Deep learning and its applications to speech processing: recognition and generation

Dr. Yu Tsao 曹昱, Associate Research Fellow
Research Center for IT Innovation, Academia Sinica

2019/11/17

Dr. Yu Tsao (曹昱), *Associate Research Fellow*

- Education

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

- Work Experience

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

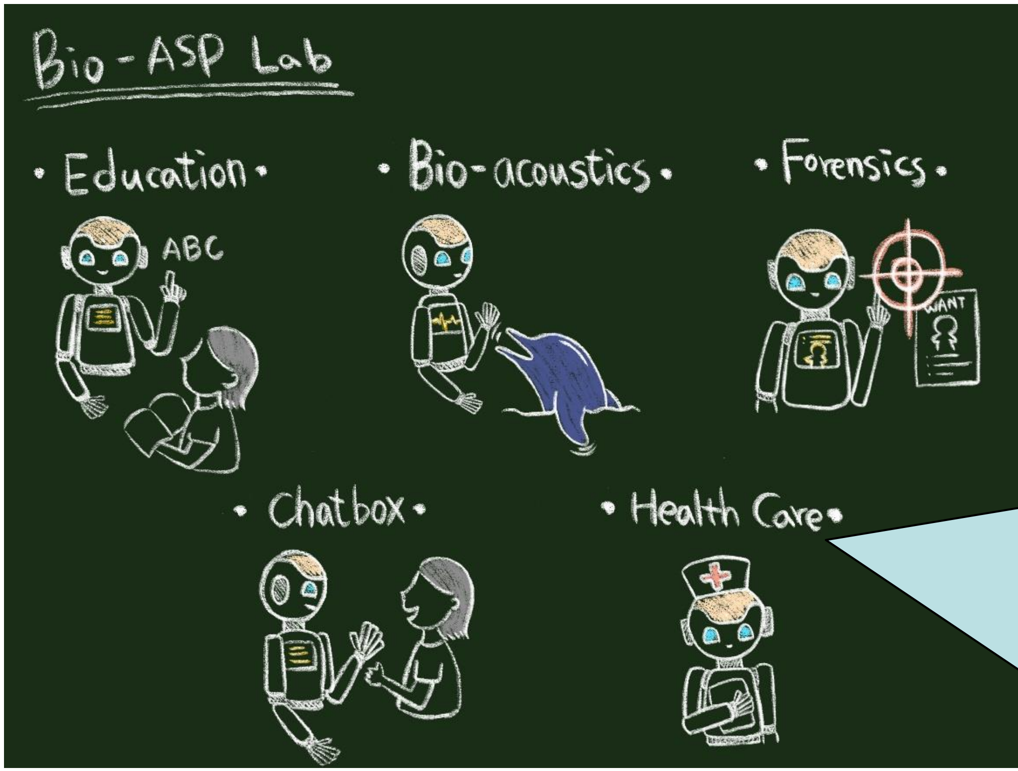
- Research Interests

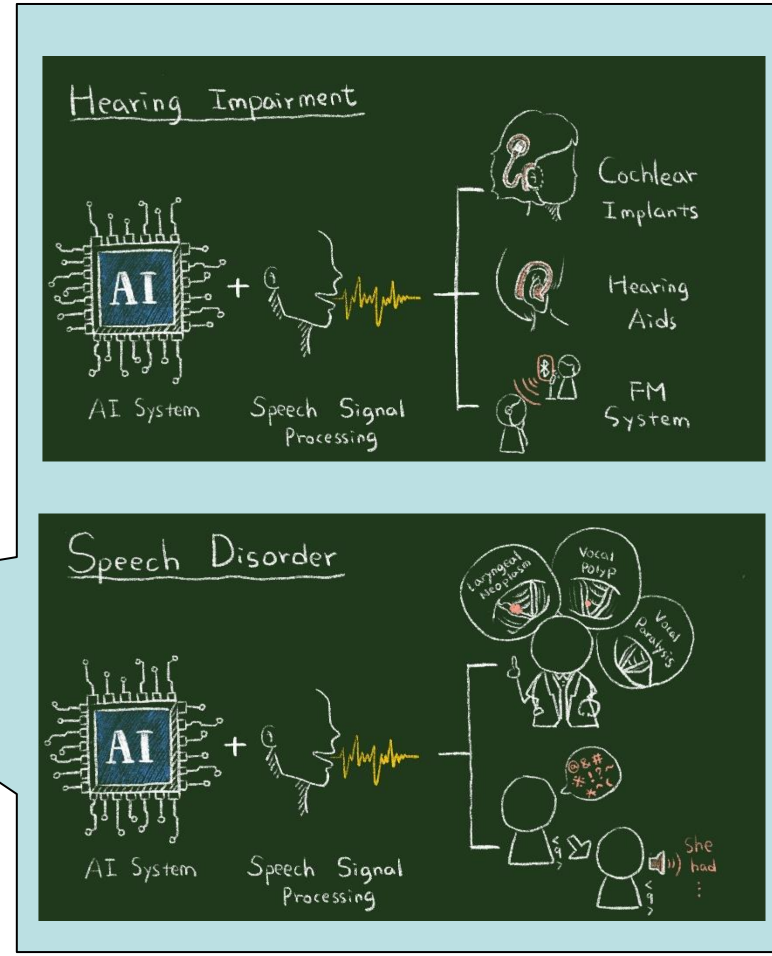
  Speech & Audio Signal Processing, Machine Learning and Pattern Recognition, Speech and Speaker Recognition

- Lab at CITI (Academia Sinica)

  Biomedical Acoustic Signal Processing (Bio-ASP) Lab

(Bio-ASP) Lab


Bio-ASP Lab
• Education •
ABC
• Bio-acoustics •
• Forensics •
WANT
• Chatbox •
• Health Care •


Hearing Impairment
AI
+
AI System
Speech Signal Processing
Cochlear Implants
Hearing Aids
FM System
Speech Disorder
AI
+
AI System
Speech Signal Processing
Laryngeal Neoplasm
Vocal Polyp
Vocal Paralysis
She had

Outline

- Deep Learning
  - Artificial intelligence, machine learning, deep learning
  - Human learning versus machine learning
  - Some histories about deep learning
  - Popular deep learning models
- Speech Signal Processing
  - Two categories of tasks: recognition and generation
  - Recognition: pathological voice recognition
  - Generation: speech enhancement

聲

言

兌

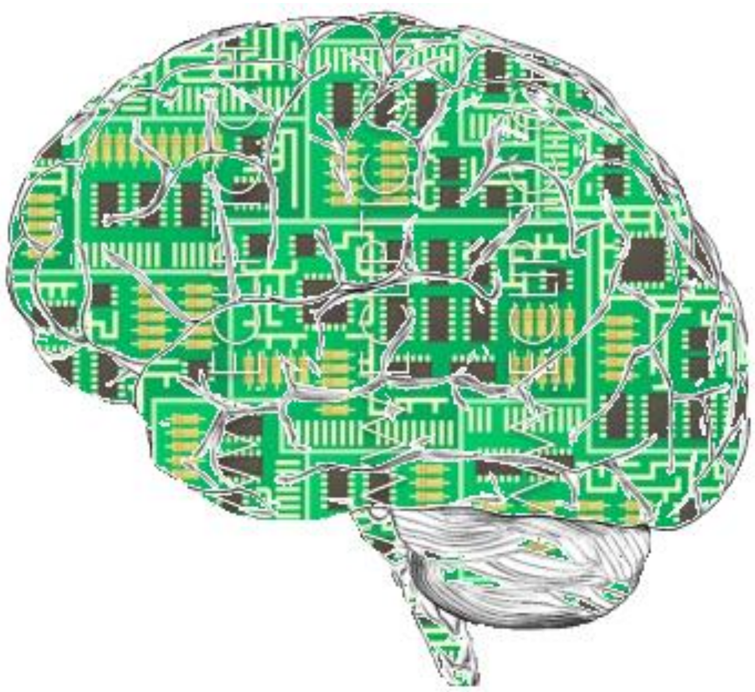
《說文解字》：兌，說也。

說

Outline

- Deep Learning
  - Artificial intelligence, machine learning, deep learning
  - Human learning versus machine learning
  - Some histories about deep learning
  - Popular deep learning models
- Speech Signal Processing
  - Two categories of tasks: recognition and generation
  - Recognition: pathological voice recognition
  - Generation: speech enhancement

