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

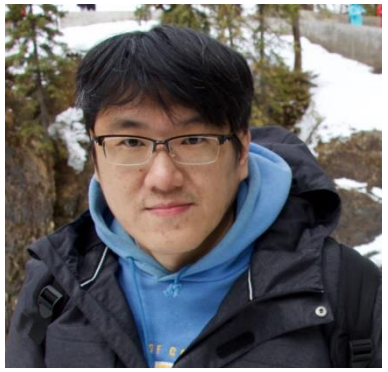
Yu Tsao

Research Center for Information Technology Innovation

Academia Sinica

yu.tsao@citi.sinica.edu.tw

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)

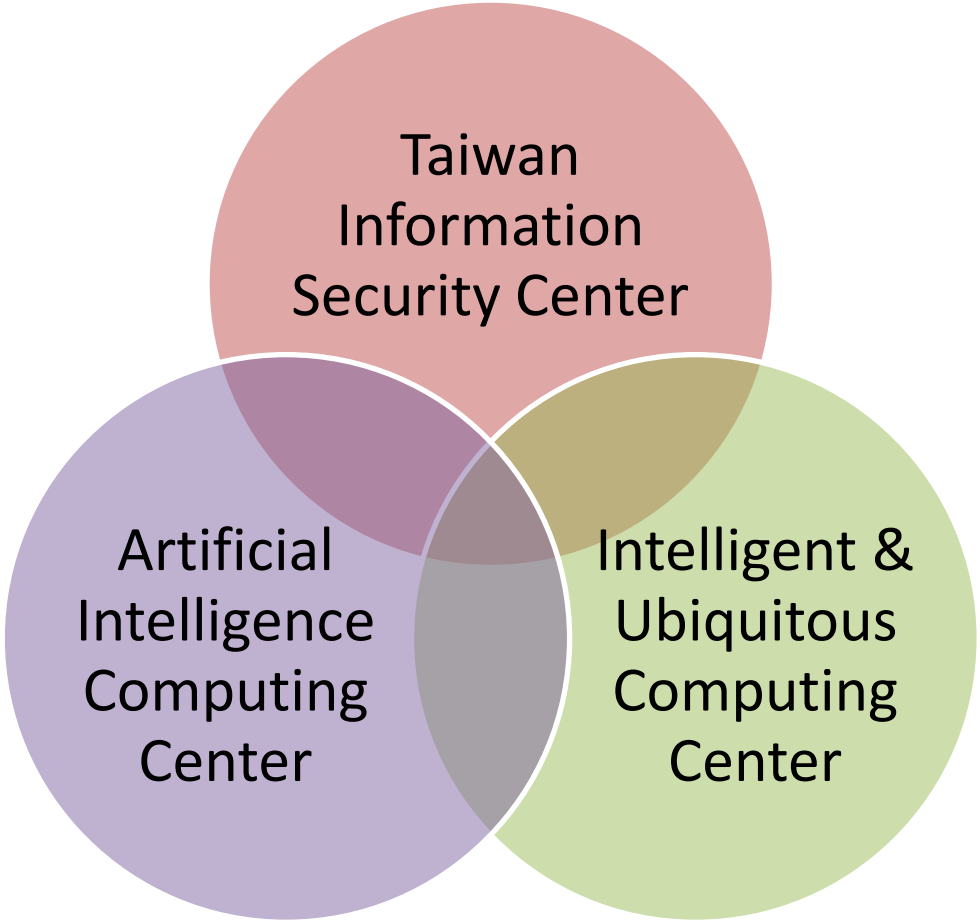
Research Fellow, Deputy Director of CITI, Academia Sinica

Biomedical Acoustic Signal Processing (Bio-ASP) Lab

– Research Interests

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

Research Center for Information Technology Innovation (CITI)



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

Current Director: Dr. Yennun Huang

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

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

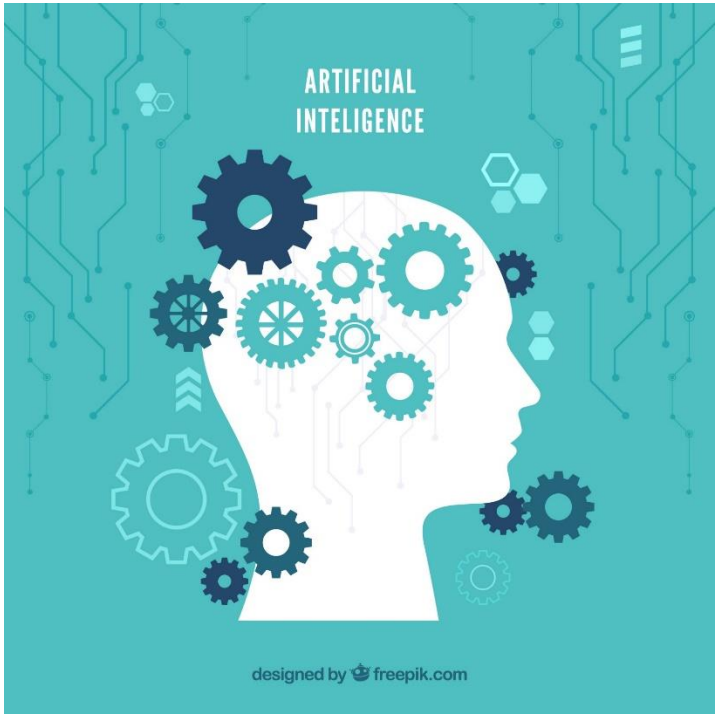
Outline

- 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

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

Machine Learning and Artificial Intelligence



Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

1. Deduction, reasoning, problem solving
2. Knowledge representation
3. Default reasoning and the qualification problem
4. Machine planning
5. **Machine learning**

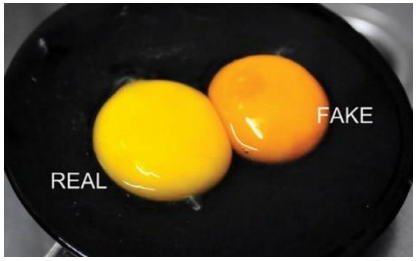
Pattern recognition
Density estimation
Linear models for regression
Linear models for classification
Deep **Neural networks**
Kernel methods
Sparse kernel machines

From M. Svensen & C. Bishop, "Pattern recognition and machine learning"

Artificial?


REAL
VS
FAKE
leather jacketshop


Vs
BUTTER
MARGARINE


FAKE
REAL


?

Artificial Intelligence?

CAMBRIDGE UNIVERSITY PRESS

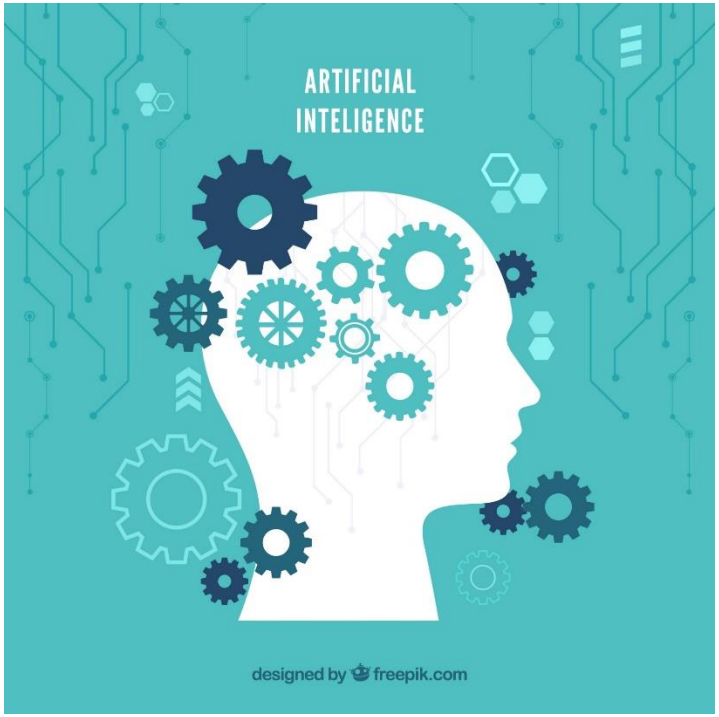
Intelligence

The ability to **understand** and **learn** well, and to **form judgments** and opinions based on reason.

Artificial Intelligence

The study of how to produce **machines** that have some of the qualities that the **human mind** has, such as the ability to understand language, recognize pictures, solve problems, and learn.

Machine Learning and Artificial Intelligence



Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

1. Deduction, reasoning, problem solving
2. Knowledge representation
3. Default reasoning and the qualification problem
4. Machine planning
5. **Machine learning**

Pattern recognition
Density estimation
Linear models for regression
Linear models for classification
Deep **Neural networks**
Kernel methods
Sparse kernel machines

From M. Svensen & C. Bishop, "Pattern recognition and machine learning"

Artificial Neural Network (ANN)

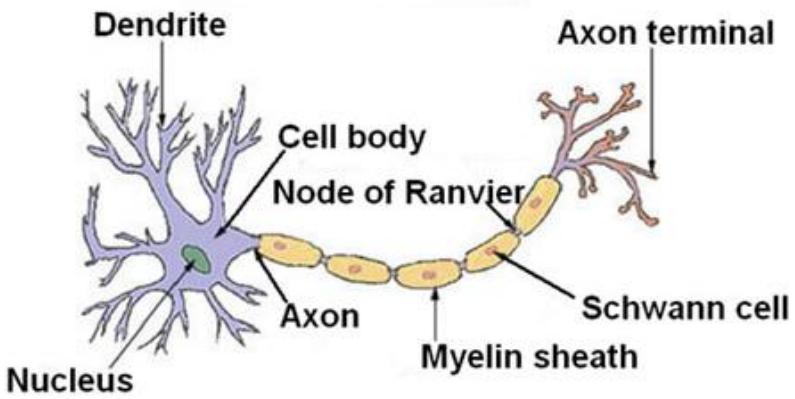
- Artificial neural network (ANN) is a **computational model** that **mimics** brain functionality with artificial means.

From Youtube

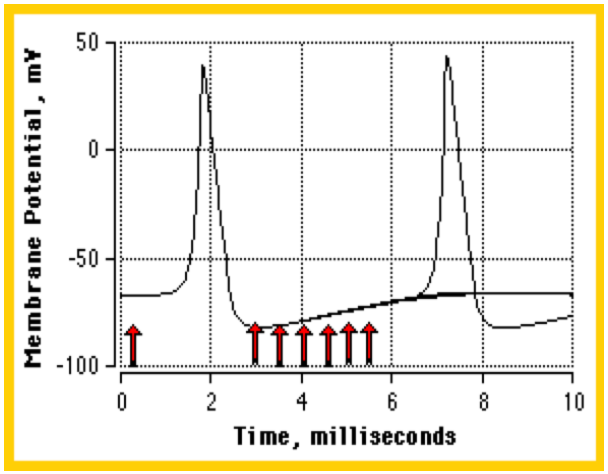
Artificial Neural Network (ANN)

- Structure of a Typical **Neuron**.

