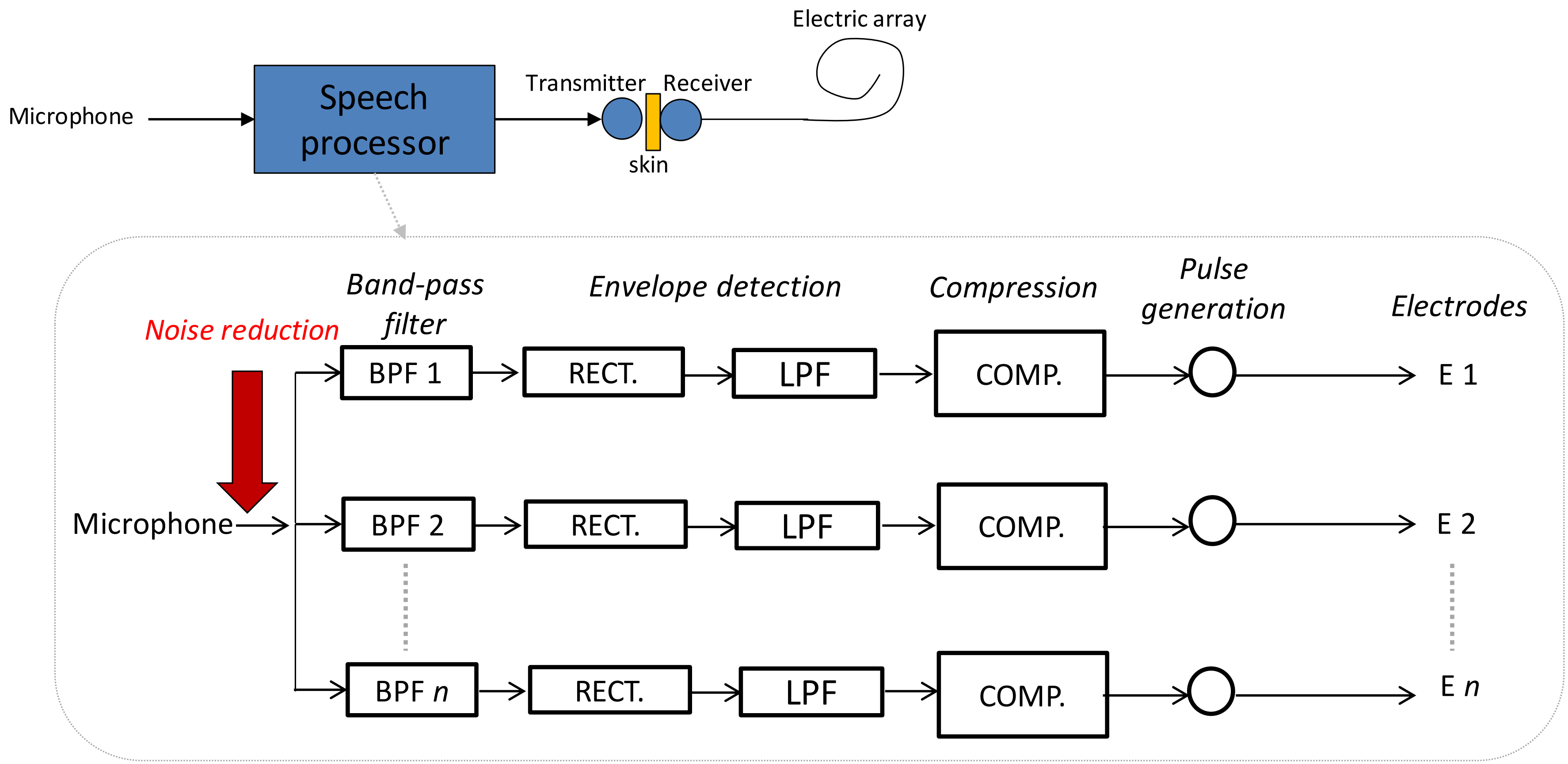
人工電子耳: A Modern Miracle

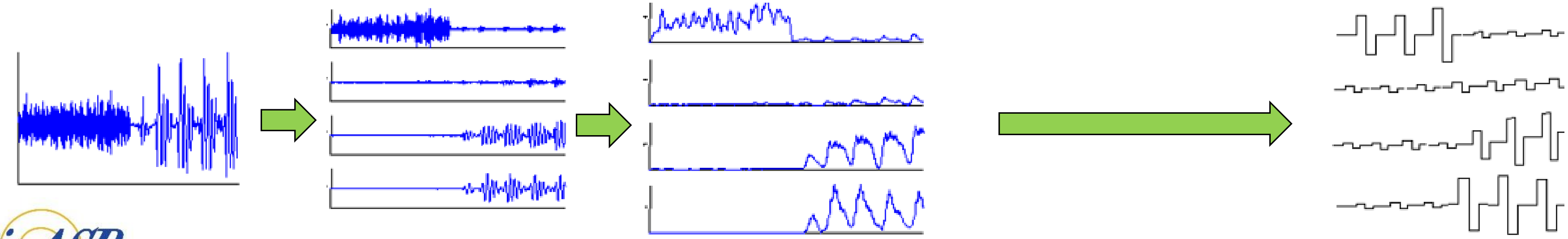
- 人工電子耳可以讓全聾者重新聽到聲音 (2018年人工電子耳納入健保)
- 在乾淨情況下，配戴者有高度辨識度，在有干擾情況下(特別是背景雜訊)，配戴者的理解度明顯降低