Machine Learning and Artificial Intelligence

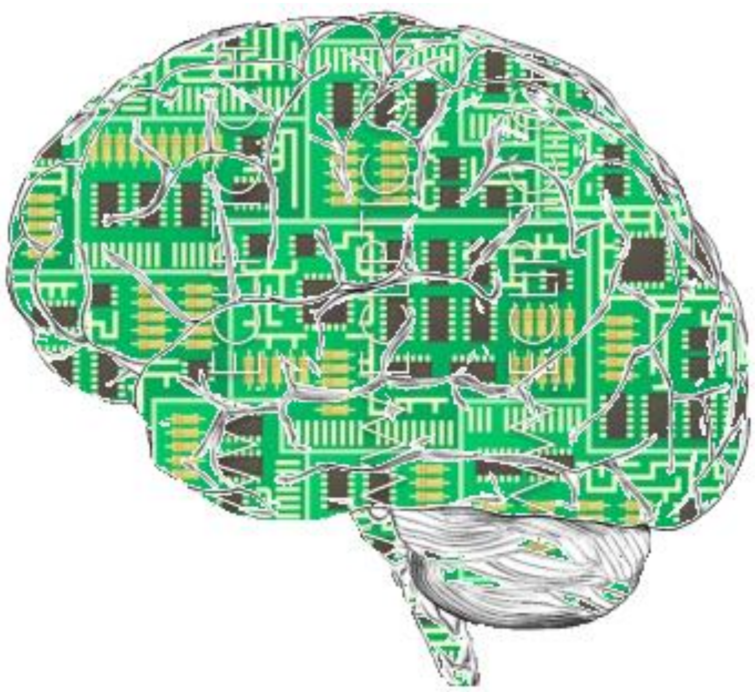


Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

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

From wiki

Machine Learning and Artificial Intelligence

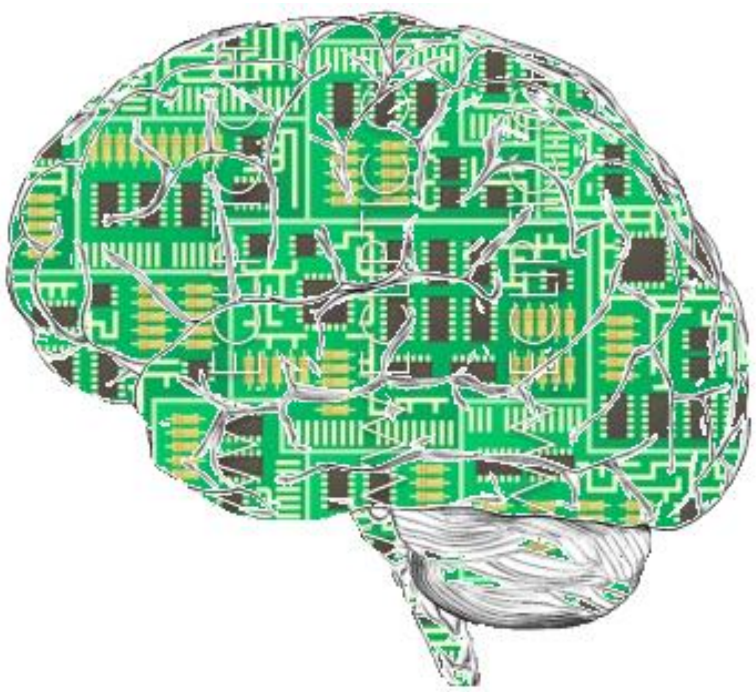


Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

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

From wiki

Machine Learning and Artificial Intelligence



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

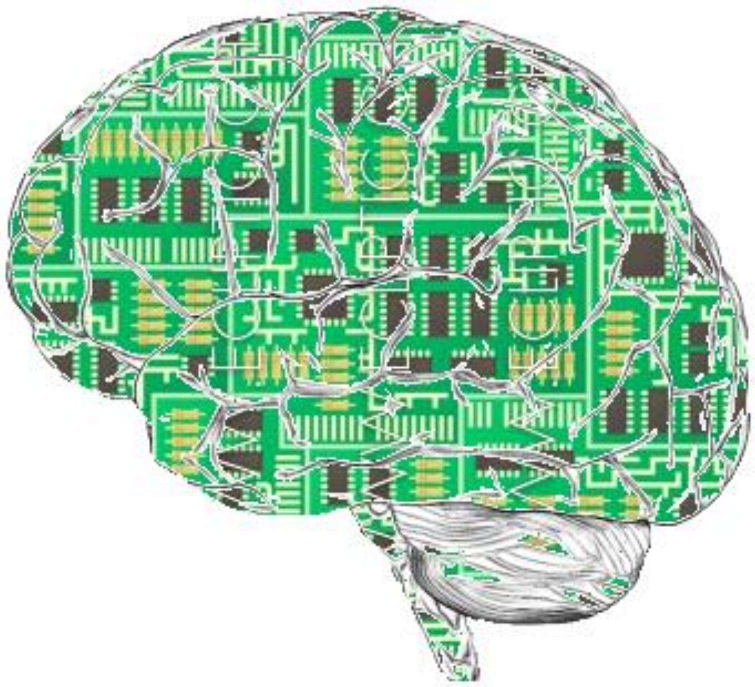
1. Deduction, reasoning, problem solving
2. Knowledge representation
3. Default reasoning and the qualification problem
4. Machine planning
5. **Machine learning**

⋮

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

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

Machine Learning and Artificial Intelligence



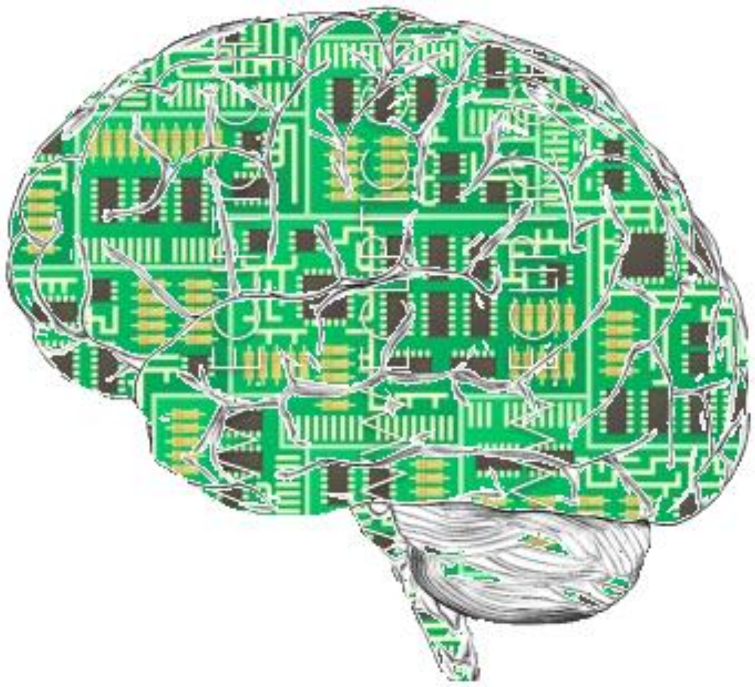
Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

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
Neural networks
Kernel methods
Sparse kernel machines