突觸、軸突、髓鞘、細胞核



Adrian, 1932



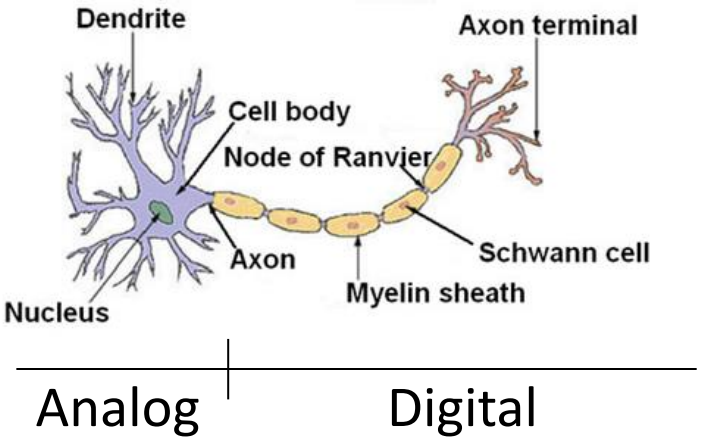
Analog | Digital

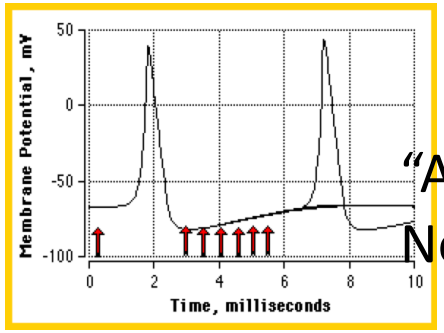
"All-or-None Nerve Firing"

Bio-ASP

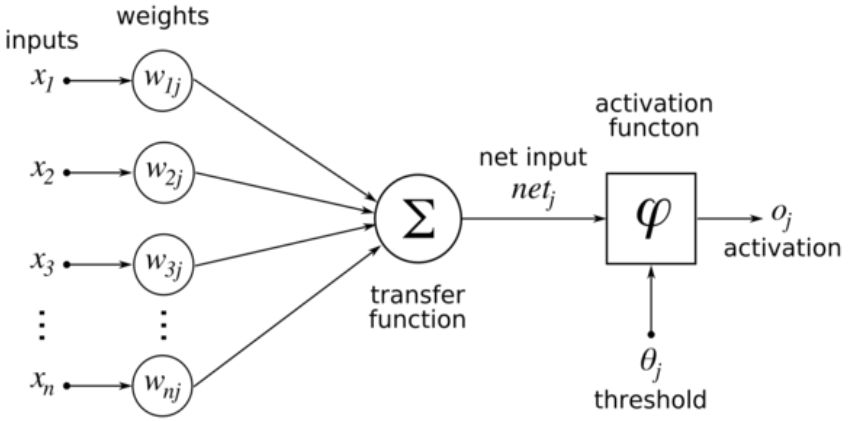
Artificial Neural Network (ANN)

- Structures of **Neuron** and **Artificial Neuron**.





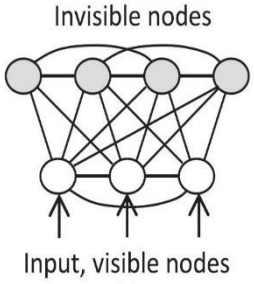
"All-or-None Nerve Firing"



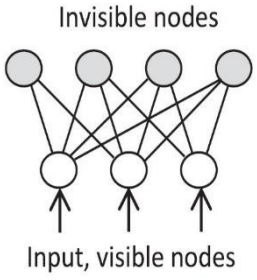
Bio-ASP

Artificial Neural Network (ANN)

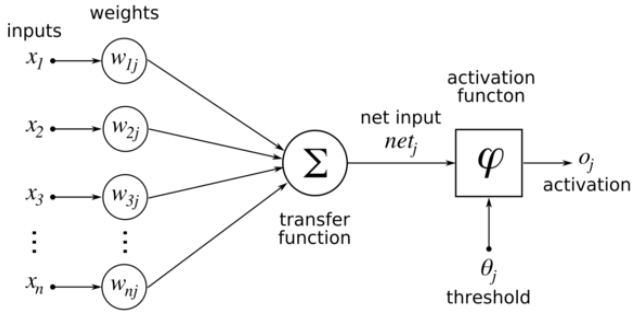
- Artificial Neural Network (Multiple Artificial Neurons) for **Generation** and **Discrimination** (Classification).



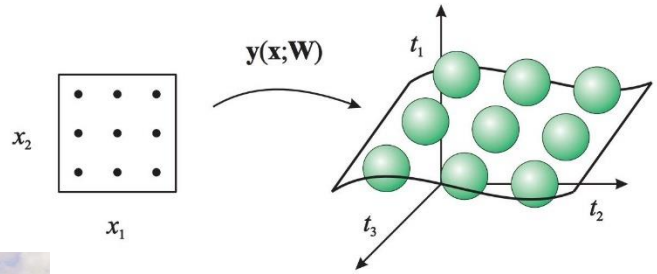
Boltzmann machine



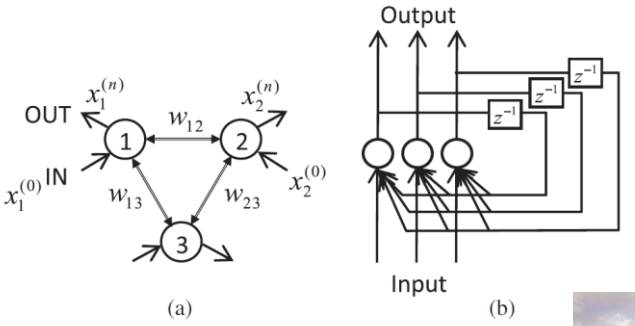
Restricted Boltzmann machine



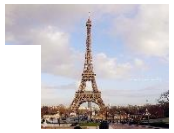
McCulloch–Pitts model



Generative topographic map



Hopfield network

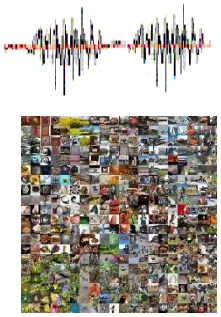


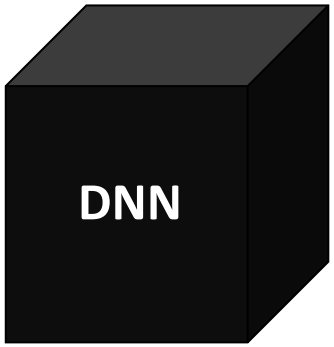
From B.-H. Juang, "Deep neural networks – a developmental perspective," APSIPA Trans. on SIP

Bio-ASP

Analyzing DNN Model and Human Brain

- Difficult to fully understand what is inside
- Analyzing functions of DNN/brain by sending input signals and investigating activations → performance prediction