- 使用深度學習語音增強模型，提升配戴者的語音理解度



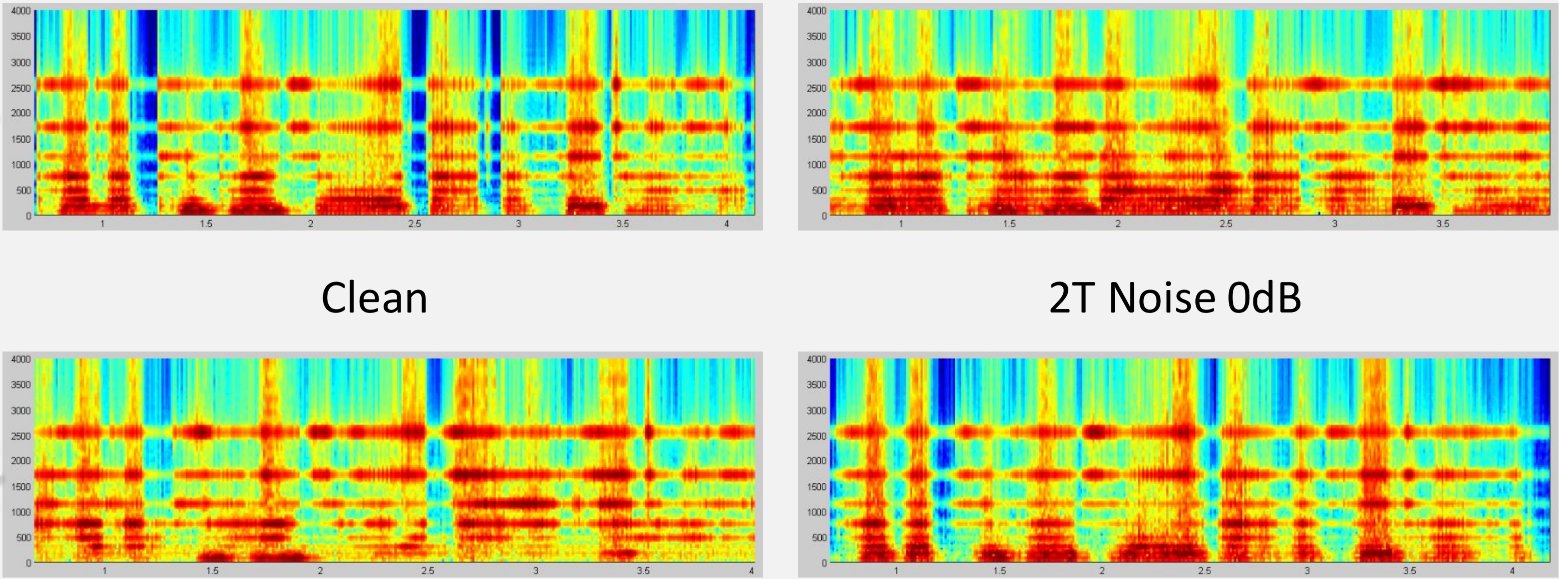
SE for Cochlear Implant





SE for Cochlear Implant Simulation

- Vocoded speech

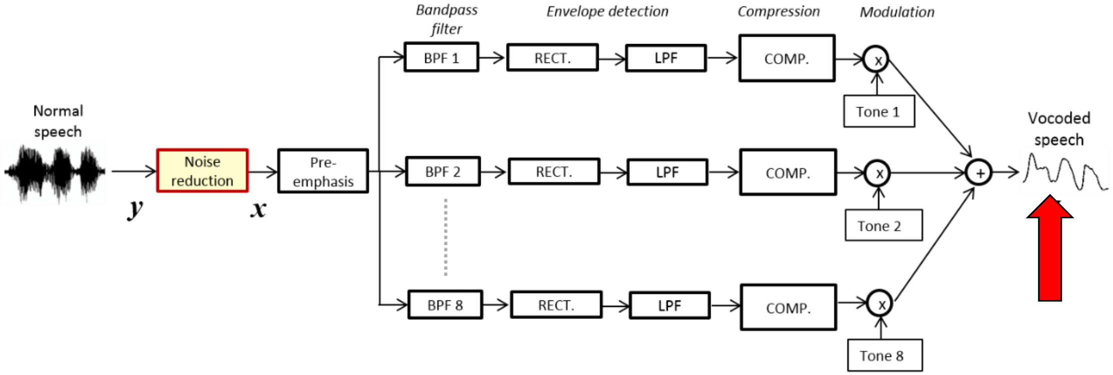


Clean

2T Noise 0dB

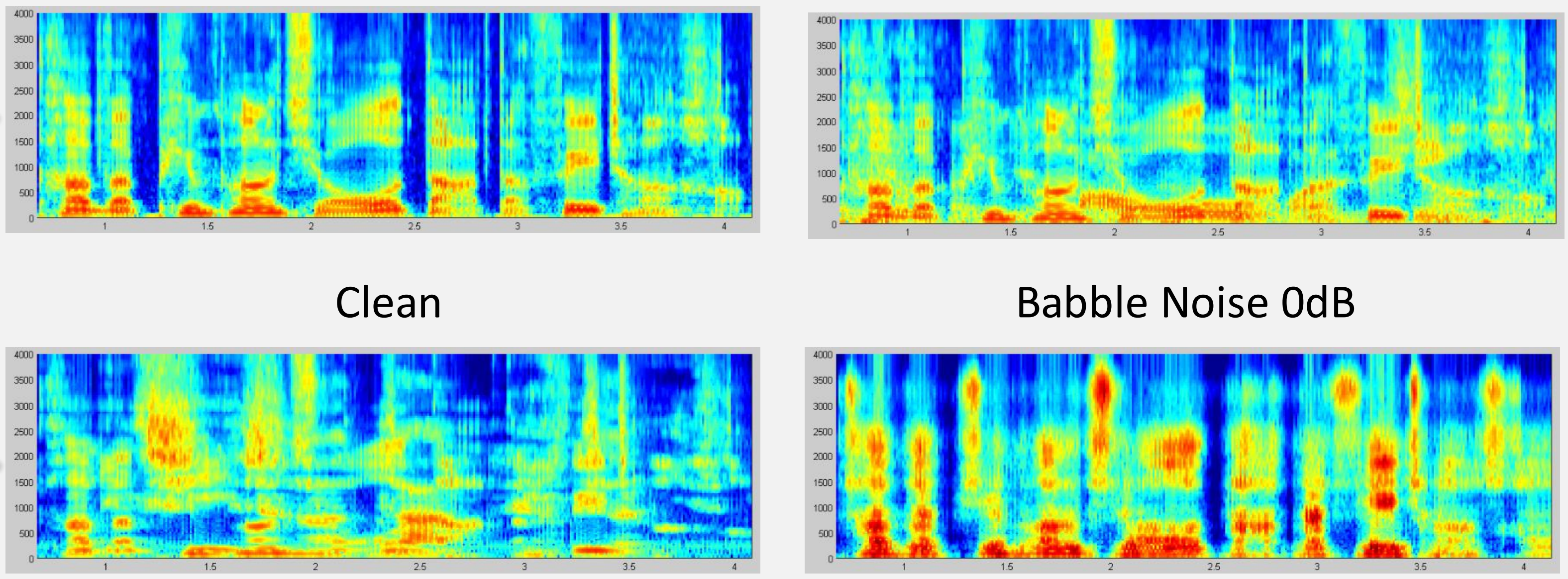
MMSE

DDAE