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

Machine Learning and Artificial Intelligence



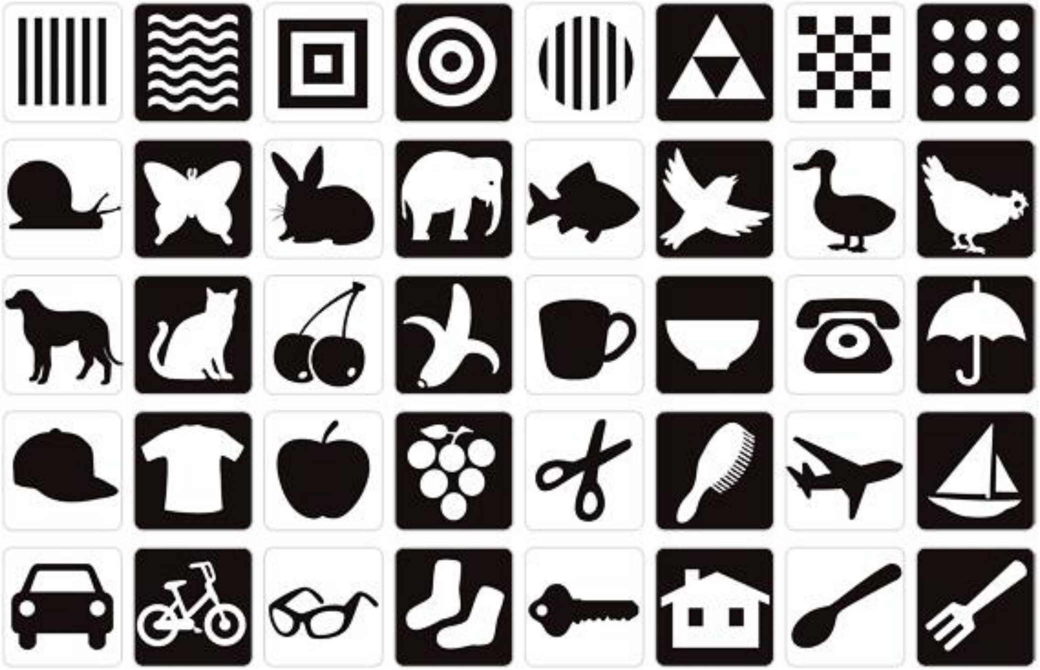
Artificial intelligence(AI) is intelligence exhibited by machines, mainly covers:

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

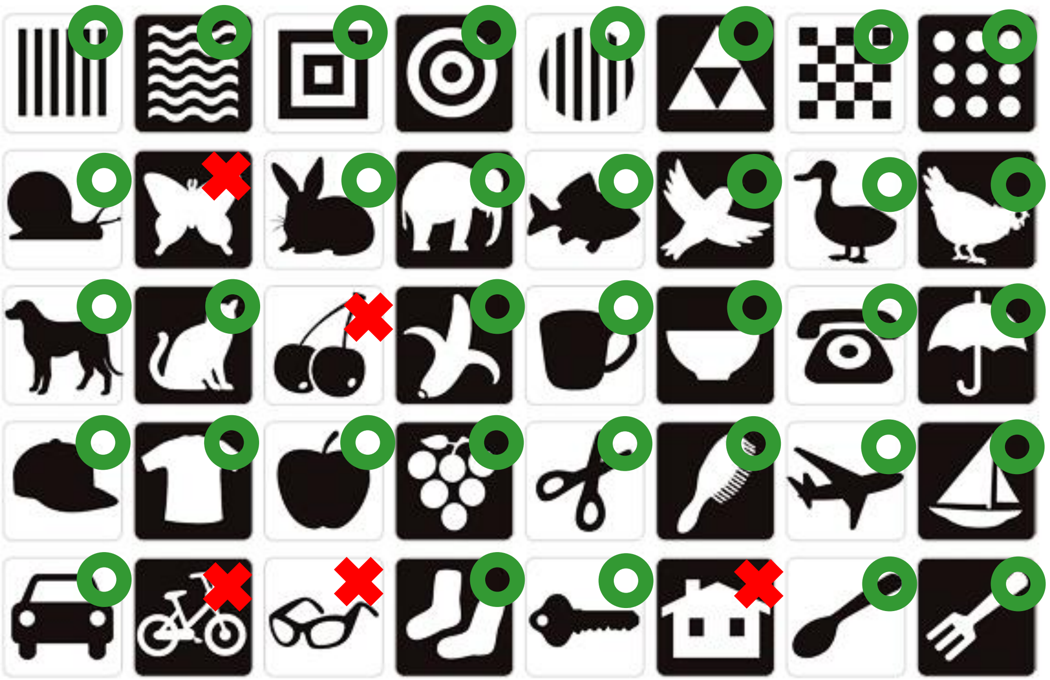
From M. Svensen & C. Bishop, "Pattern recognition and machine learning"