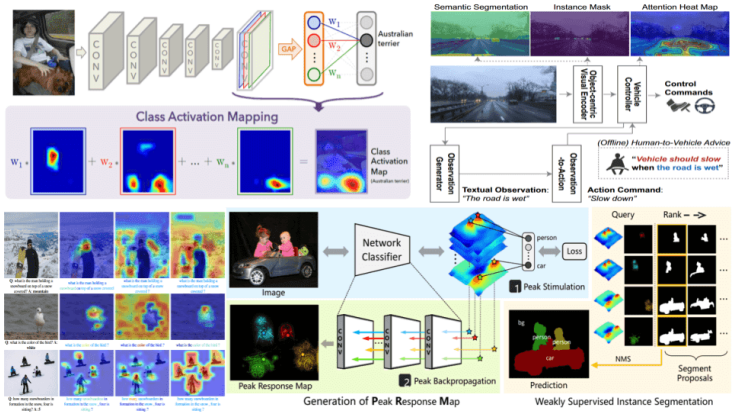


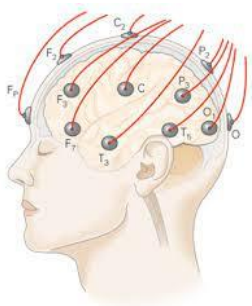
Classification

Generation

Classification

Generation





EEG

MEG

fMRI

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

Speech ….Easily Got Distorted

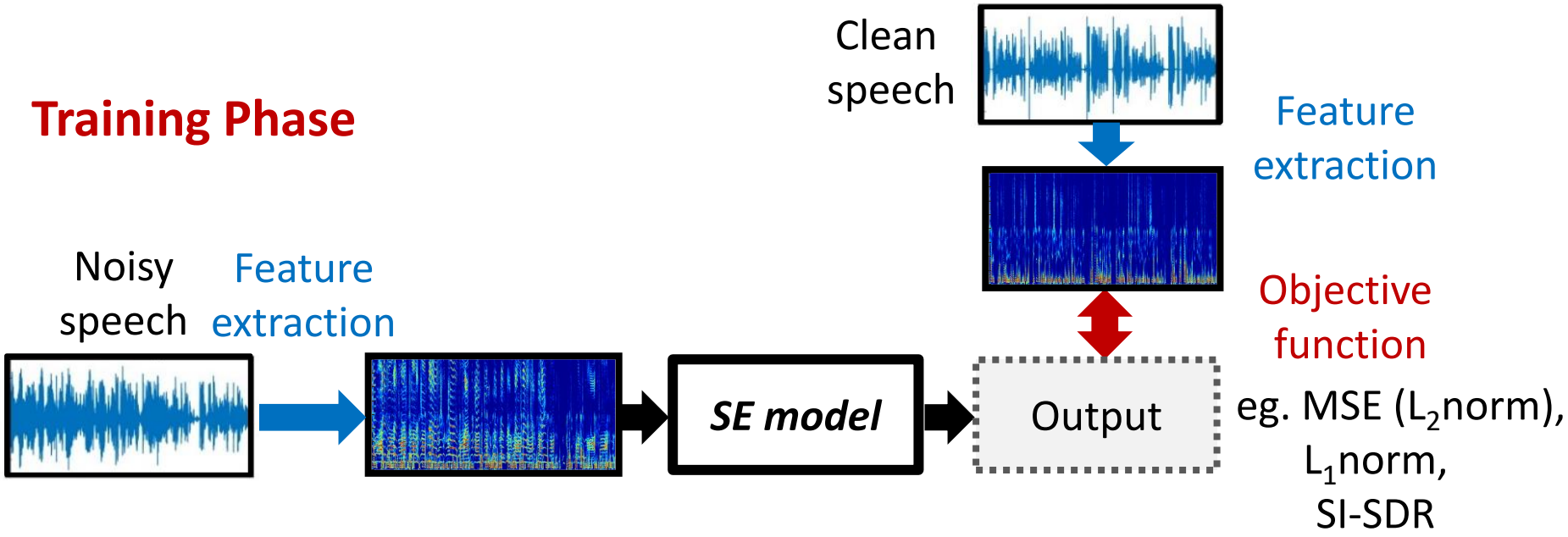




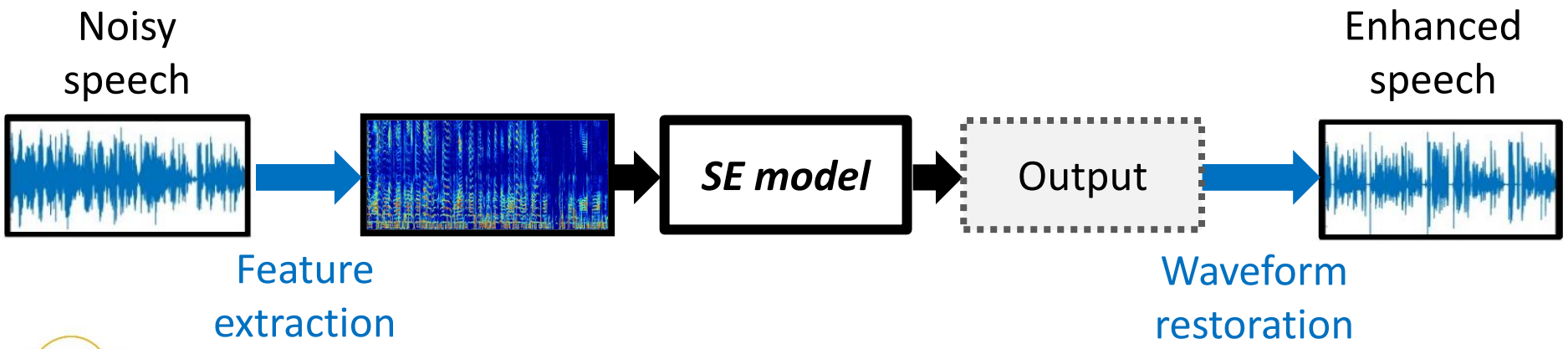

Bio-ASP

Deep Learning Based SE System

Training Phase

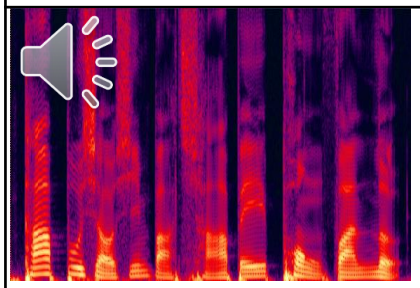
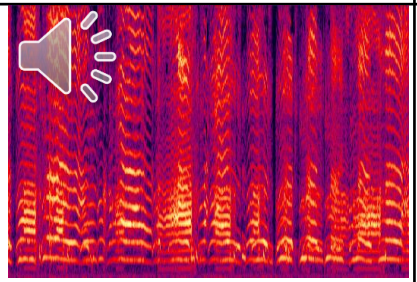
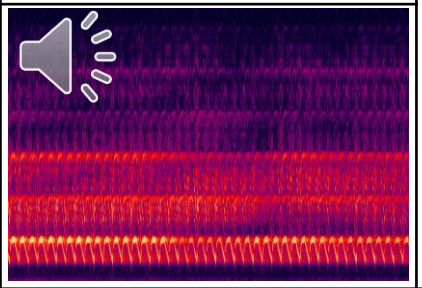
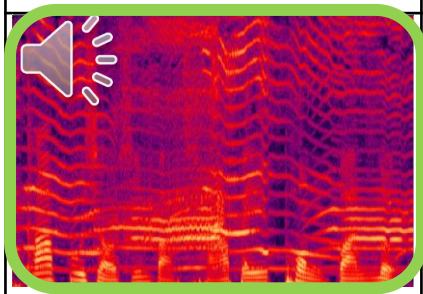
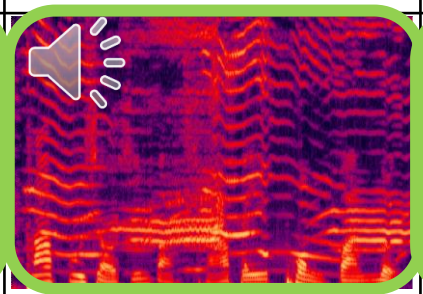
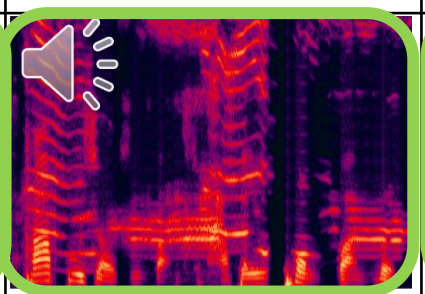
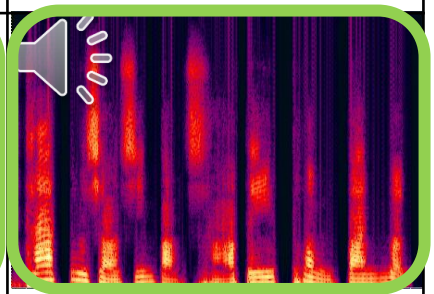
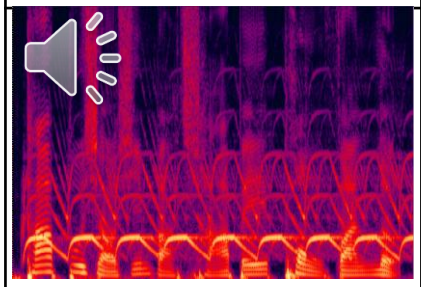
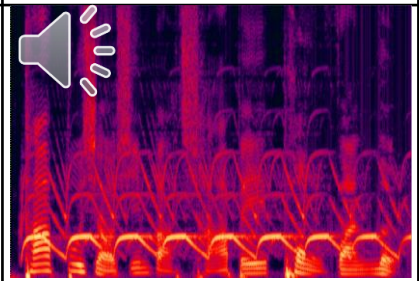
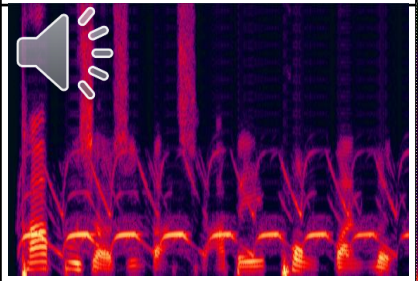
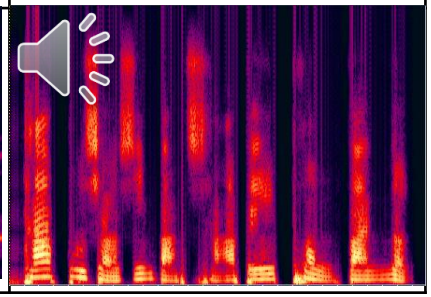


Testing Phase



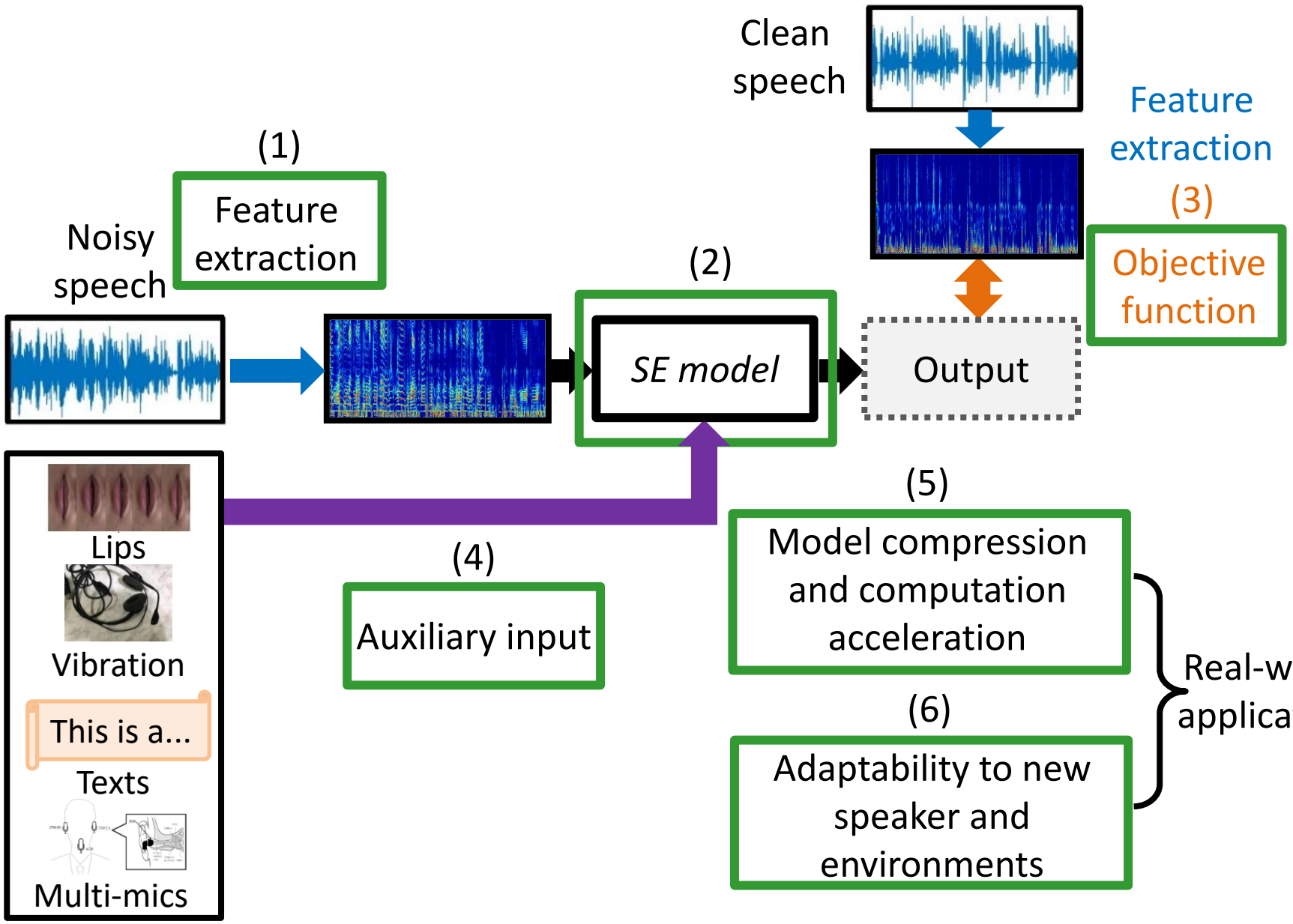
The first work of DL-based SE system: [Lu et al, Interpseech 2013].

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**
				

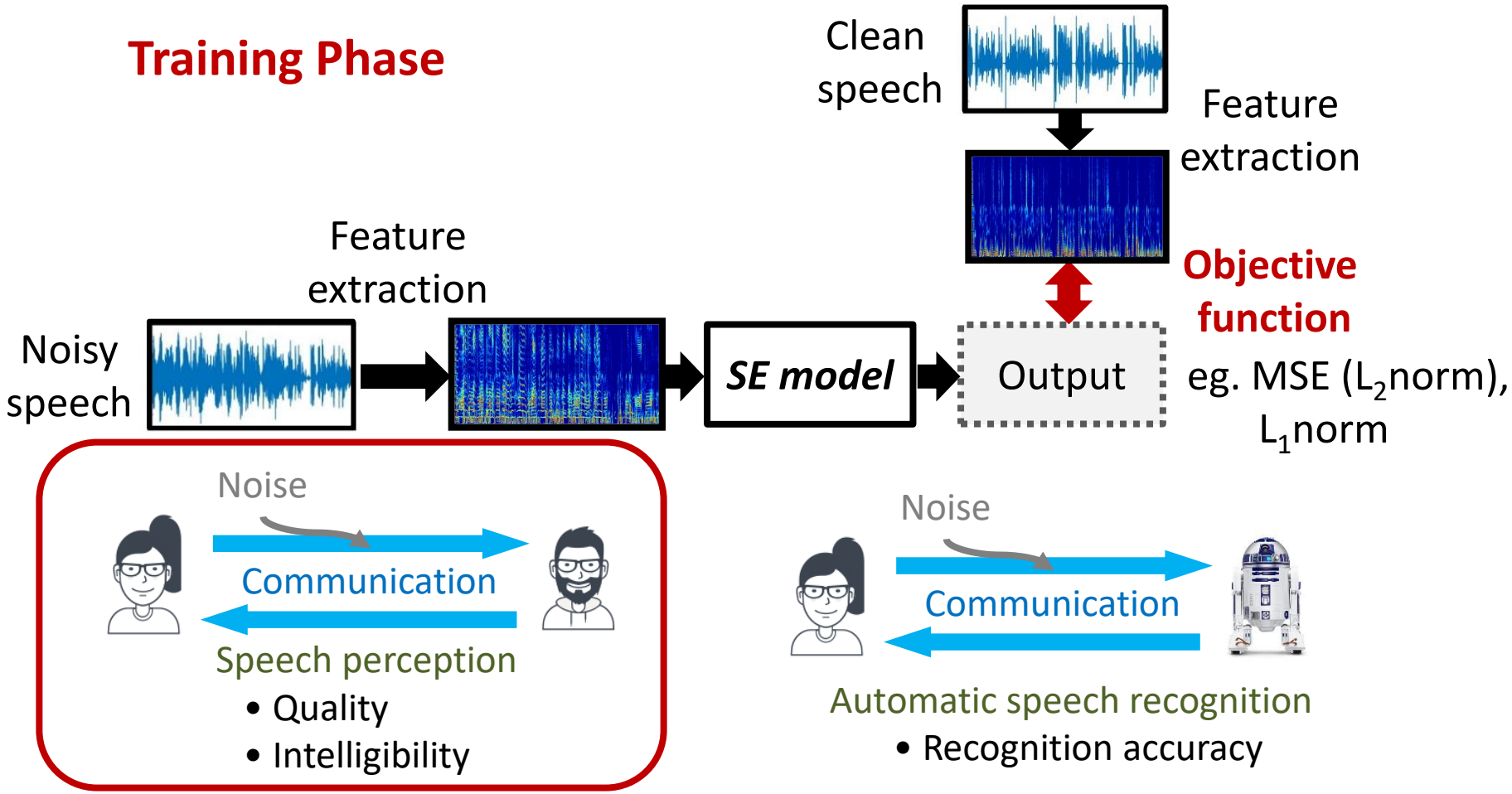
Bio-ASP

Deep Learning Based SE System



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.