SE for Cochlear Implant Simulation

- Normal speech

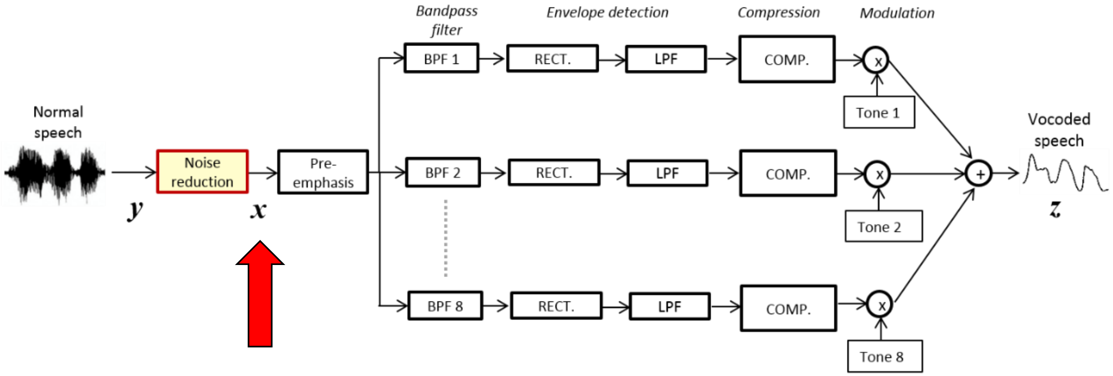


Clean

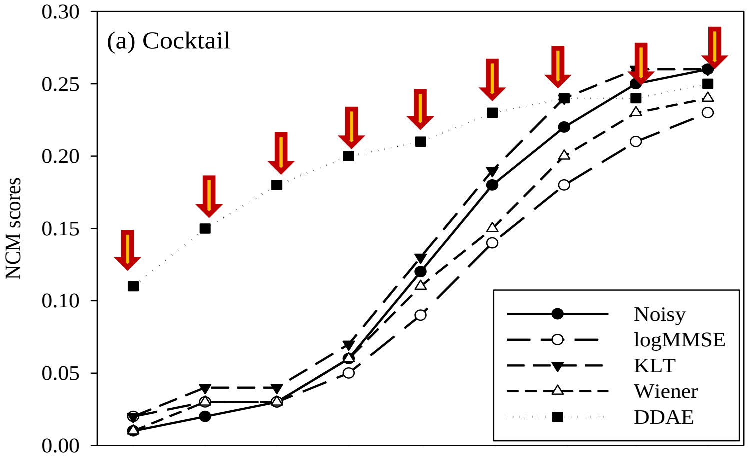
Babble Noise 0dB

MMSE

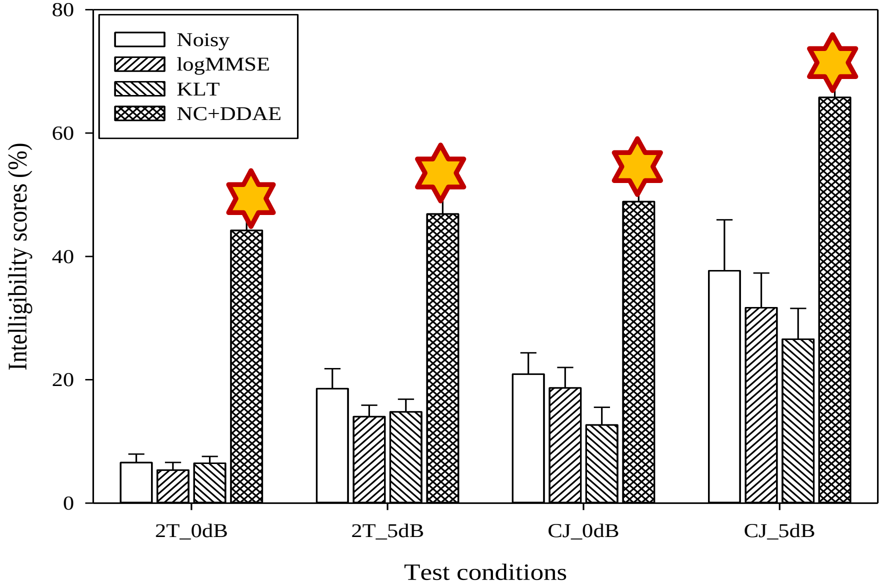
DDAE



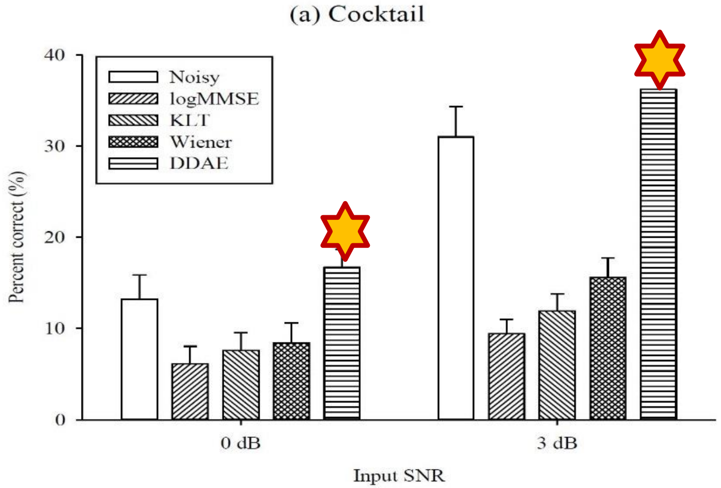
實驗結果



客觀測量



主觀測量(臨床實測).



主觀測量(模擬器)