Human Perception: Classification



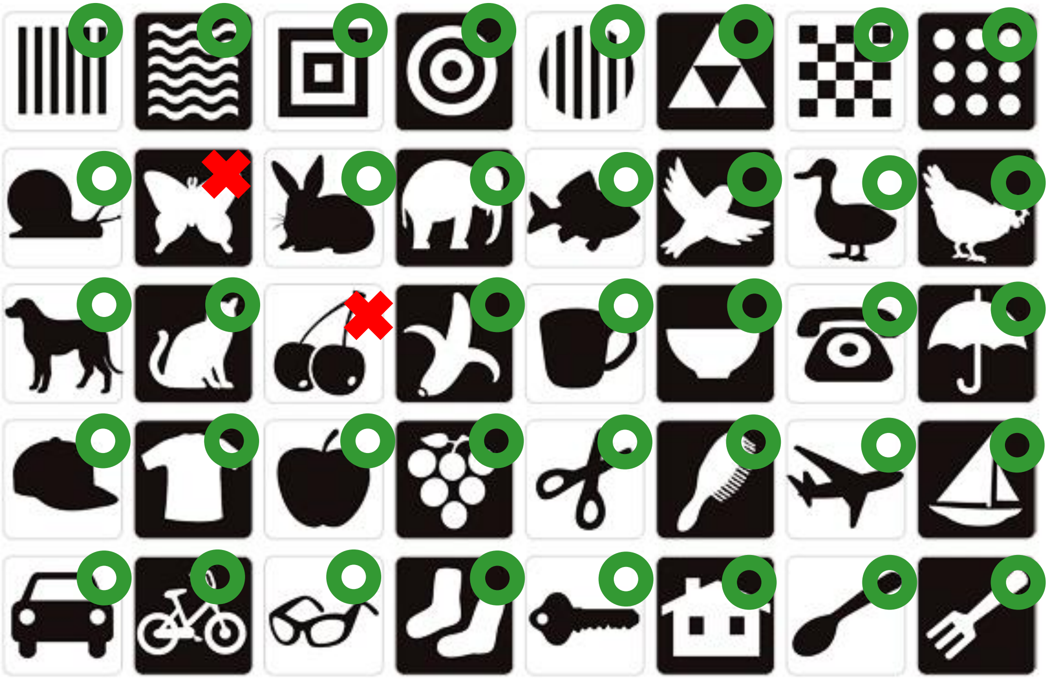
Discriminative learning

Human Perception: Classification



Performance Evaluation: Accuracy=(40-5)/40

Human Perception: Classification



Performance Evaluation: Accuracy=(40-2)/40

Training

Human Perception: Regression

Human Perception: Regression

Performance Evaluation: Correlation

Human Learning: Exemplar Theory

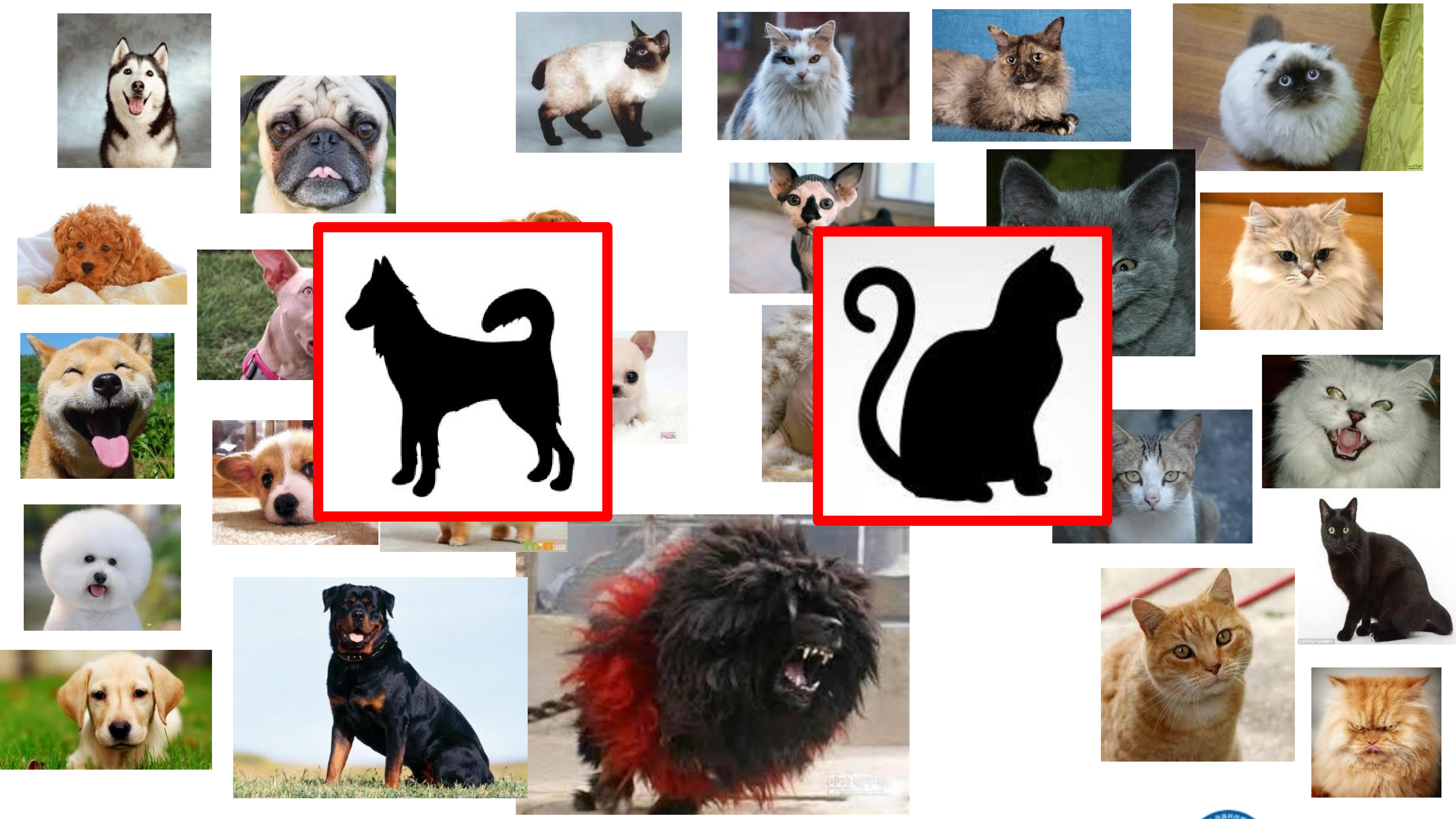
Human Learning: Exemplar Theory

Human Learning: Exemplar Theory

Human Learning: Exemplar Theory

Machine Learning:
Discriminative models, such as:
- Support vector machine (SVM),
- Artificial neural networks (ANN),
- Deep neural network (DNN).