Bio-ASP

聞

聽

大學曰：心不在焉，聽而不聞

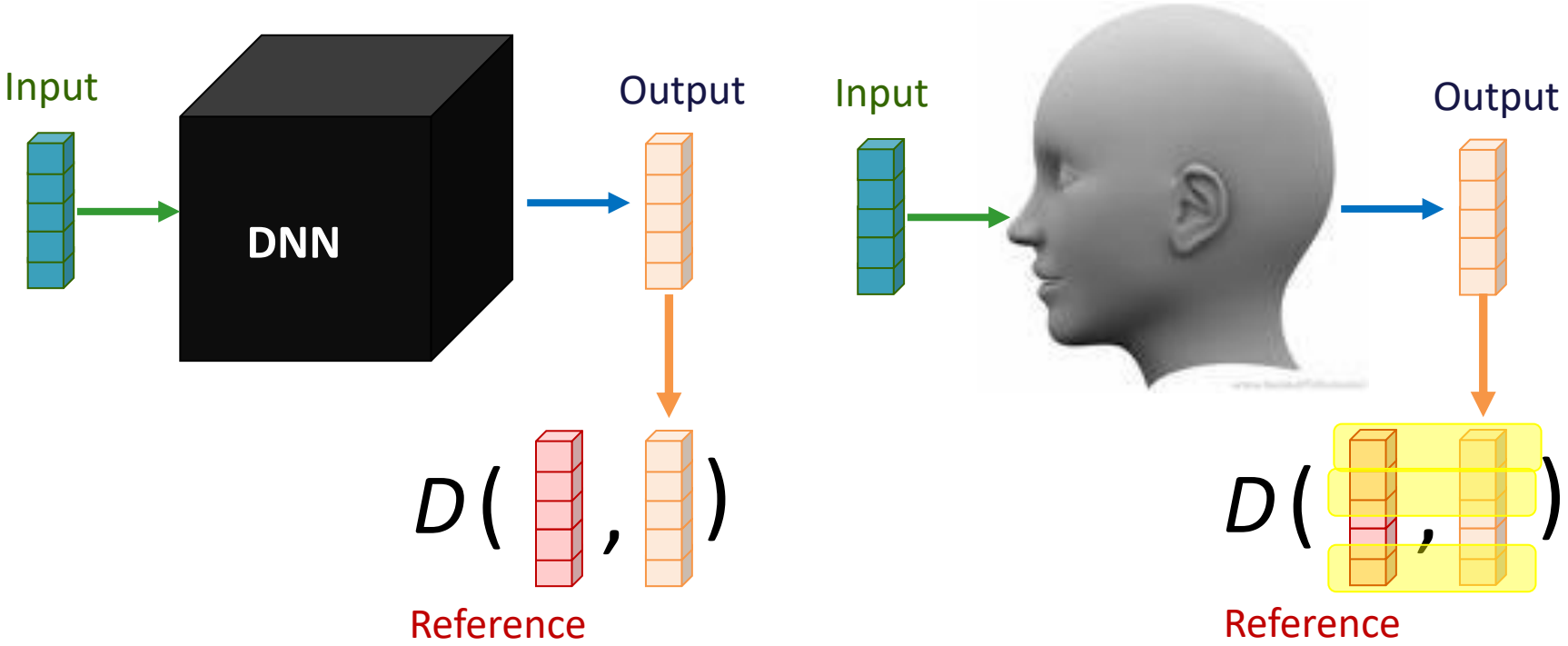
Hear but pay no attention; listen but not hear

Intelligibility and Quality are different

Intelligibility is more important than quality for assistive listening devices

Objective Functions for DNN and Brain

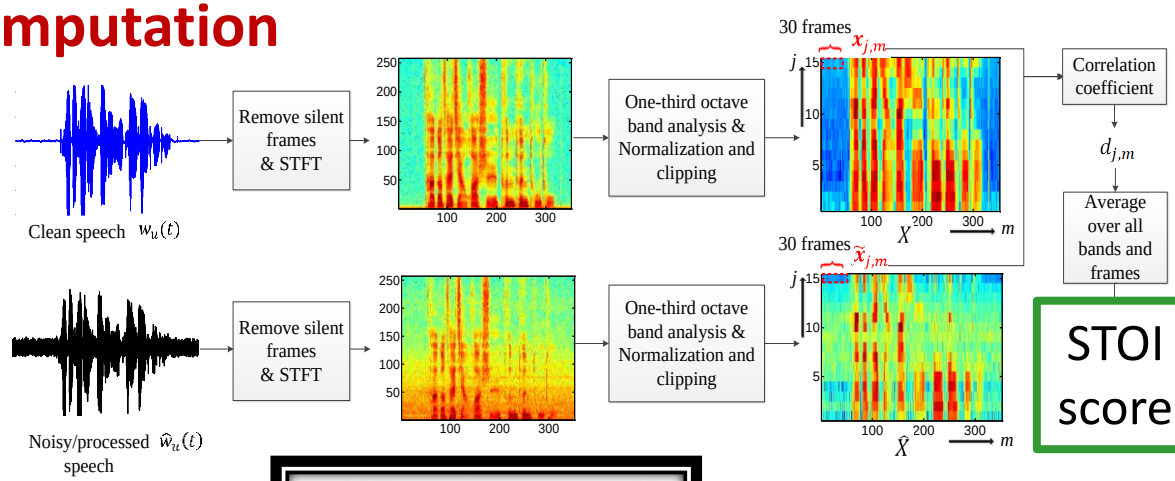
- DNN Model vs. Human Brain
  - ➢ Difficult to fully understand what is inside
  - ➢ What we can control: input, output, objective function



Objective Function

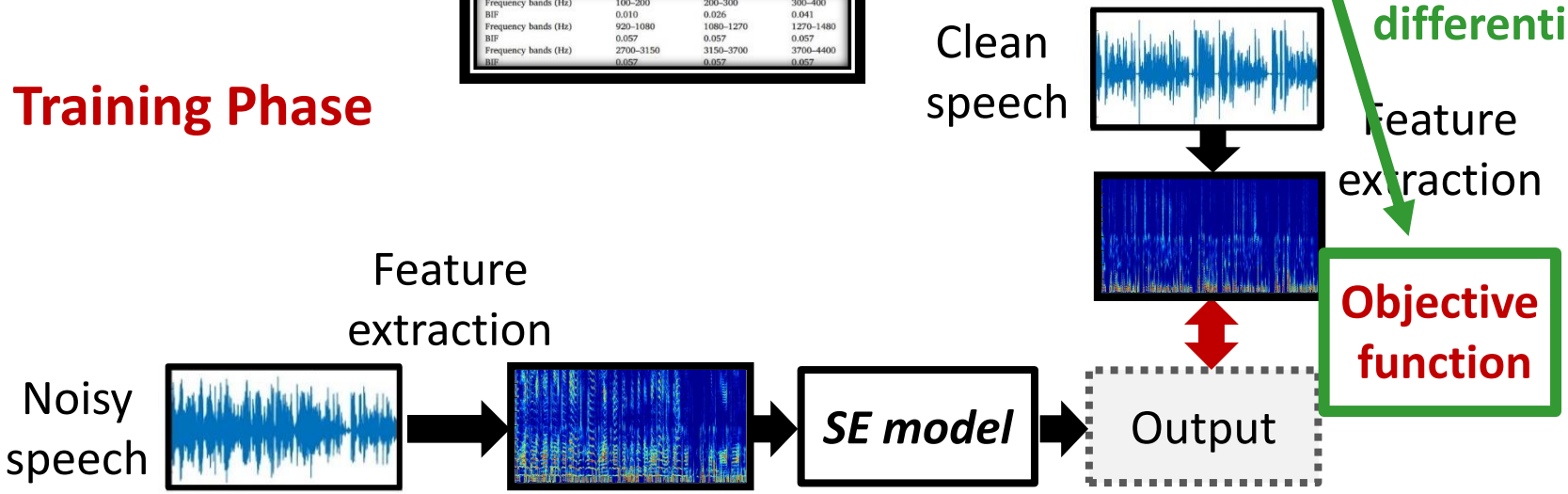
- STOI-based Objective Function [Fu et al, TASLP 2018]

STOI Computation



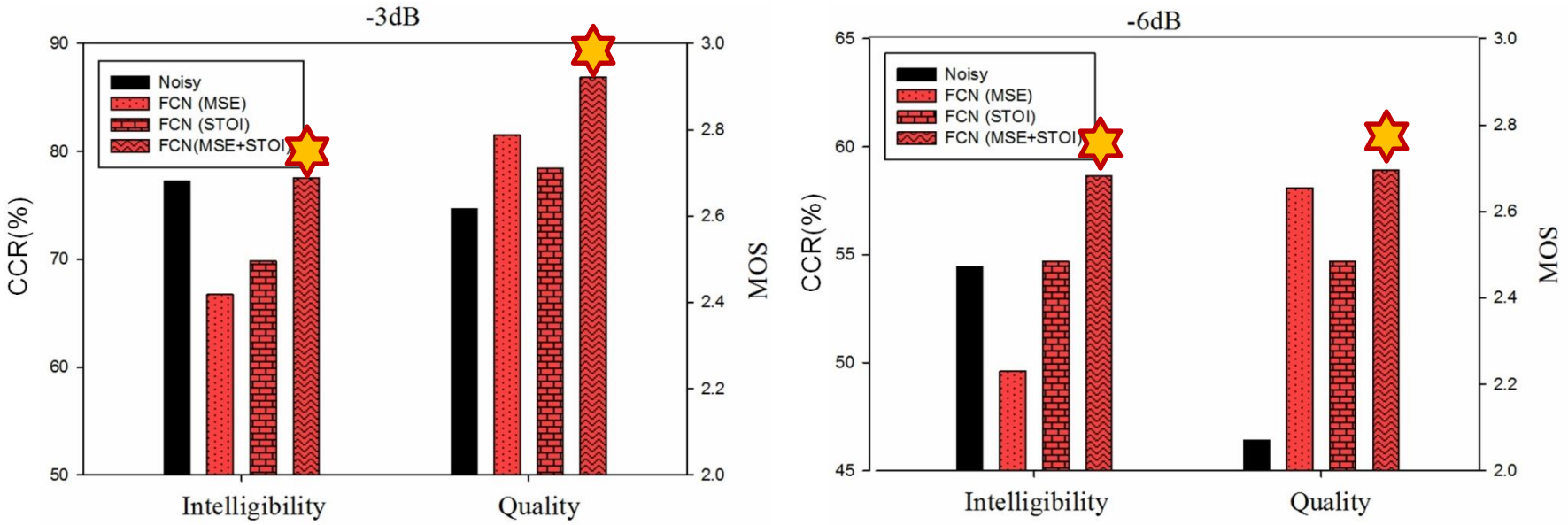
Linear and differentiable

Training Phase



Objective Function (STOI)

- Experimental Results (Human Listening Test)