(1) 基於深度學習的語音增強可以有效在**客觀評估**、**主觀評估(模擬器)**、**主觀測量(臨床實測)**均獲得明顯的進步。

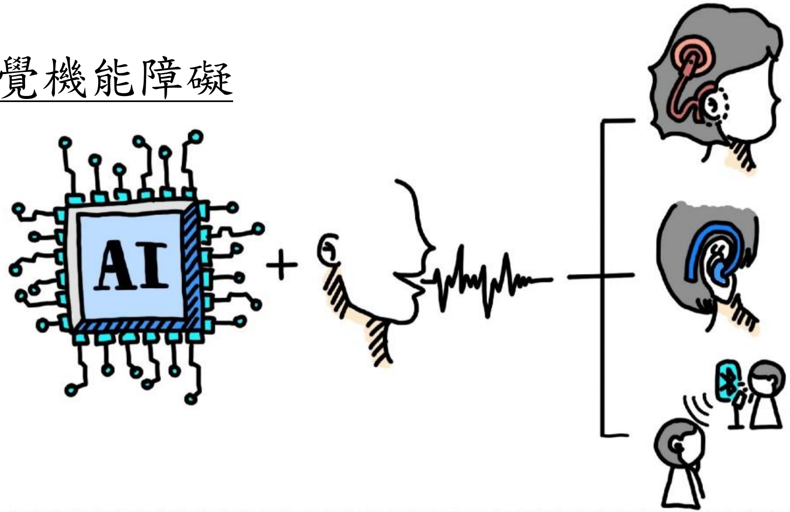
(2) 此研究成果為**全球首項**應用深度學習語音增強於人工電子耳的研究。

(3) 此研究成果獲得**國家新創獎** (2018-2020) 2022**未來科技獎**。

基於AI的口語溝通輔助系統

- 聽 AI

聽覺機能障礙



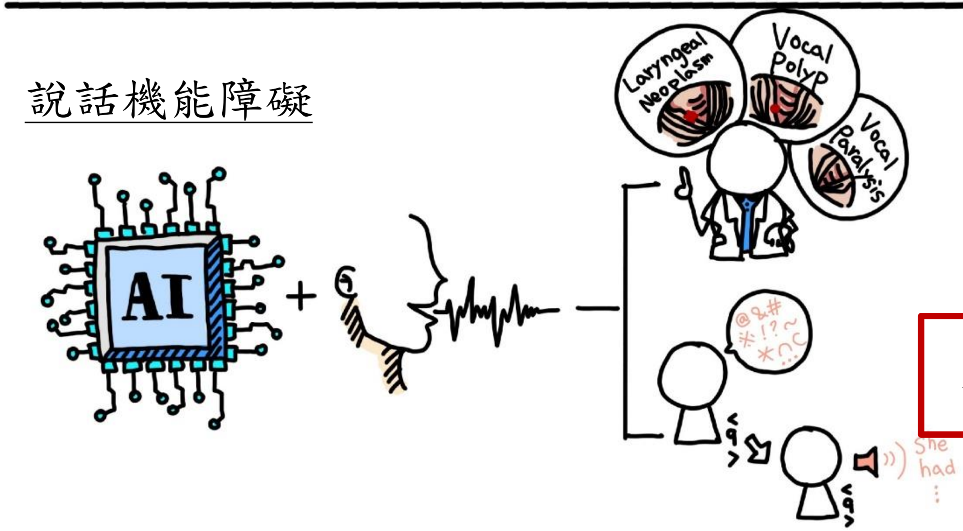
人工電子耳

助聽器

輔聽器

- 說 AI

說話機能障礙



異常語音偵測

異常語音增強

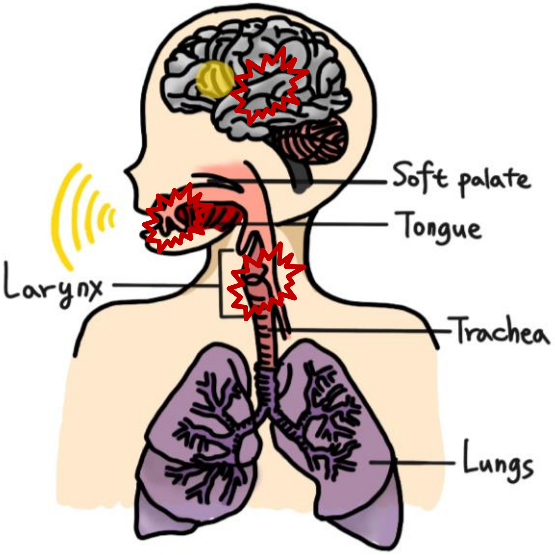
吶

文言版《說文解字》：訥，言難也