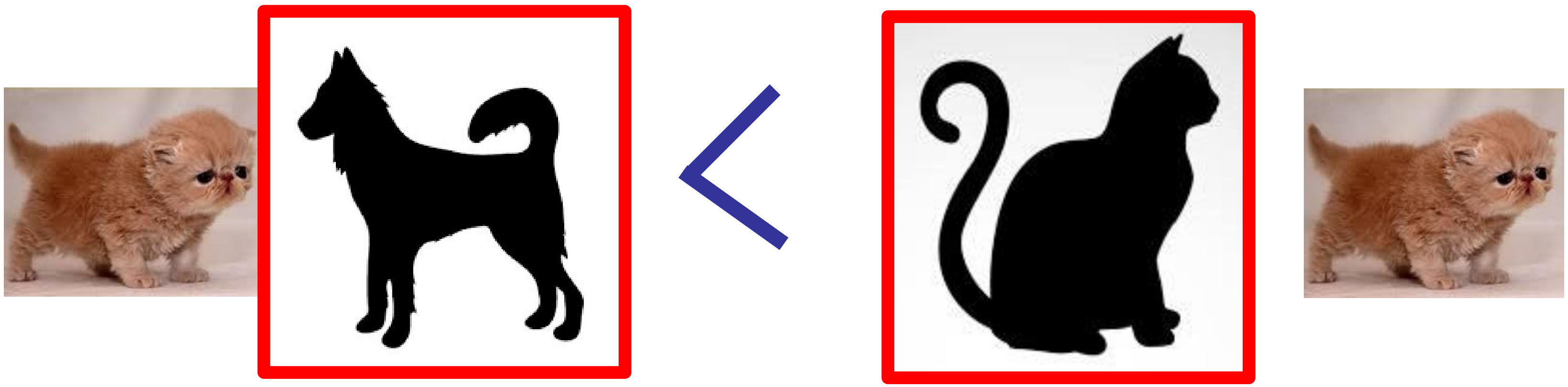
Human Learning: Prototype



Human Learning: Prototype

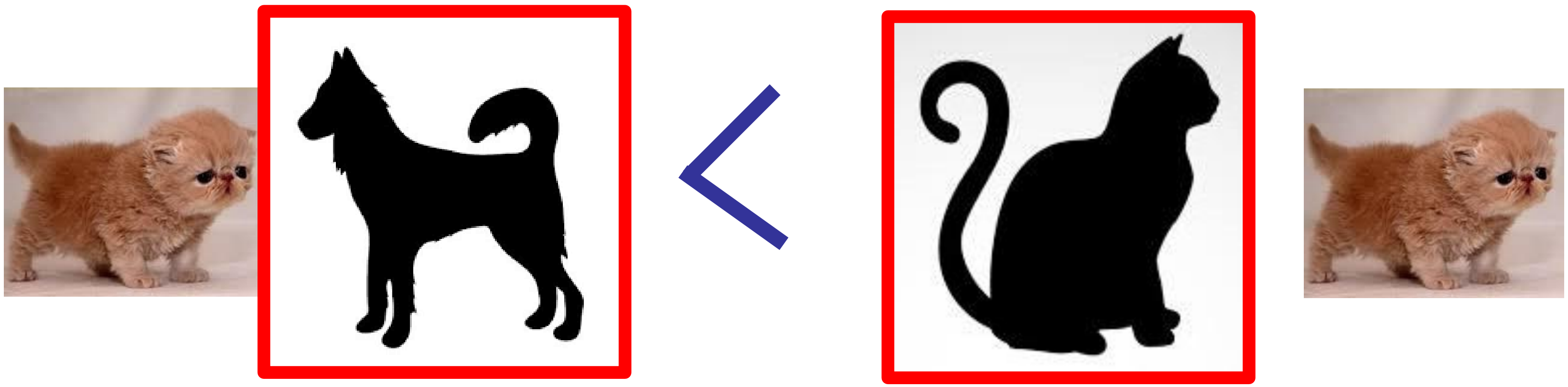
Human Learning: Prototype

$P(\mathrm{O}|\Lambda_{\mathrm{dog}})$ $P(\mathrm{O}|\Lambda_{\mathrm{cat}})$



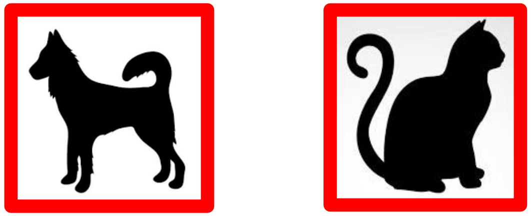
Human Learning: Prototype

$P(\mathrm{O}|\Lambda_{\mathrm{dog}})$ $P(\mathrm{O}|\Lambda_{\mathrm{cat}})$

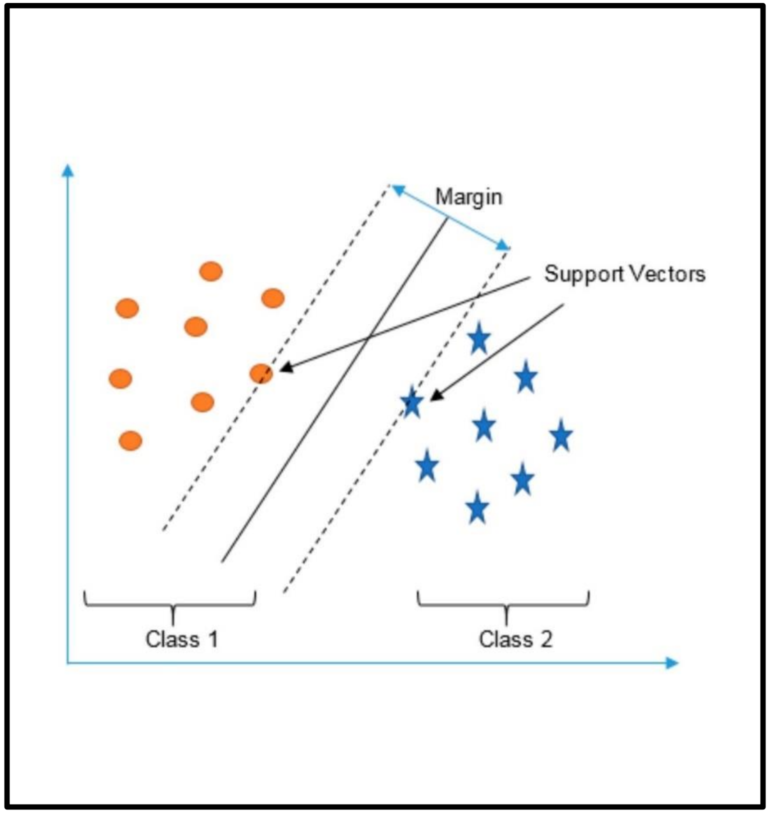


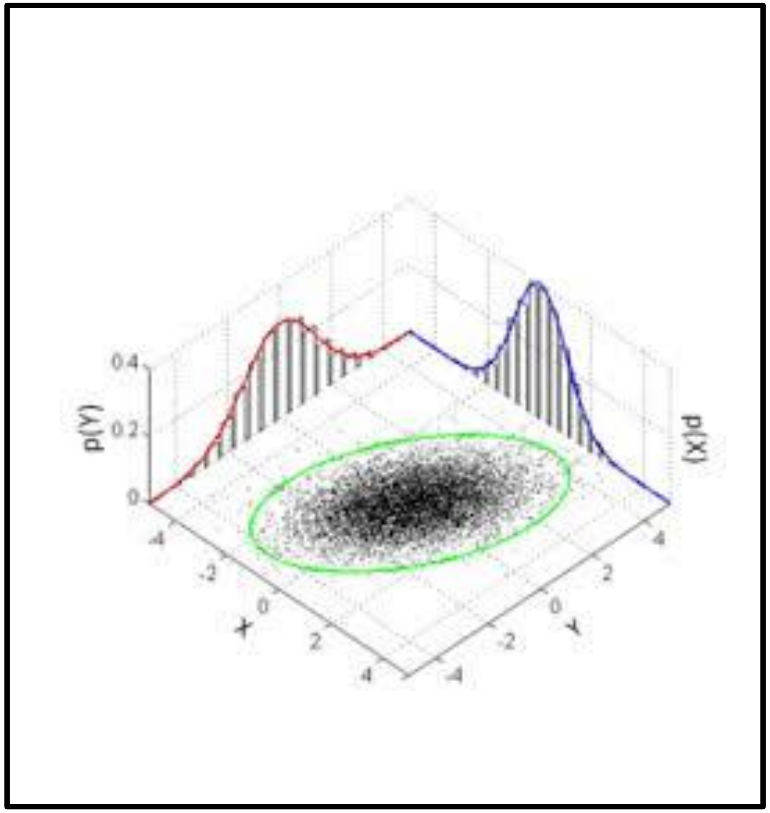
Machine Learning:
Generative models, such as:
- Gaussian mixture models (GMM),
- Restricted Boltzmann machine (RBM),
- Deep belief network (DBN).

Human Learning: Prototype vs. Exemplar



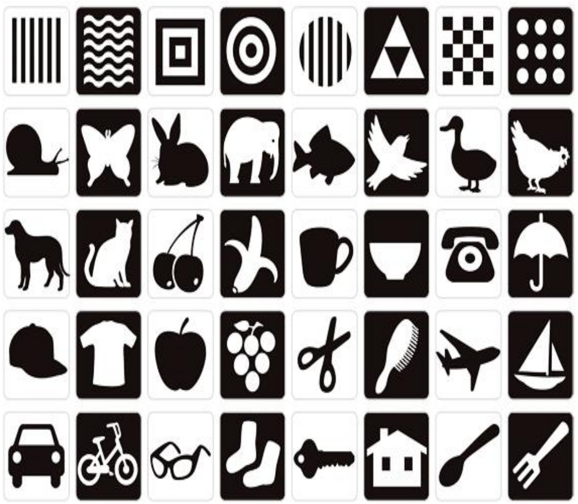
Machine Learning: Discriminative vs. Generative Models