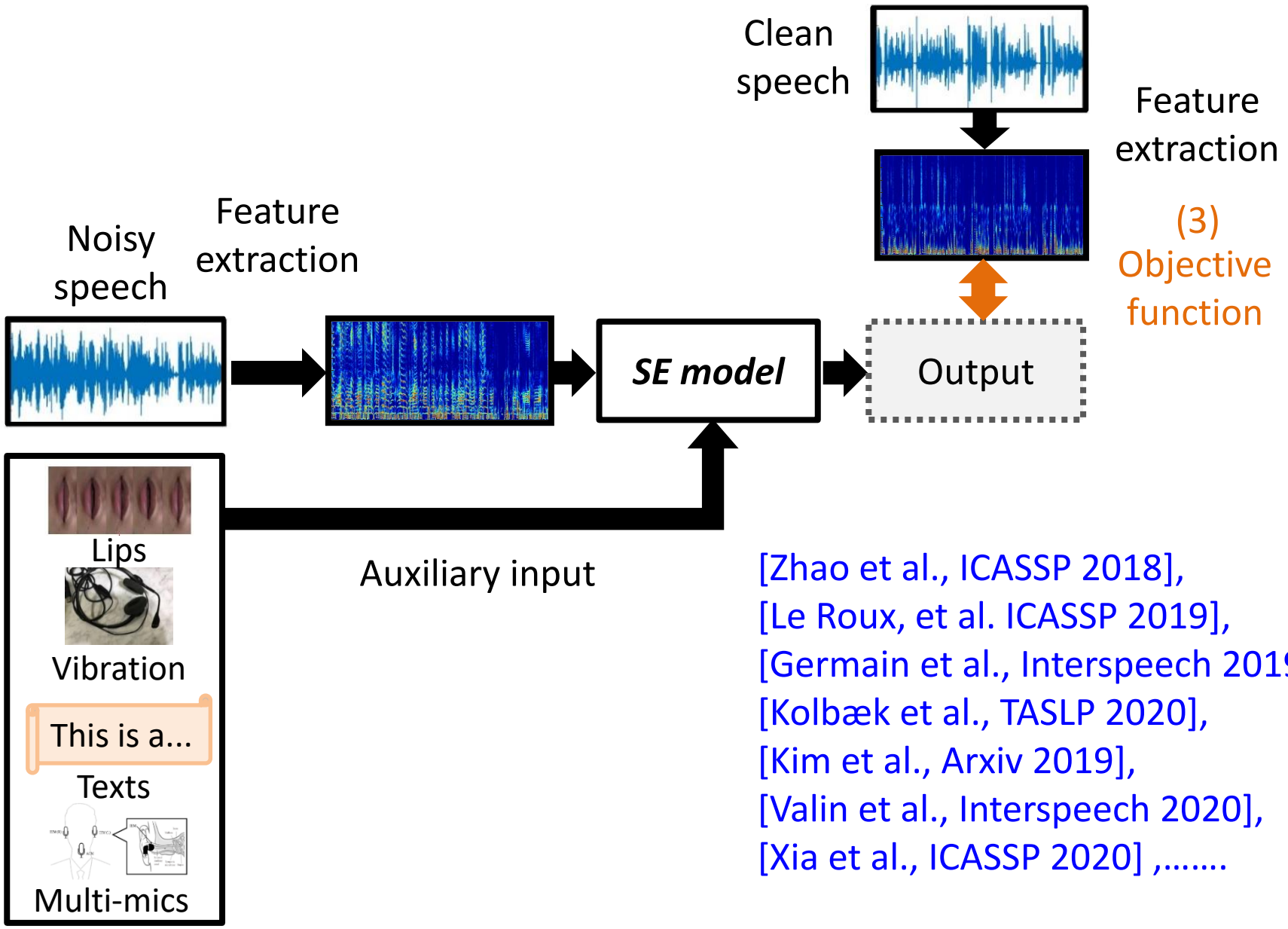


Average character error rate (CCR) and quality scores (MOS) of human subjects for (a) −3 dB and (b) −6 dB SNR.

(1) Intelligibility: FCN (MSE+STOI)> FCN (STOI)>FCN (MSE).
(2) Quality: FCN (MSE+STOI) performs the best.
(3) STOI and PESQ are not highly correlated.

Bio-ASP

Deep Learning Based SE System



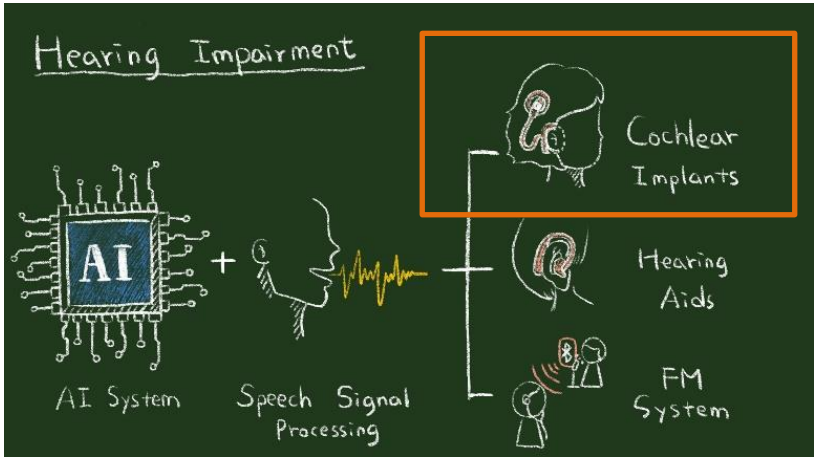
[Zhao et al., ICASSP 2018],
[Le Roux, et al. ICASSP 2019],
[Germain et al., Interspeech 2019],
[Kolbæk et al., TASLP 2020],
[Kim et al., Arxiv 2019],
[Valin et al., Interspeech 2020],
[Xia et al., ICASSP 2020] ,…….

Outline

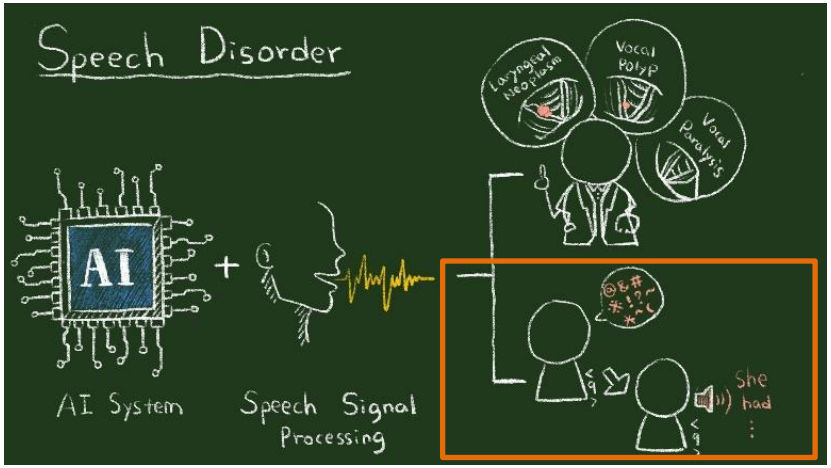
- Deep Learning (DL) based Speech Enhancement (SE)
  - Artificial intelligence and deep neural networks
  - Basic DL-based SE system architecture
  - Key factors to the DL-based SE performance
- **Assistive Oral Communication Technologies**
- Summary

Assistive Voice Communication

- Assistive listening



- Assistive speaking



Cochlear Implant

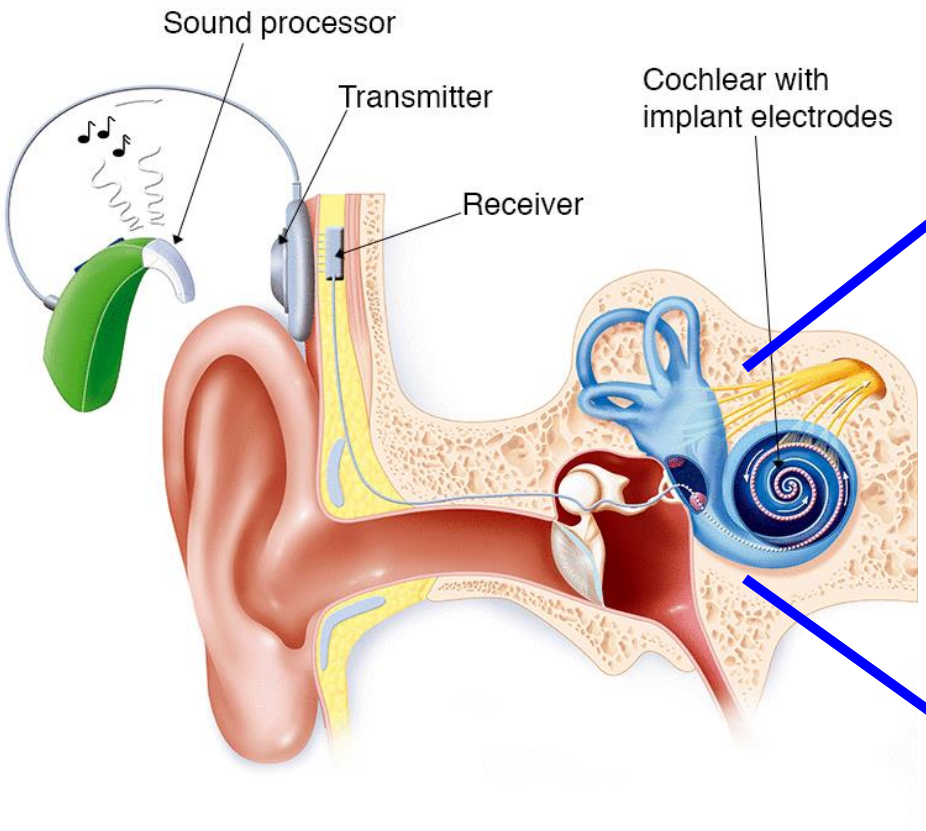
Transmitter coil

Speech processor:

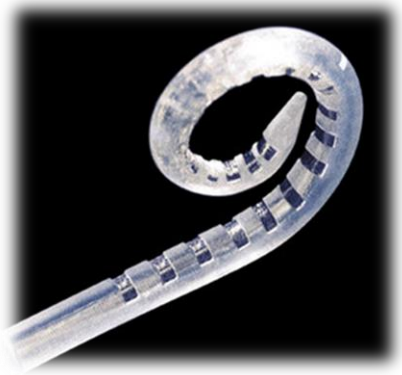
1. Microphone.
2. DSP chip.
3. Battery
4. Others…

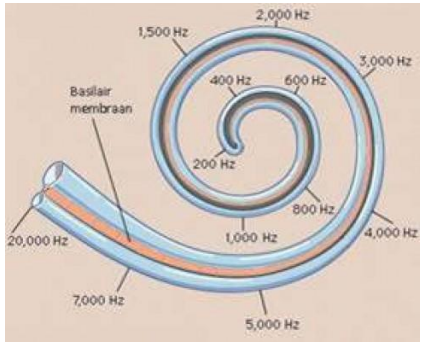
Source from:
https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cochlear-implant-surgery

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

Bio-ASP

SE for Cochlear Implant

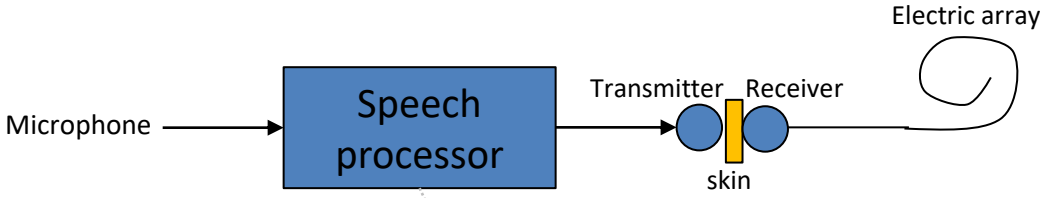
- The tremendous progress of CI technologies in the past three decades has enabled many CI users to enjoy **high level** of speech understanding **in quiet**.
- For most CI users, however, the performance of speech understanding **in noise** still remains **challenging**.

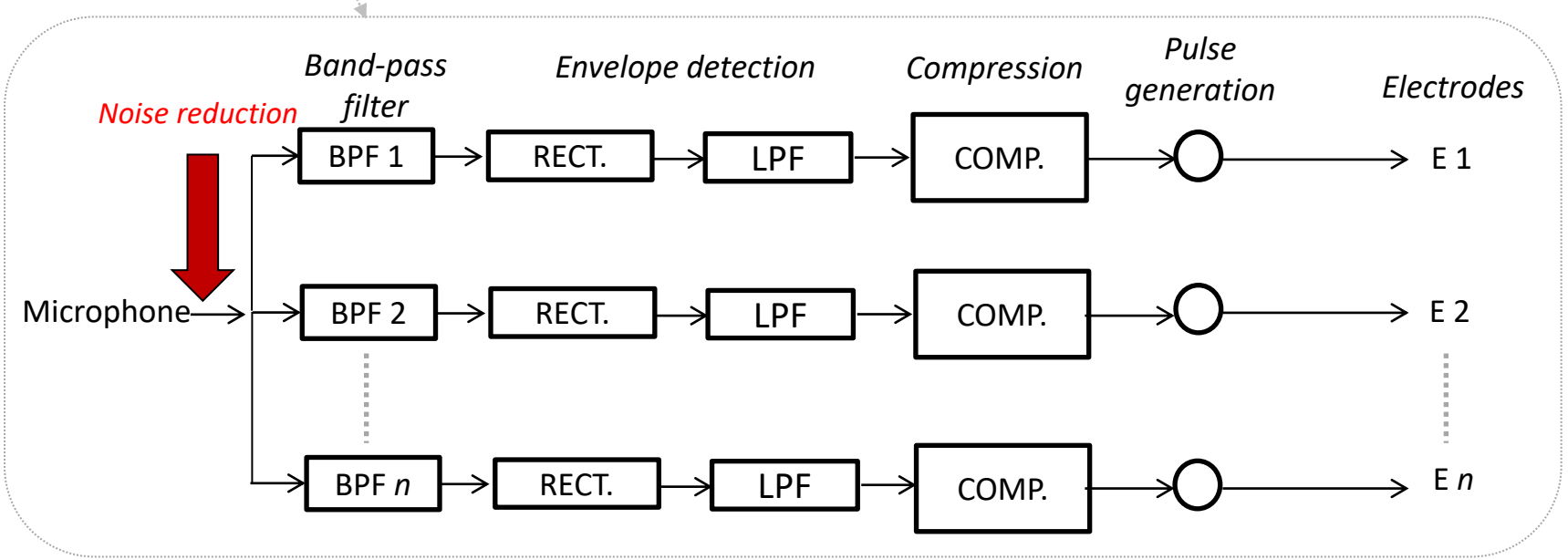
  ➢ F. Chen, Y. Hu, and M. Yuan, "Evaluation of Noise Reduction Methods for Sentence Recognition by Mandarin-Speaking Cochlear Implant Listeners," Ear and hearing, vol. 36, no. 1, pp. 61-71, 2015.