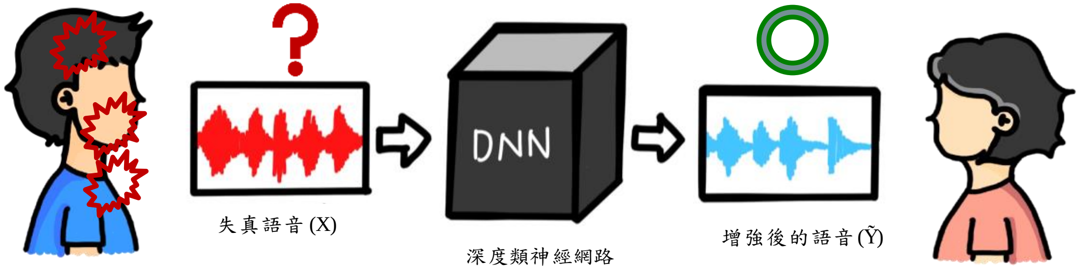
發音障礙:
構音異常、失語、口吃、口腔手術、聲帶損傷

語音增強於改善說話障礙

- 成果1：口腔癌術後語音理解度改善
- 成果2：構音障礙語音理解度改善
- 成果3：人工電子喉語音理解度改善



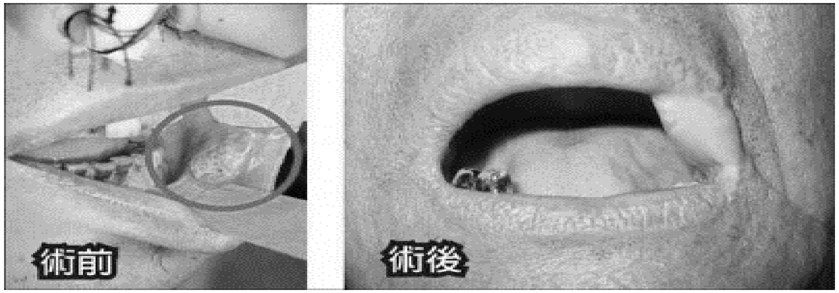
- 解決方法：基於深度學習的語音增強



Bio-ASP

SE for Speaking Disorder

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



Before After

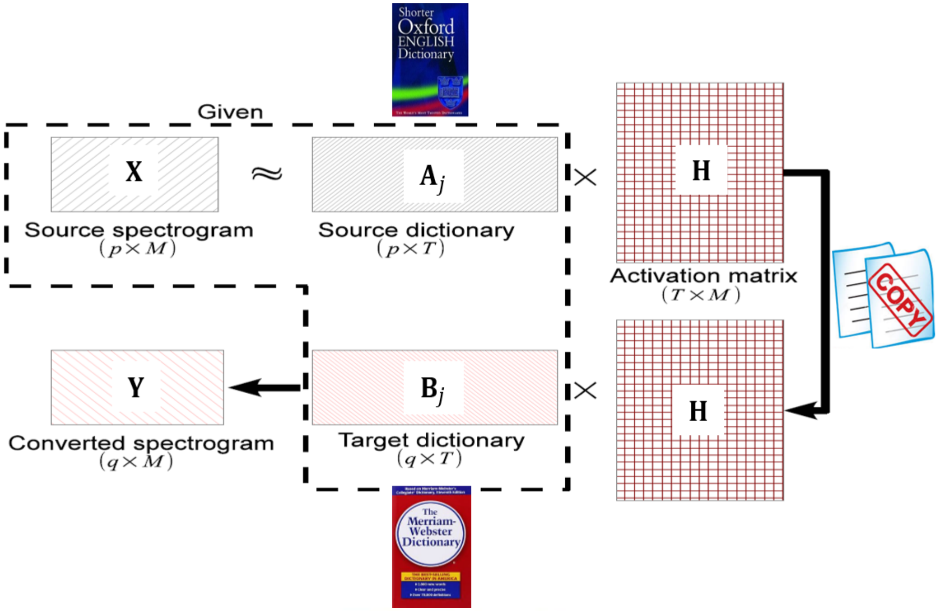
Liberty Times Ltd..