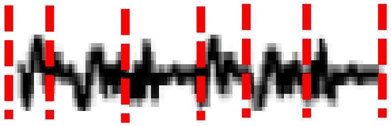
Machine Learning: Data and Labels

Image

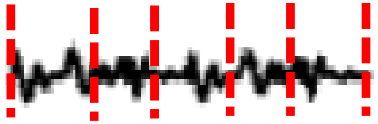


Speech

I want to play a game

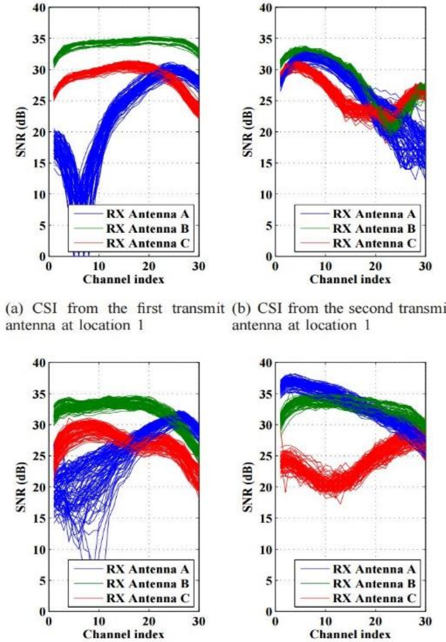


The move is so scary



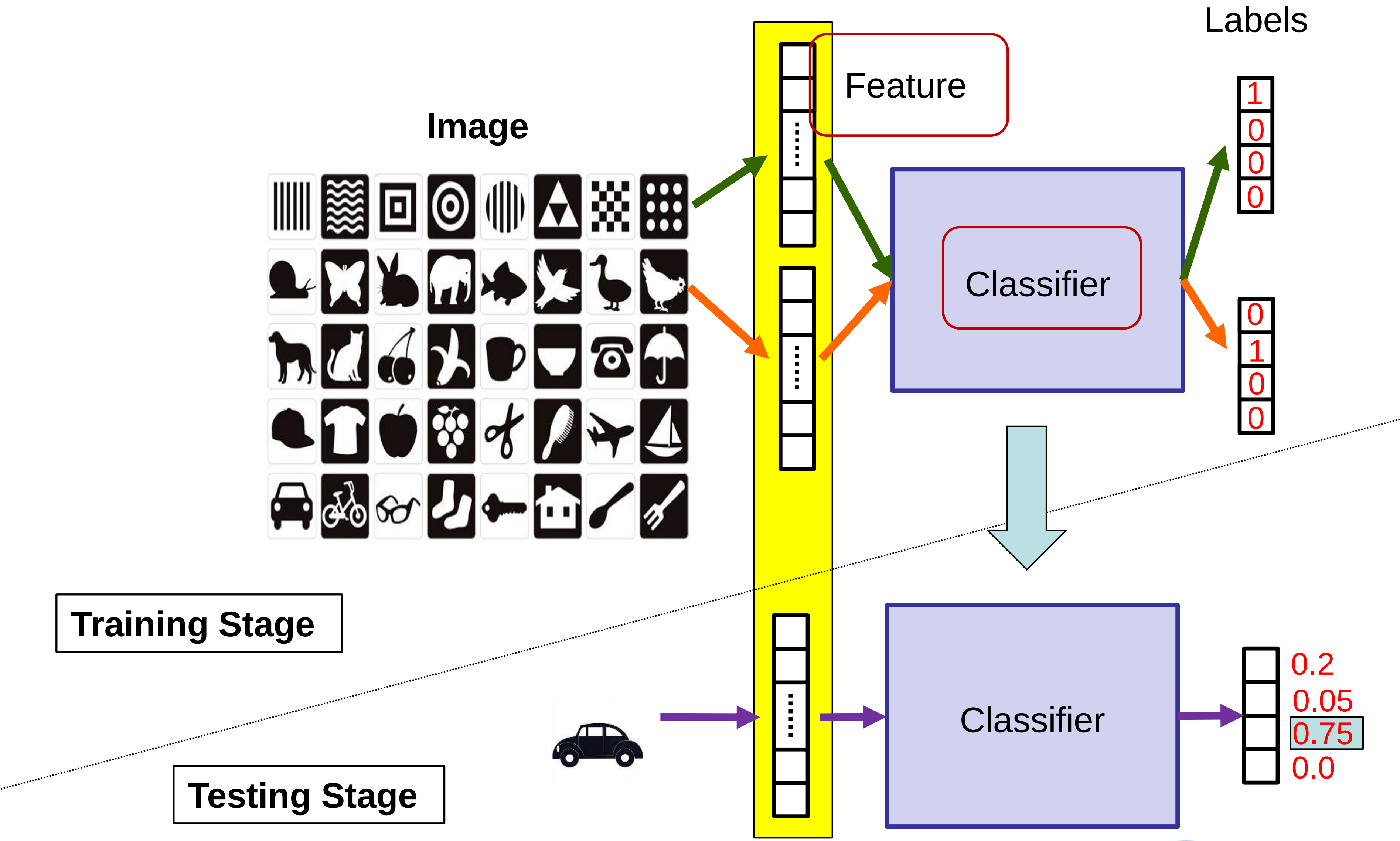
He enjoys watching it

WIFI



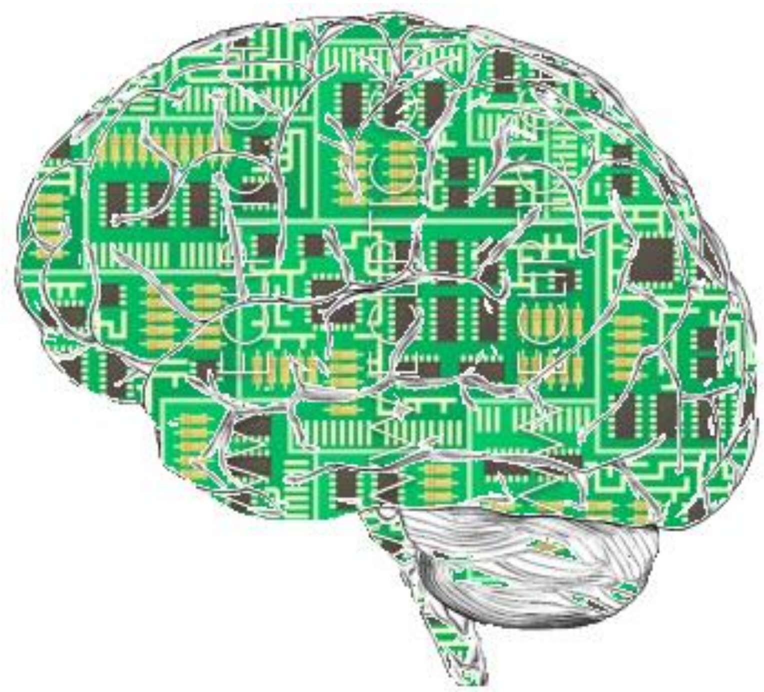
(a) CSI from the first transmit antenna at location 1 (b) CSI from the second transmit antenna at location 1

Machine Learning: Data and Labels



Labeling Error

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.

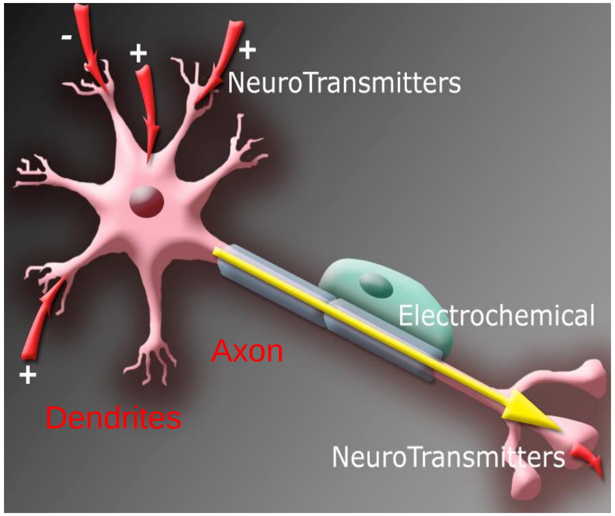
From Youtube

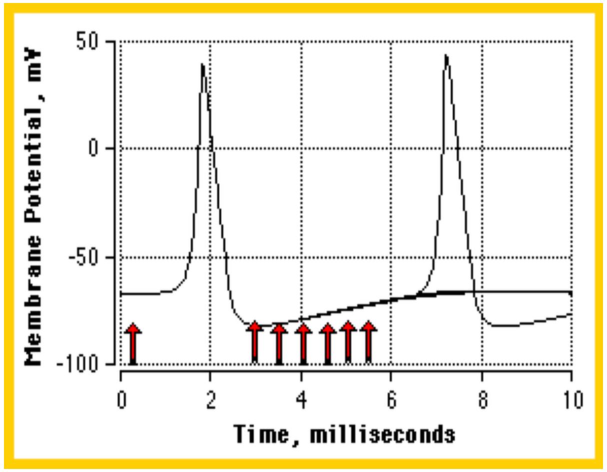
Artificial Neural Network (ANN)

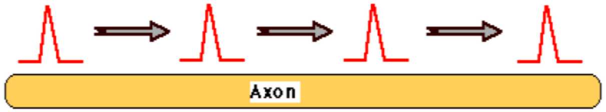
- Artificial neural network (ANN) is a **computational model** that **mimics** brain functionality with artificial means.

Adrian, 1932

(1) Receptive
(2) Trigger
(3) Conducting
(4) Output



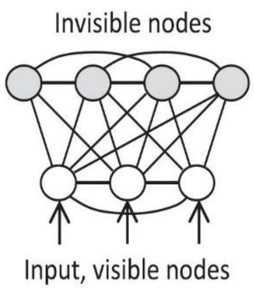




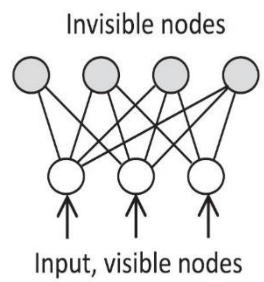
"All-or-None Nerve Firing"

Artificial Neural Network (ANN)

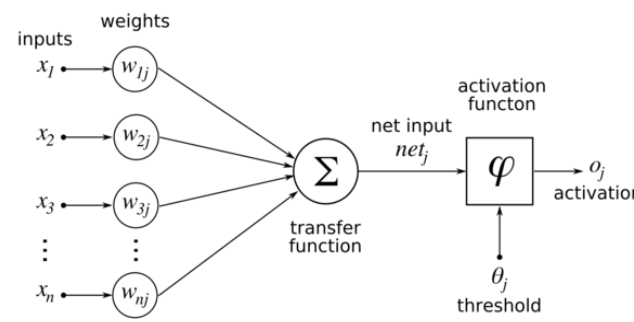
- Artificial neural network (ANN) is a **computational model** that **mimics** brain functionality with artificial means.