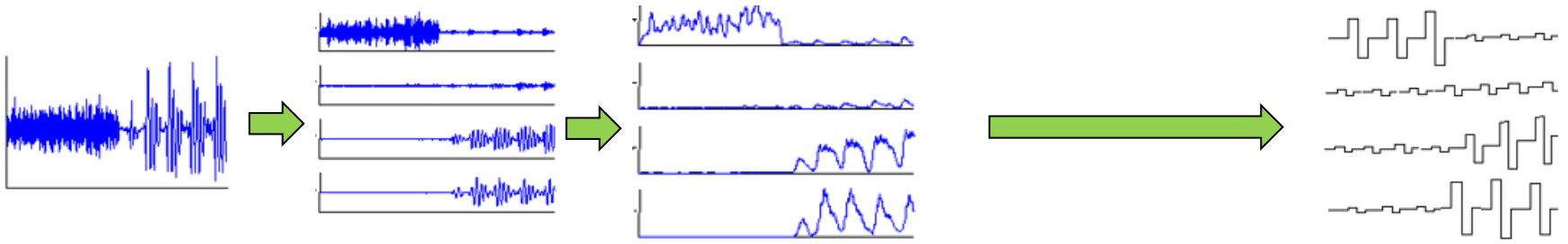
- **Deep learning** based speech enhancement (SE) for CI.