Before After

Taipei Veterans General Hospital

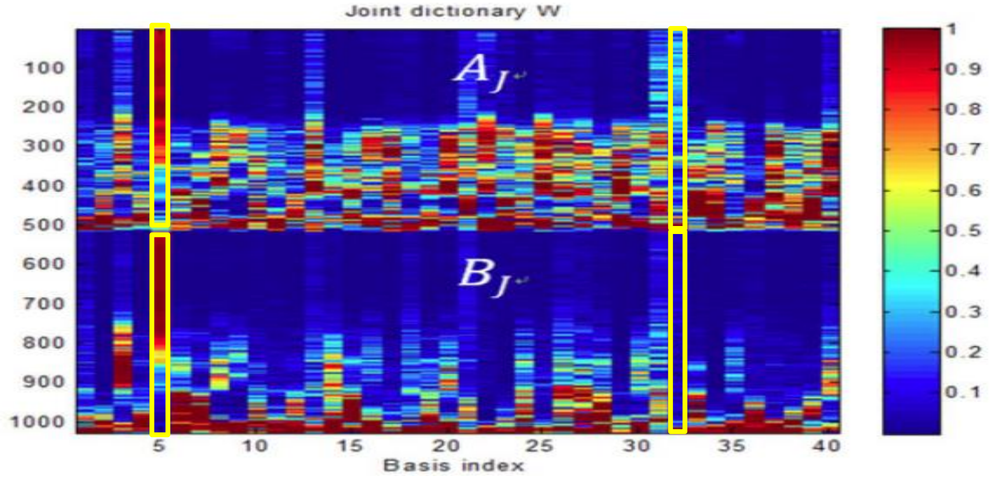
SE for Speaking Disorder

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

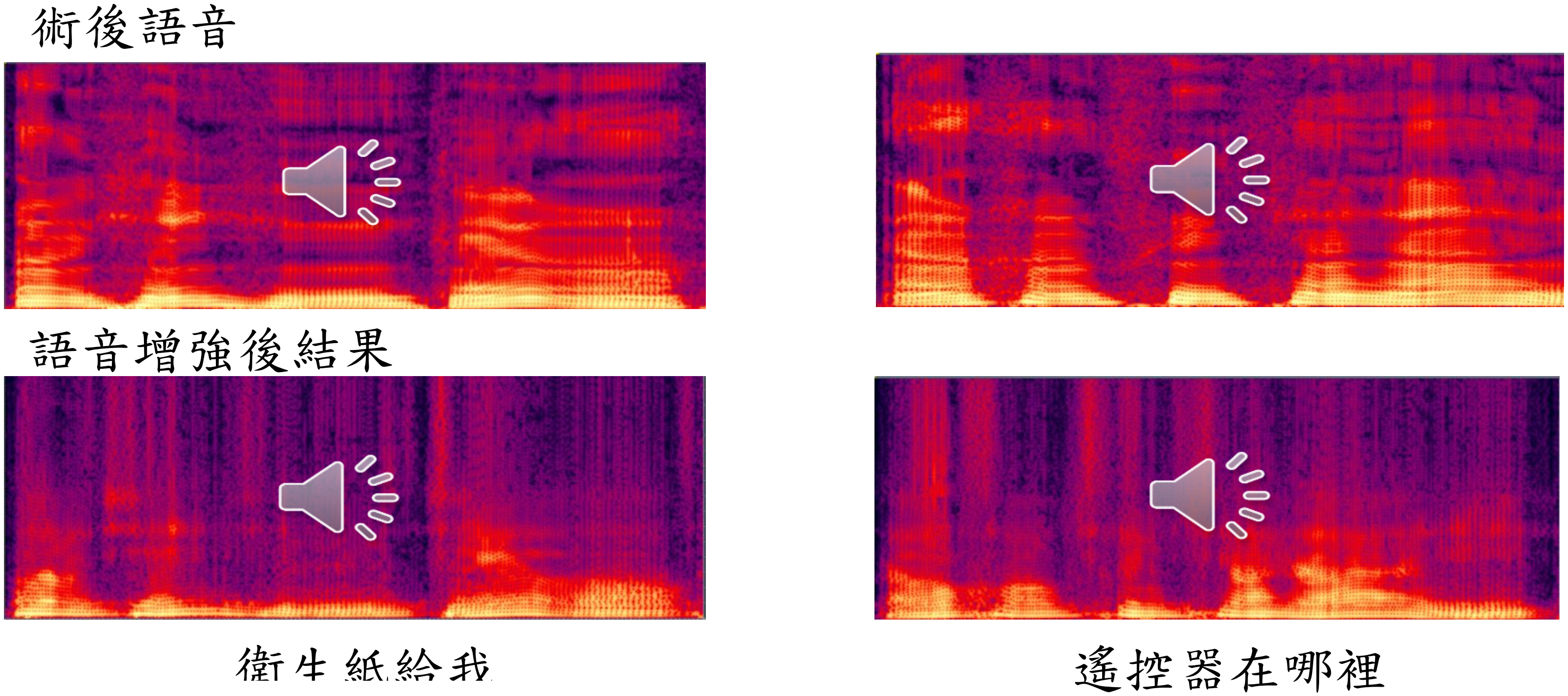


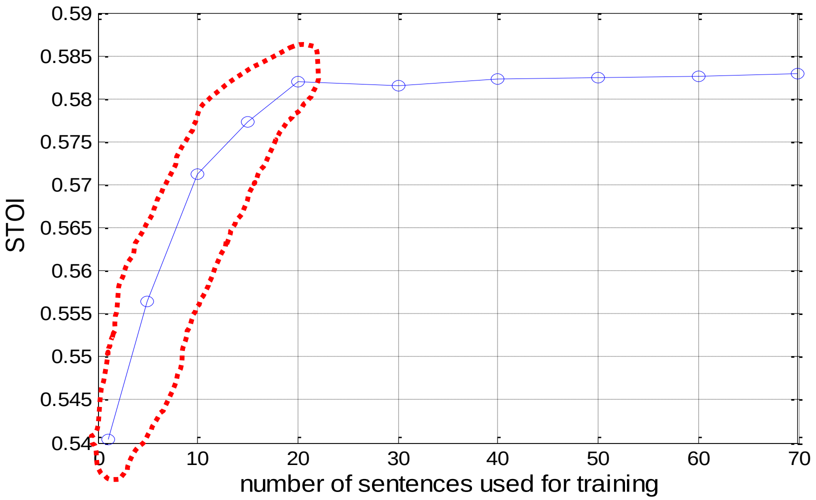
After Surgery

Before Surgery



語音增強於改善說話障礙(口腔癌術後)