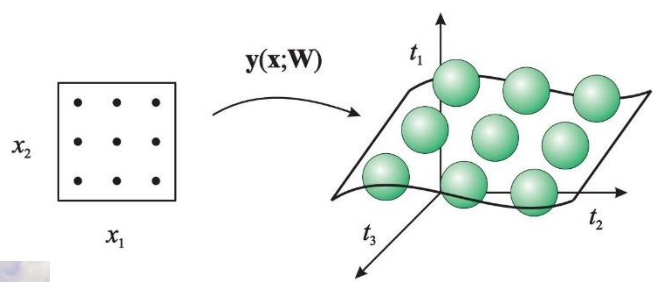
Boltzmann machine

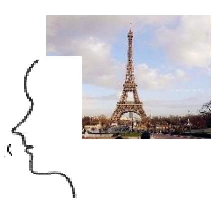


Restricted Boltzmann machine

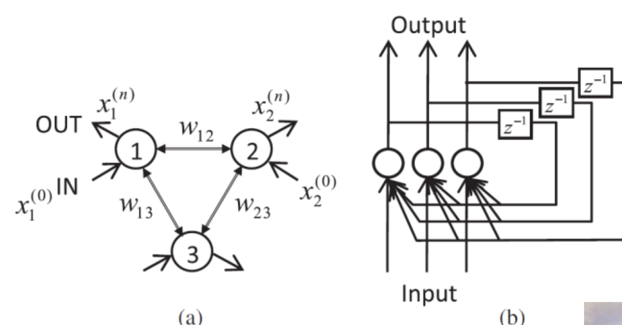


McCulloch–Pitts model

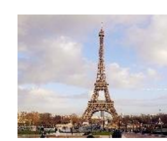




Generative topographic map



Hopfield network



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

Artificial Neural Network (ANN)

- Artificial neural network (ANN) is a **computational model** that **mimics** brain functionality with artificial means.
- Deep architecture was not successful at first, because
  - **Insufficient** labeled data.
  - **Limited** computation power.
- Learning from **unlabeled data** (generative models)
  - To make use of huge amount of unlabeled data.
- Followed by a **fine-tuning** to perform classification
  - Generative model serves as a good initial point.
- Deep models achieve current **state-of-the-art performances** in object recognition, speech recognition,…etc.

Deep Neural Network (DNN)

- DNN: **layered neural nets** with many hidden layers.
- Adding extra layers **increases representational power** of the overall model.
- DNNs were somewhat disappointing **20 years ago**, because
  - Labeled data was insufficient.
  - Computation power was limited.
  - The problem of random initials and vanishing gradients.
  - Support vector machine (SVM) has been proposed.

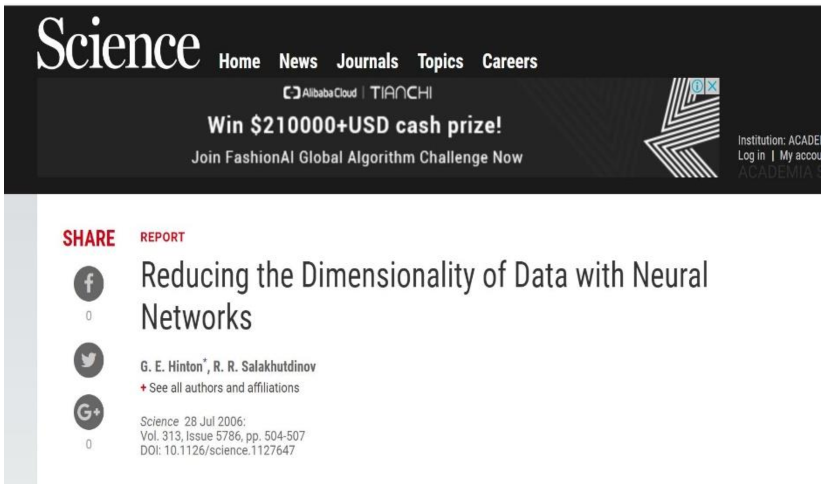
Pre-training for DNN

- With advanced computation power, current issue of DNN:
  - The problem of **random initials** and **vanishing gradients**.
- How can we overcome the issue?
  - By '**pre-training**' the networks, such RBMs, auto-encoder.
- Why pre-training ?
  - Utilizing large amount of unlabeled-data effectively.
  - Providing a good **initial point**. ?
- Procedure and assumptions of pre-training
  - Learn **one layer** at a time and **stack** them up.
  - **Shallow models** are easier to train, and can be used for initialization of deep models.

Fine-tuning on DNN Parameters based on Back-propagation

- After pre-training, back-propagation is performed to fine-tune the model parameters
  - A **big difference** to original approach of initializing with **random weights** then back-propagation.
  - Because now we already have a **sensible initialization points** before performing back-propagation.
  - This is the **key difference** between deep learning and traditional neural network 20 years ago.
  - **Deep belief network (DBN)** is generally used for pre-training, and **restricted Boltzmann machine (RBM)** is the function box of DBN.

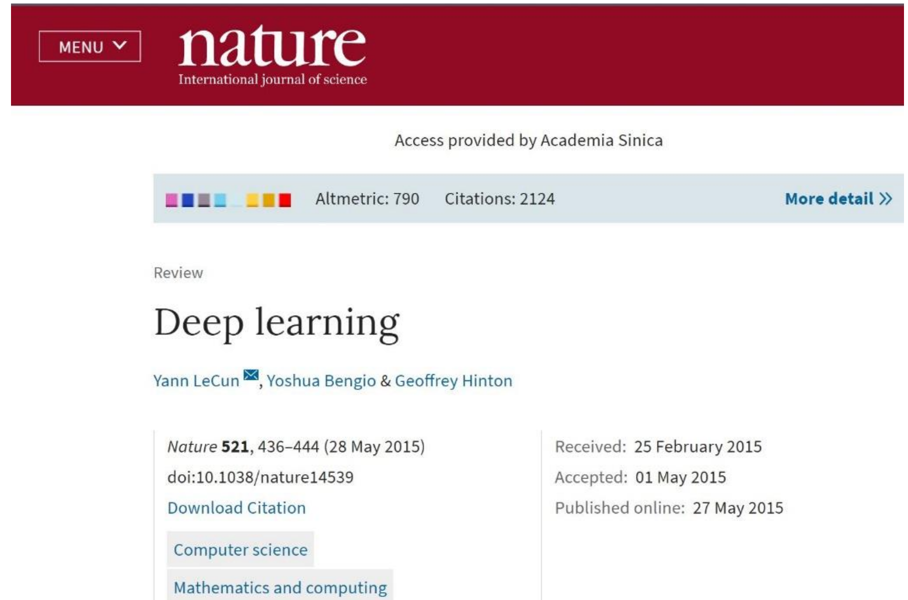
Recent Advances in Deep Learning


Science Home News Journals Topics Careers

Win $210000+USD cash prize!
Join FashionAI Global Algorithm Challenge Now

SHARE
REPORT
Reducing the Dimensionality of Data with Neural Networks
G. E. Hinton*, R. R. Salakhutdinov
+ See all authors and affiliations
Science 28 Jul 2006:
Vol. 313, Issue 5786, pp. 504-507
DOI: 10.1126/science.1127647

10601 citations


MENU
nature
International journal of science

Access provided by Academia Sinica

Altmetric: 790 Citations: 2124 More detail

Review
Deep learning
Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Nature **521**, 436–444 (28 May 2015)
doi:10.1038/nature14539
Download Citation
Computer science
Mathematics and computing

Received: 25 February 2015
Accepted: 01 May 2015
Published online: 27 May 2015

20227 citations

Turing Award Winners

<1>

Recent Advances in DL and AI

MENU

nature
International journal of science

Altmetric: 3197 Citations: 569 **More detail »**

Article

Mastering the game of Go with deep neural networks and tree search

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis

Nature **529**, 484–489 (28 January 2016)
doi:10.1038/nature16961
Download Citation

Computational science Computer science Reward

Received: 11 November 2015
Accepted: 05 January 2016
Published online: 27 January 2016

<2>

MENU

nature
International journal of science

Altmetric: 2152 Citations: 1 **More detail »**

Article

Mastering the game of Go without human knowledge

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

Nature **550**, 354–359 (19 October 2017)
doi:10.1038/nature24270
Download Citation

Computational science Computer science Reward

Received: 07 April 2017
Accepted: 13 September 2017
Published online: 18 October 2017