SE for Cochlear Implant

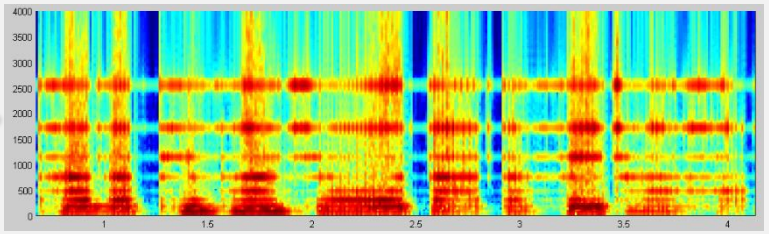




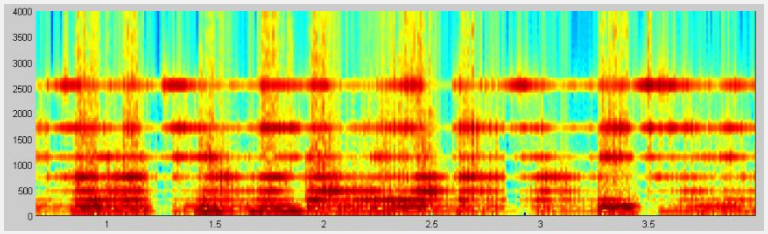


SE for Cochlear Implant Simulation

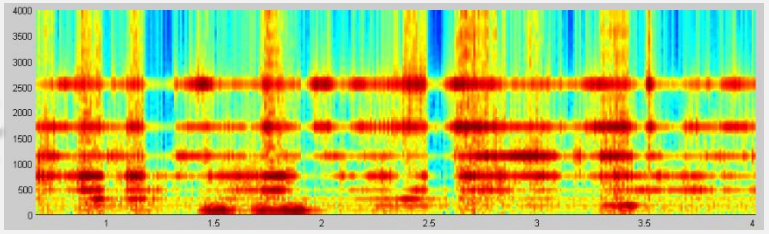
- Vocoded speech



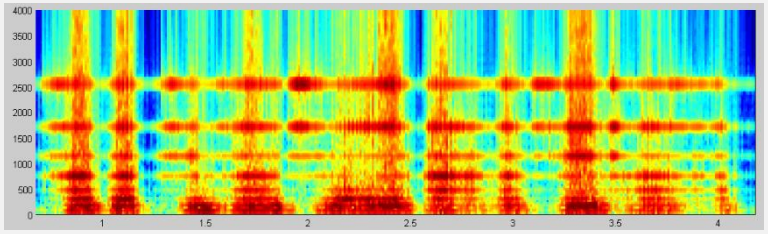
Clean



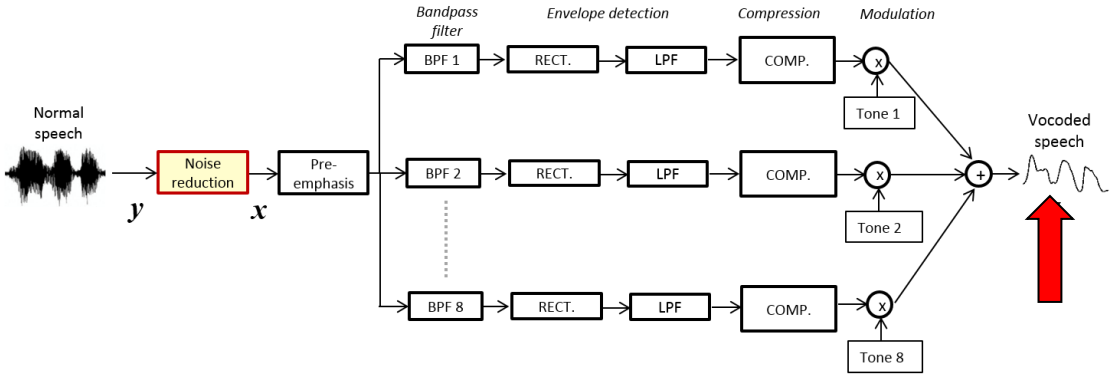
2T Noise 0dB



MMSE

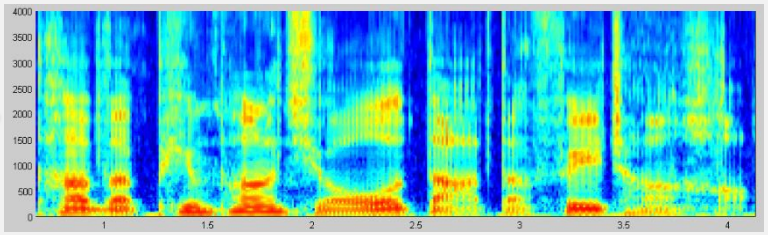


DDAE

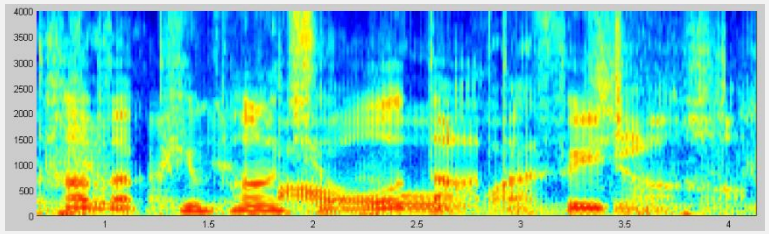


SE for Cochlear Implant Simulation

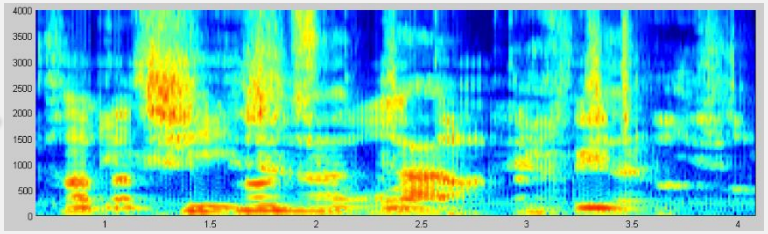
- Normal speech



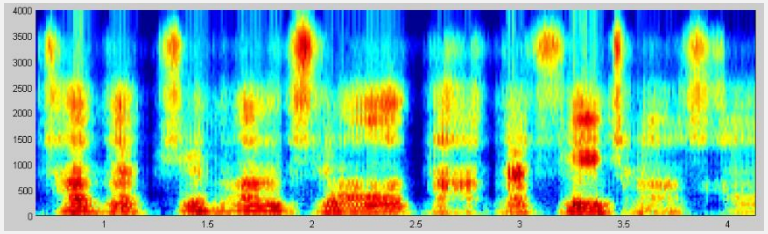
Clean



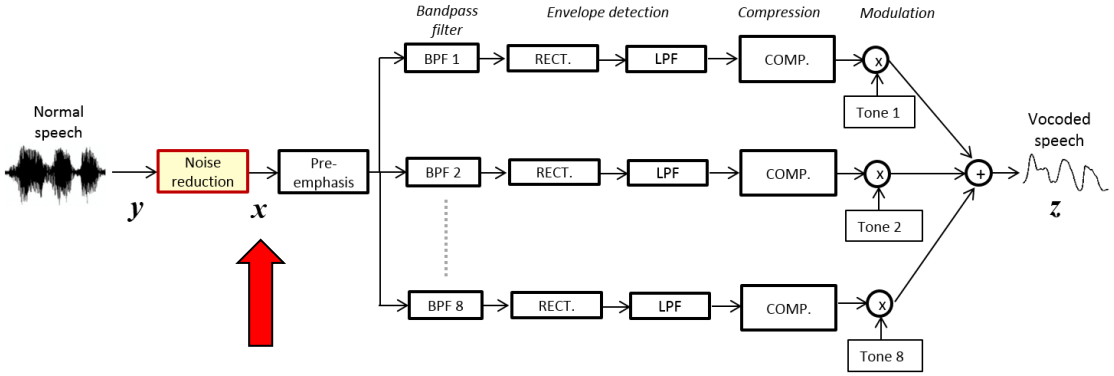
Babble Noise 0dB



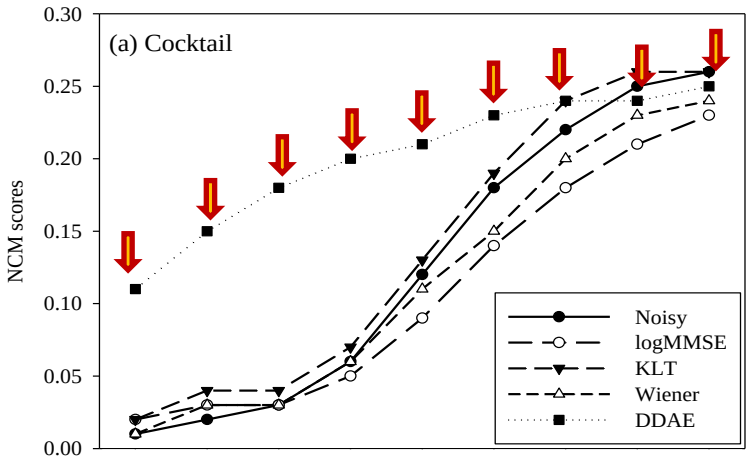
MMSE



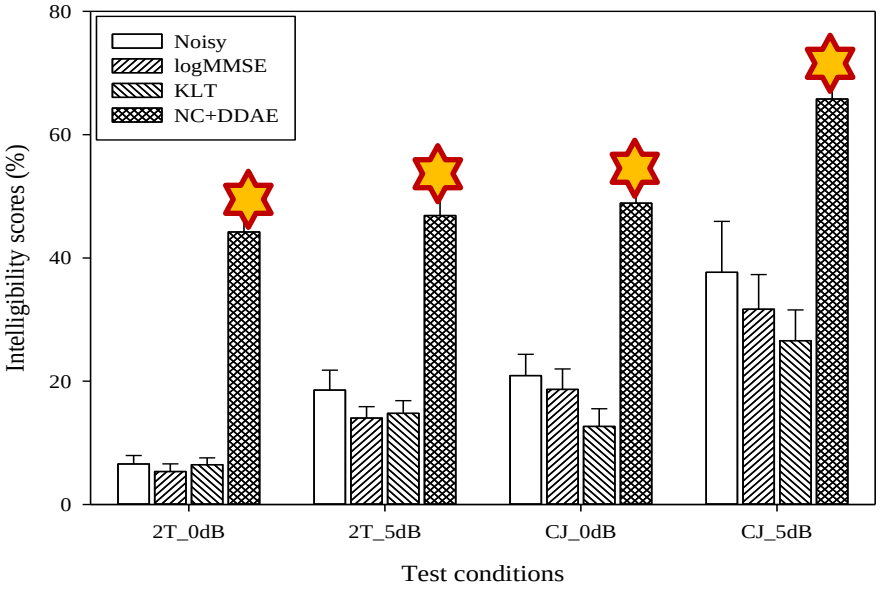
DDAE



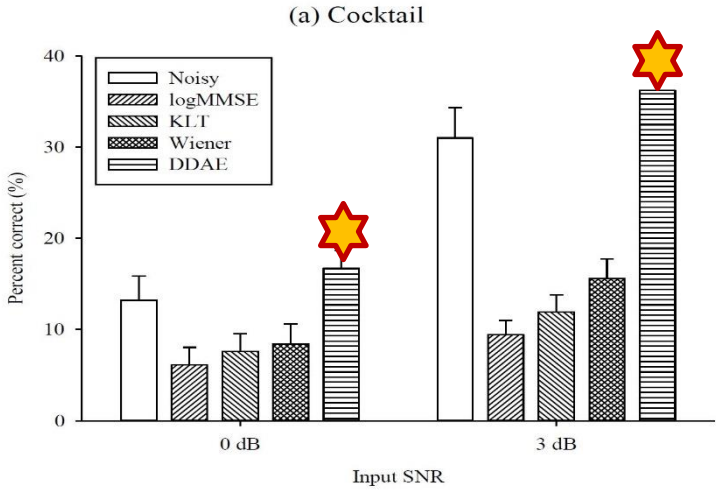
Evaluation Results



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.

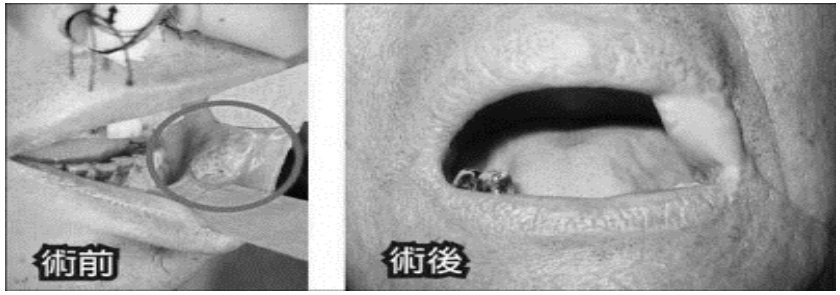
文言版《説文解字》：訥，言難也。

Speaking disorders

Dysarthria, apraxia, aphasia, stuttering,
oral surgery, vocal damage

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..

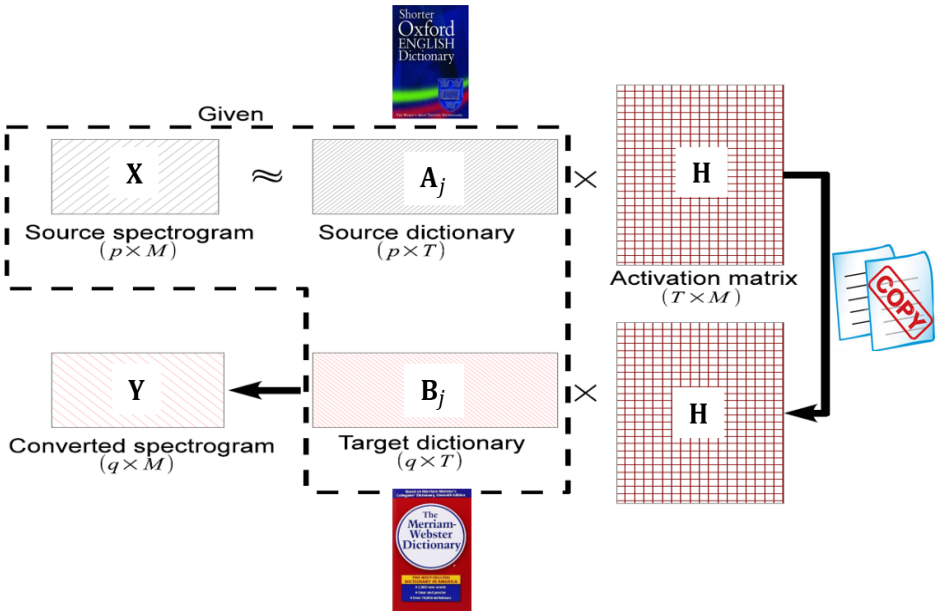
Before After

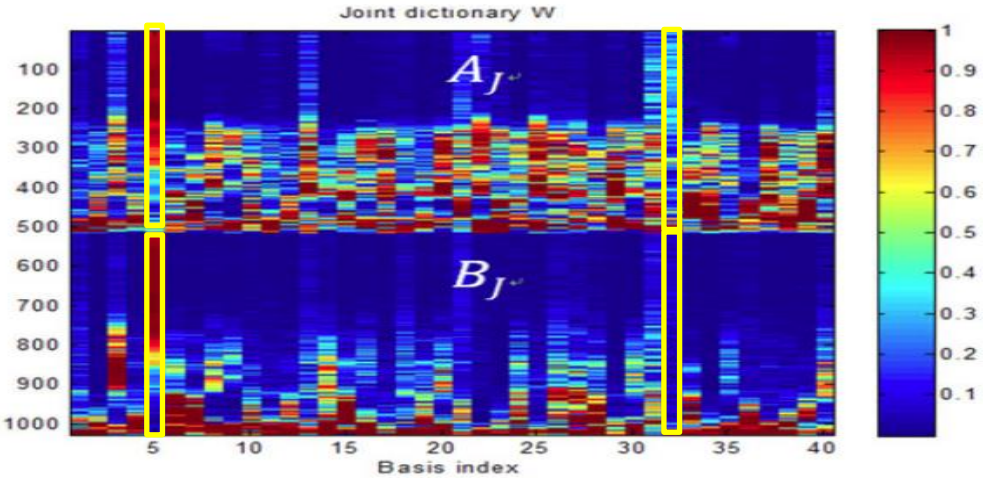
Taipei Veterans General Hospital

Bio-ASP

SE for Speaking Disorder

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

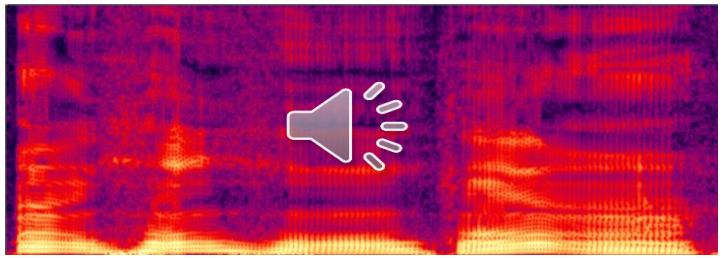


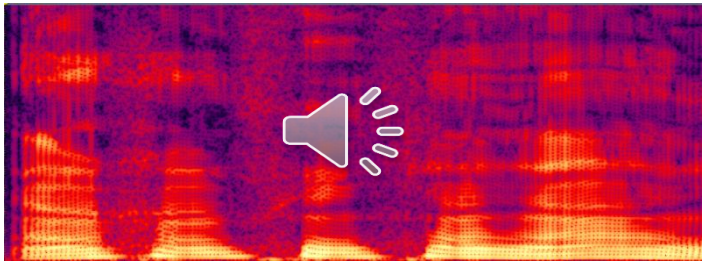


Bio-ASP

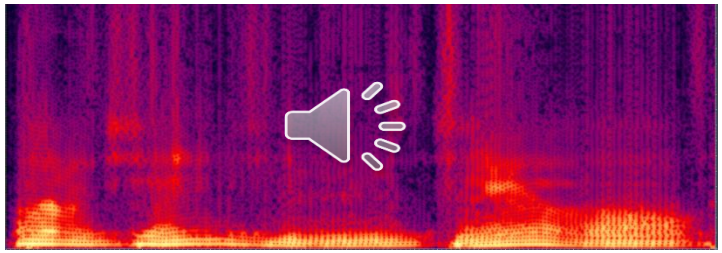
Testing Results

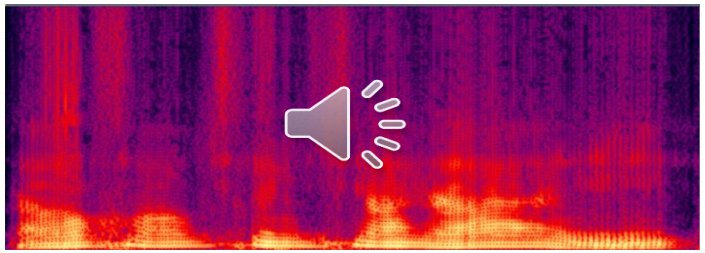
Original:





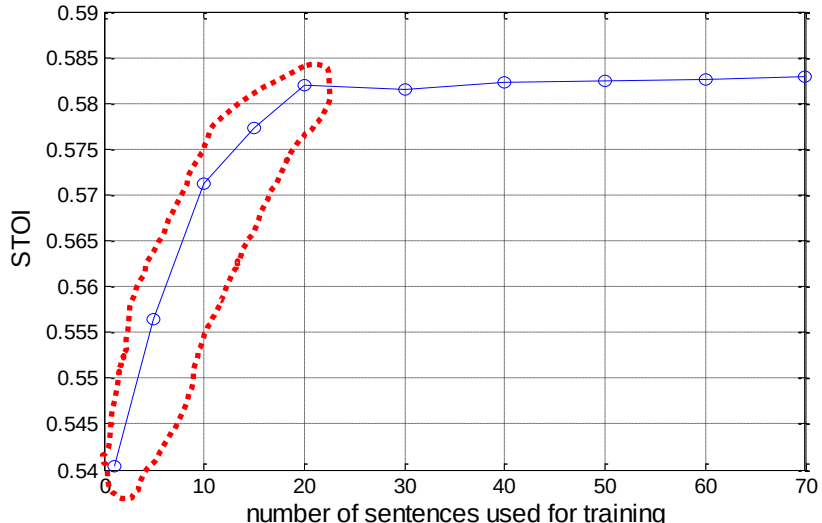
After Conversion:





衛生紙給我

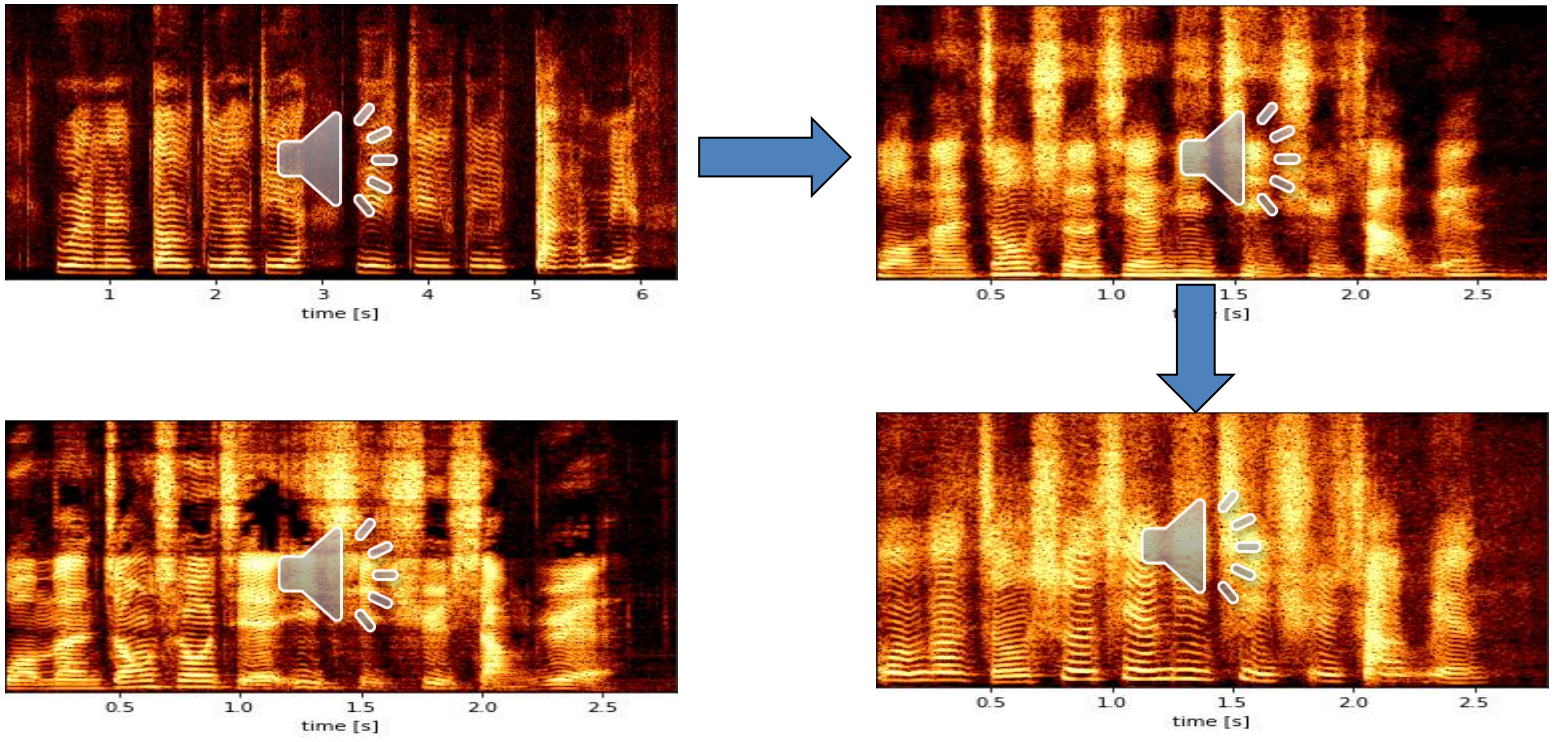
遙控器在哪裡



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

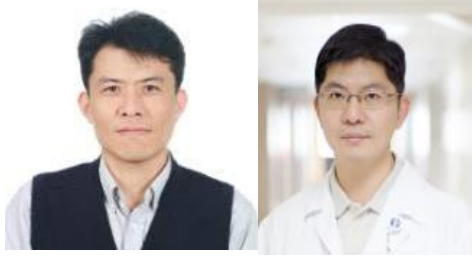
GAN-based solution
[Chen et. al., Interspeech 2019]

Dysarthric Voice Conversion



我們中秋節一起去賞月

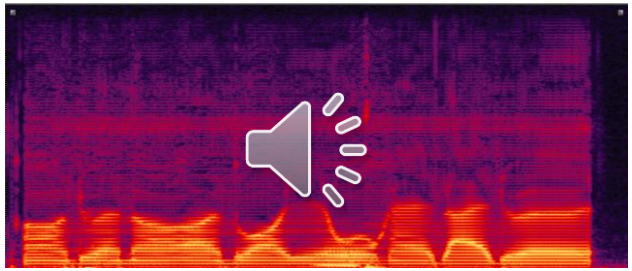
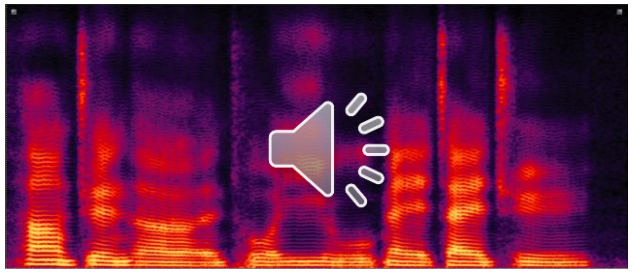
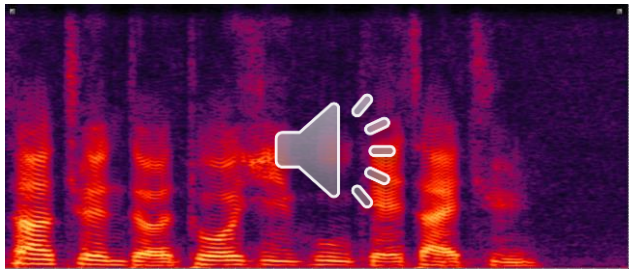
- W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, T. Toda, "A Preliminary Study of a Two-Stage Paradigm for Preserving SpeakerIdentity in Dysarthric Voice Conversion," Interspeech 2021.



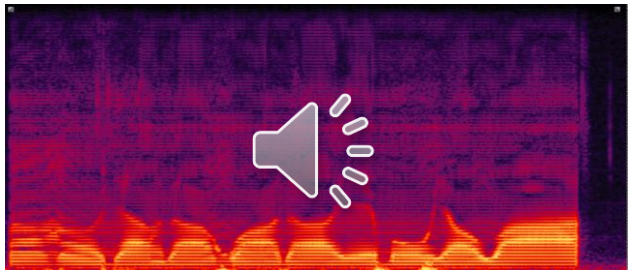
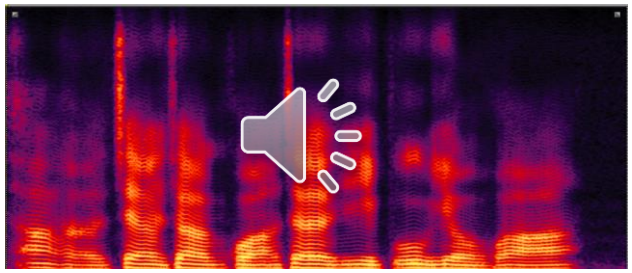
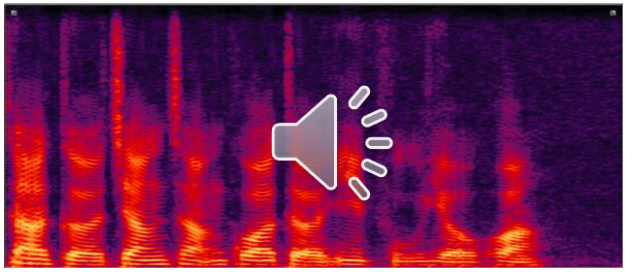
Electrolaryngeal Voice Conversion

Original | MT-CLDNN | Seq2seq ELVC

Sample 1: 他捐了很多衣物給災區

Sample 2: 那個牆上掛著一幅油畫

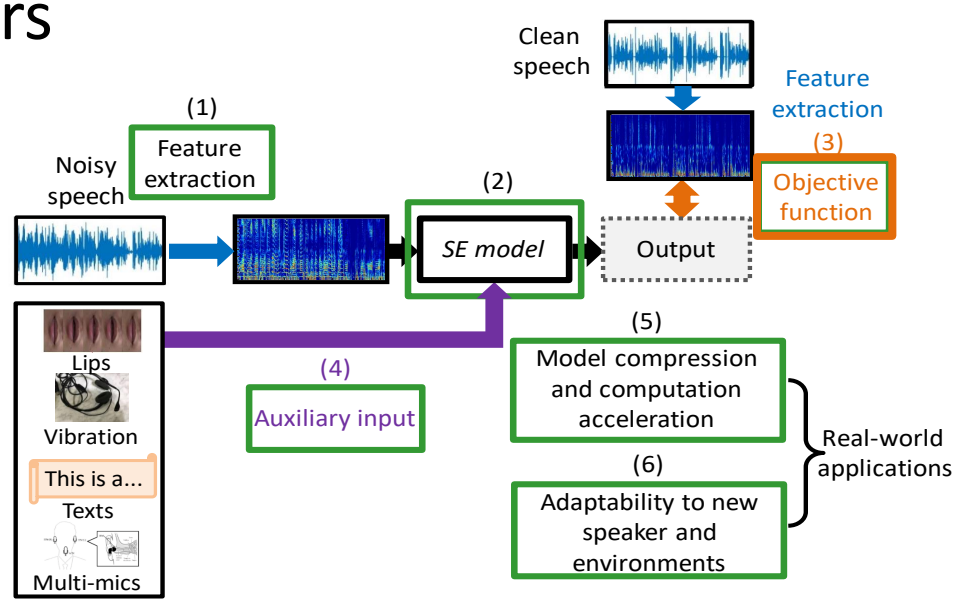
- M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. R. Jang, and H.-M. Wang, “Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling, ASRU 2021”


Outline

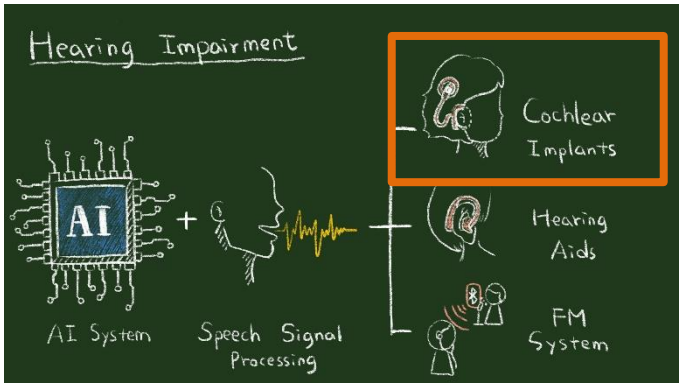
- Deep Learning (DL) based Speech Enhancement (SE)
  - Artificial intelligence and deep neural networks
  - Basic DL-based SE system architecture
  - Key factors to the DL-based SE performance
- Assistive Oral Communication Technologies
- **Summary**

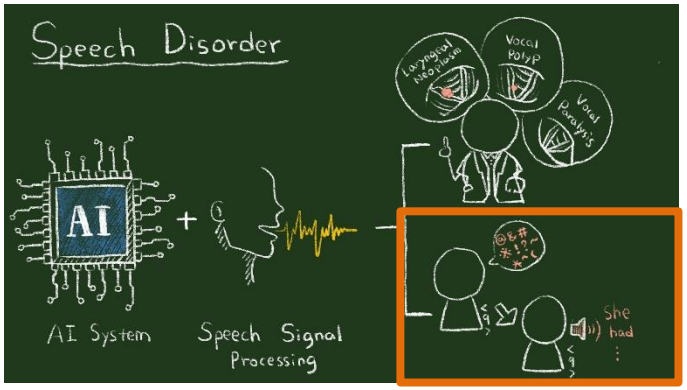
Summary

- Key Factors



- Assistive Oral Communication Technologies





Bio-ASP

Artificial Intelligent Assistance (AIA)

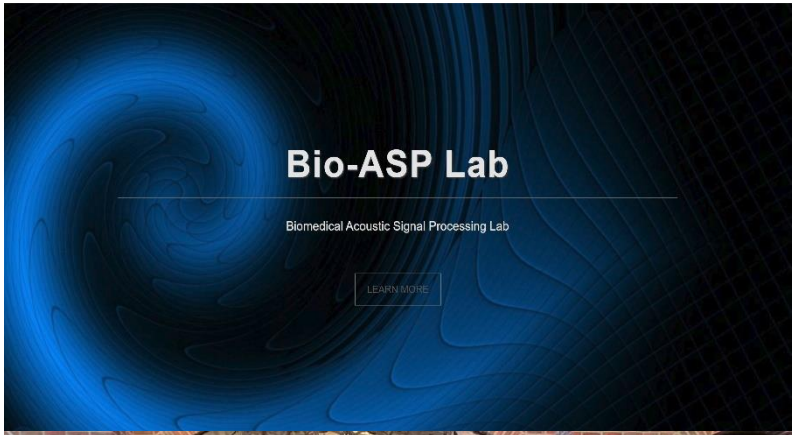
Artificial Intelligence (AI)

歡迎加入我們來做有溫度的科學研究

From Internet.

Bio-ASP

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

Thank You Very Much for Your Attention