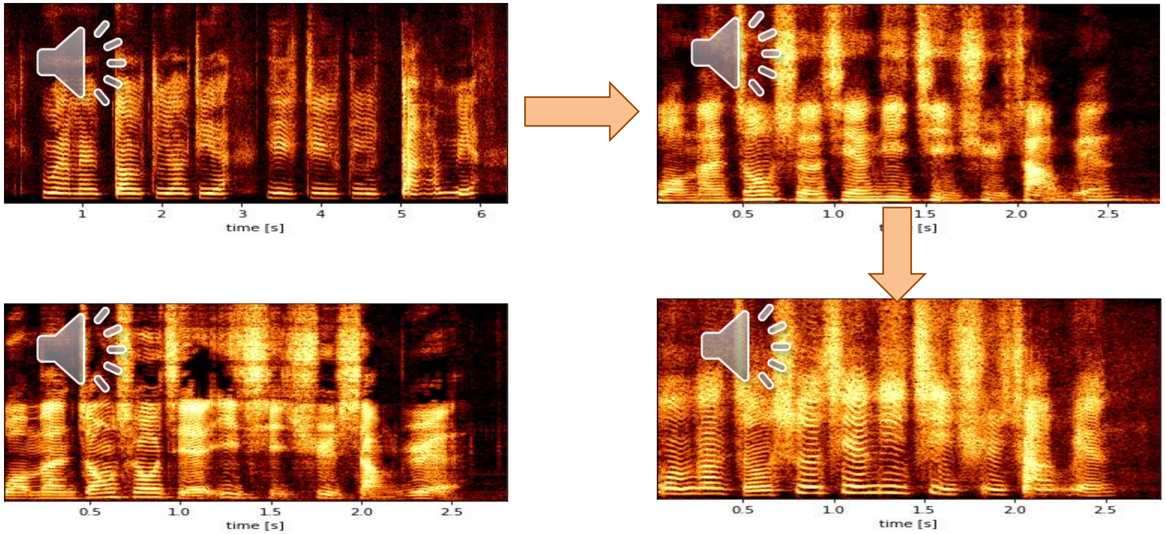


Speech samples were from [Fu et. al., TBME 2017]

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

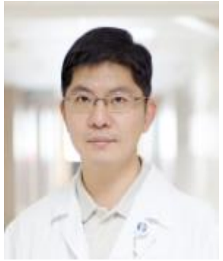
語音增強於改善說話障礙(腦性麻痺)

- 成果2：構音障礙語音理解度改善
- 使用兩階段轉換提升理解度之外並增加語音相似度



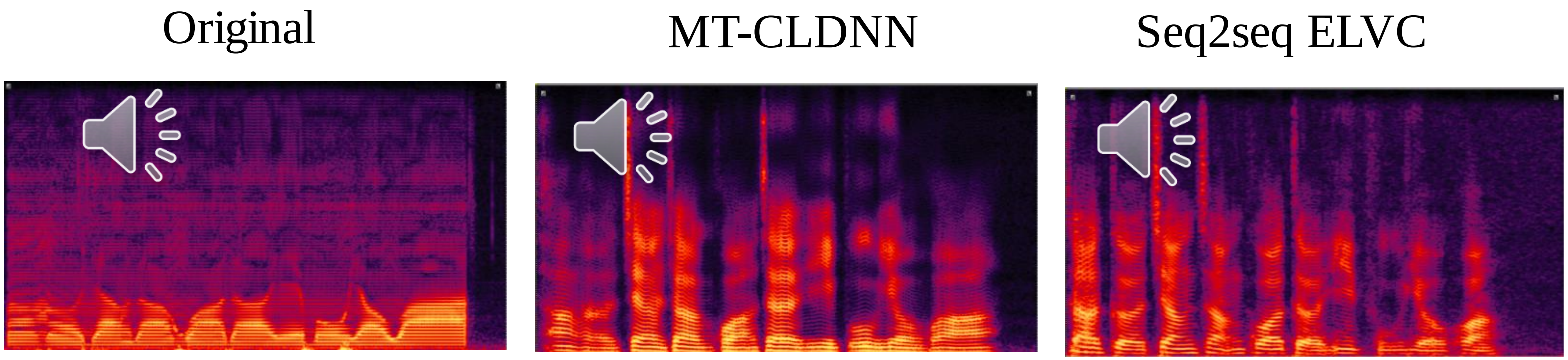
我們中秋節一起去賞月

➢ W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, T. Toda, “A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion," Interspeech 2021.



語音增強於改善說話障礙(電子喉語音)

- 成果3：人工電子喉語音理解度改善
- 使用新穎的序列對序列 (Seq2seq) 轉換演算法



Original

MT-CLDNN

Seq2seq ELVC

那個牆上掛著一幅油畫

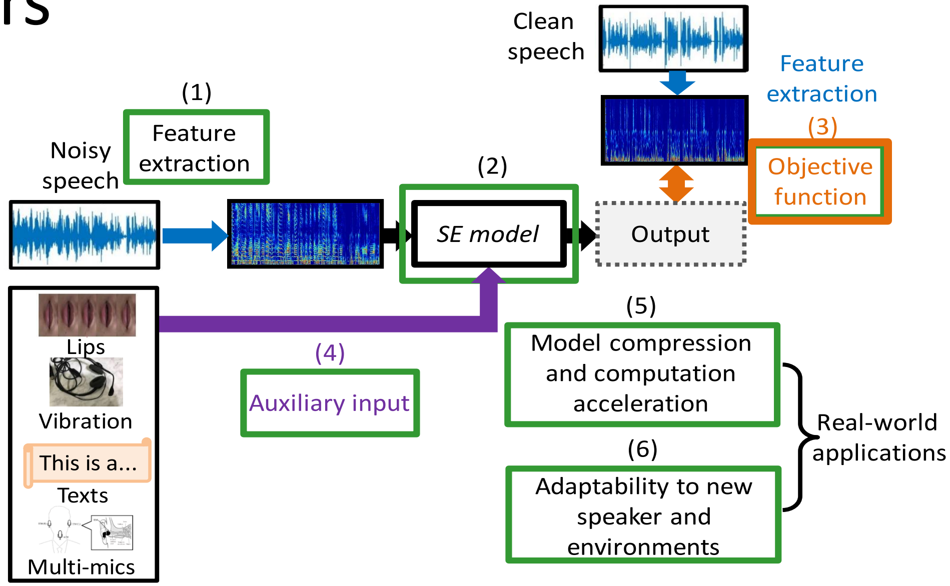
➢ M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. R. Jang, and H.-M. Wang, “Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling, ASRU 2021”

Outline

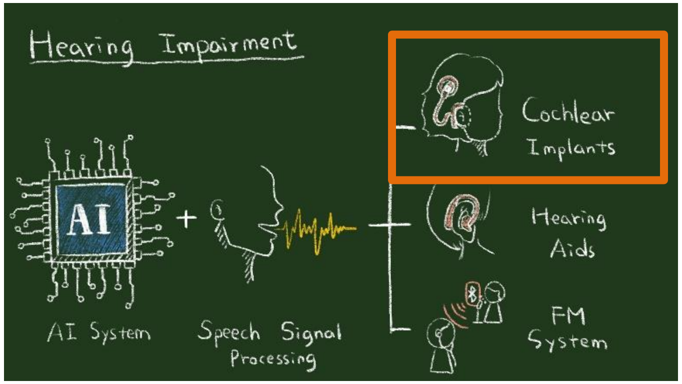
- Deep Learning (DL) based Speech Enhancement (SE)
  - Artificial intelligence and deep neural networks
  - Basic DL-based SE system architecture
  - Key factors to the DL-based SE performance
- Assistive Oral Communication Technologies
- **Summary**

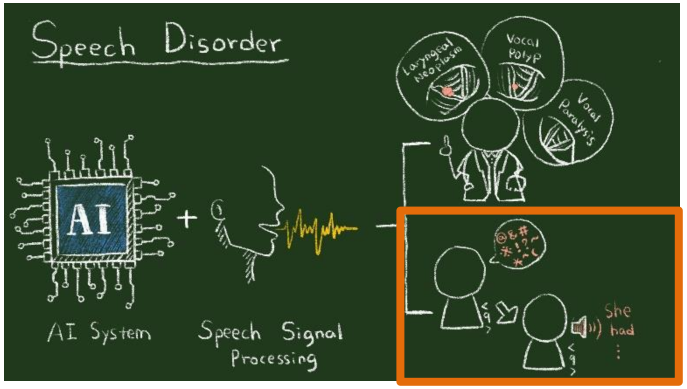
結論

- Key Factors

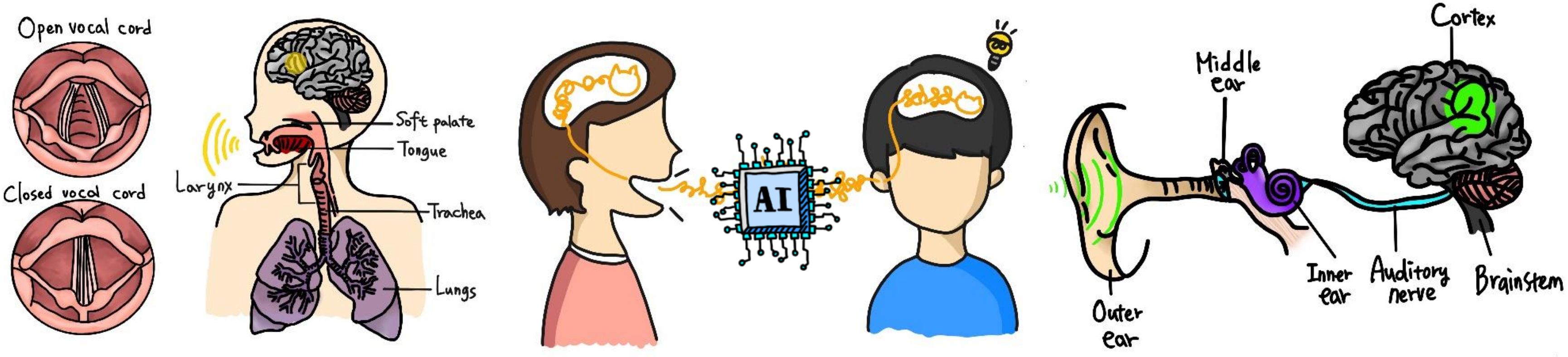


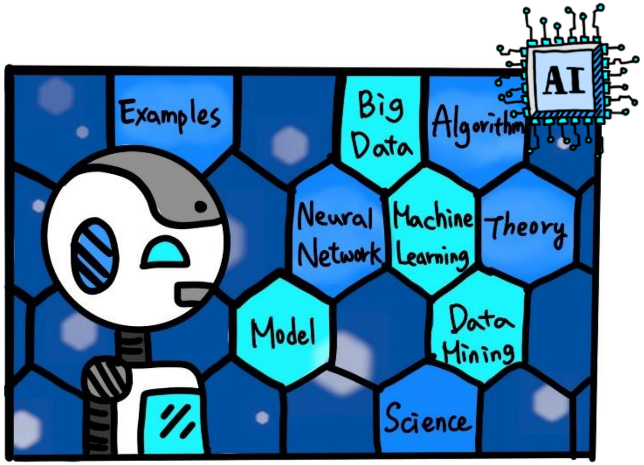
- Assistive Oral Communication Technologies





結論

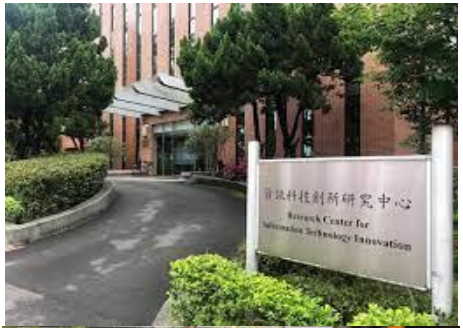




聽說 AI

歡迎一起來作
有溫度的研究

特別感謝



Bio-ASP Lab

Biomedical Acoustic Signal Processing Lab

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

Bio-ASP

Thank You Very Much for Your Attention

Bio-ASP