<3>

Recent Advances in DL and AI

nature
International journal of science

MENU

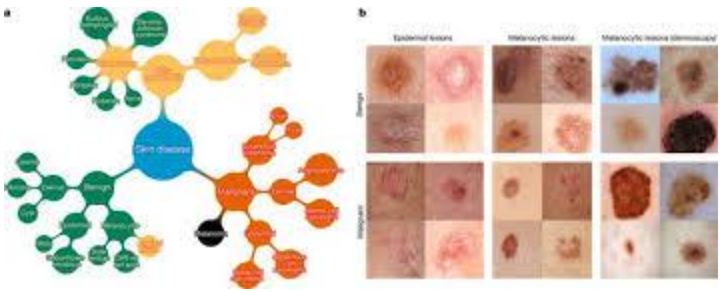
Letter

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017)
doi:10.1038/nature21056
Download Citation
Diagnosis Machine learning Skin cancer

Received: 28 June 2016
Accepted: 14 December 2016
Published online: 25 January 2017
Corrigendum: 28 June 2017



<4>

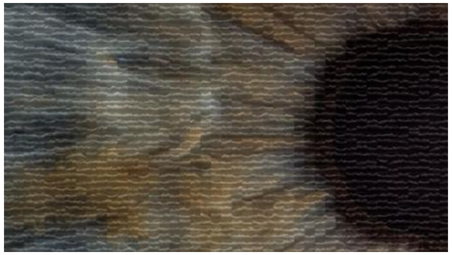
nature
International journal of science

This illustration of an iris and pupil is composed of thousands of diagnostic scans of the eye. Credit: Daniel Kermany, Guangzhou Medical University and Kang Zhang, UC San Diego Health

TECHNOLOGY · 22 February 2018

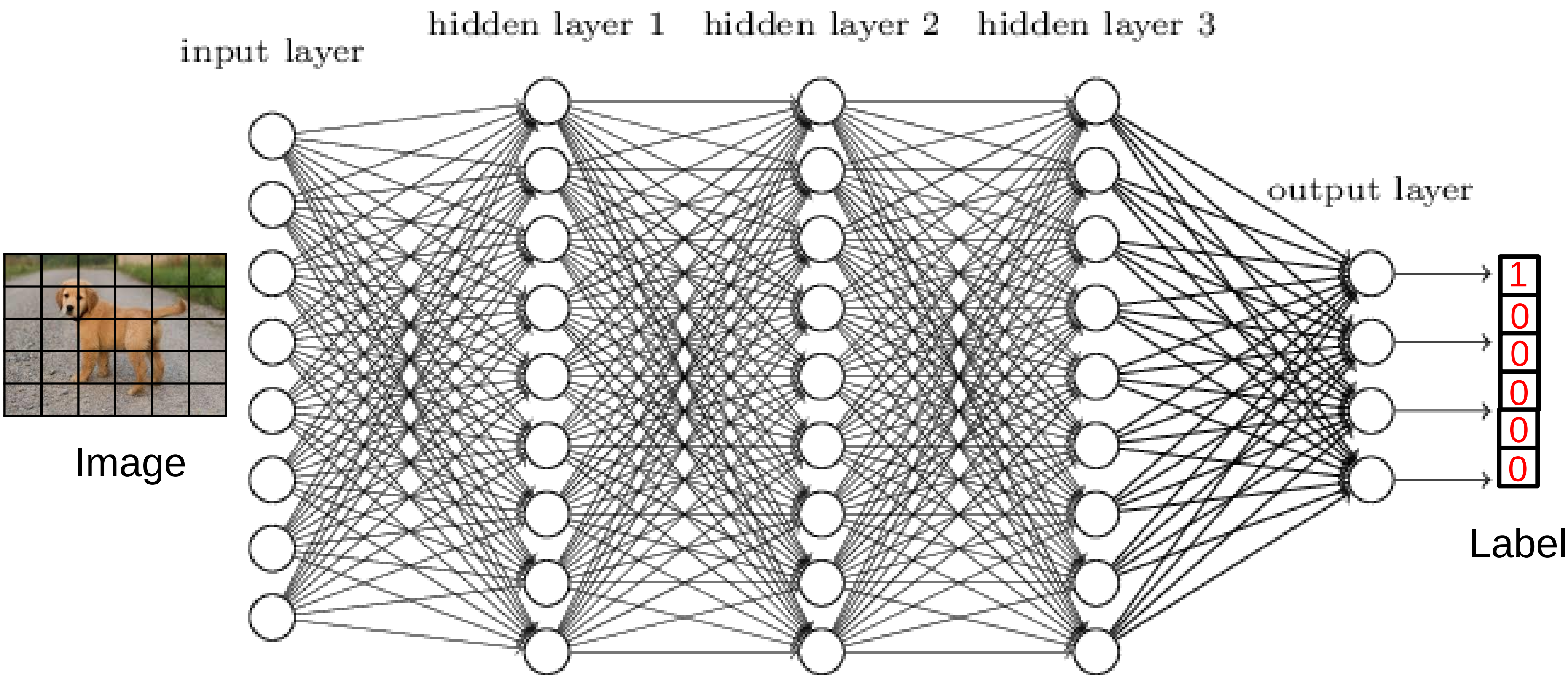
An efficient deep-learning tool for detecting eye disease

Model scans images to detect urgent signs of conditions leading to blindness.



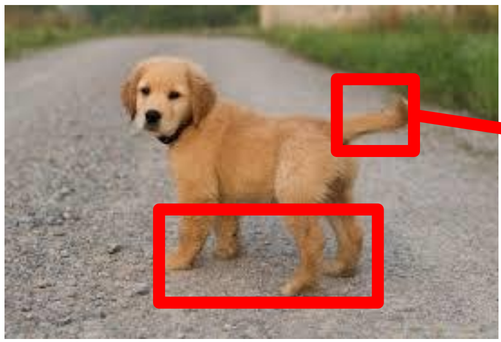
<5>

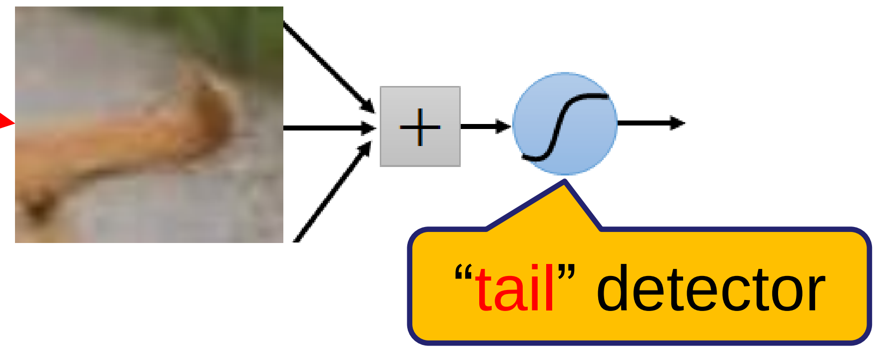
Fully Connected NNs



Different from human perception

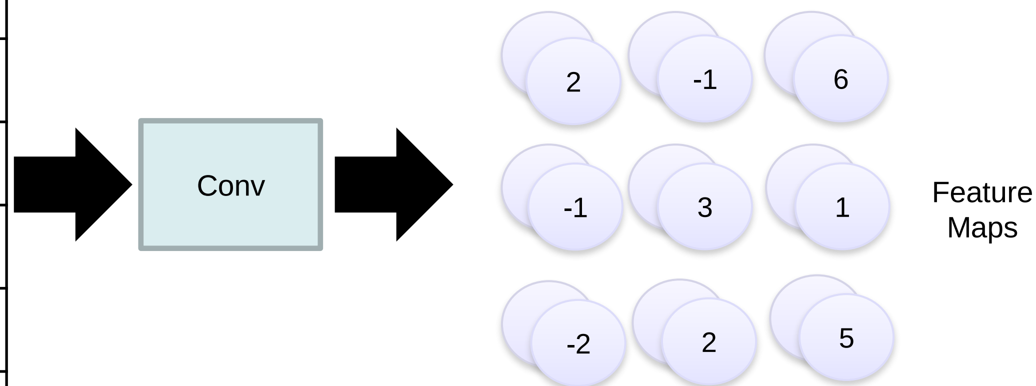
Convolutional NNs





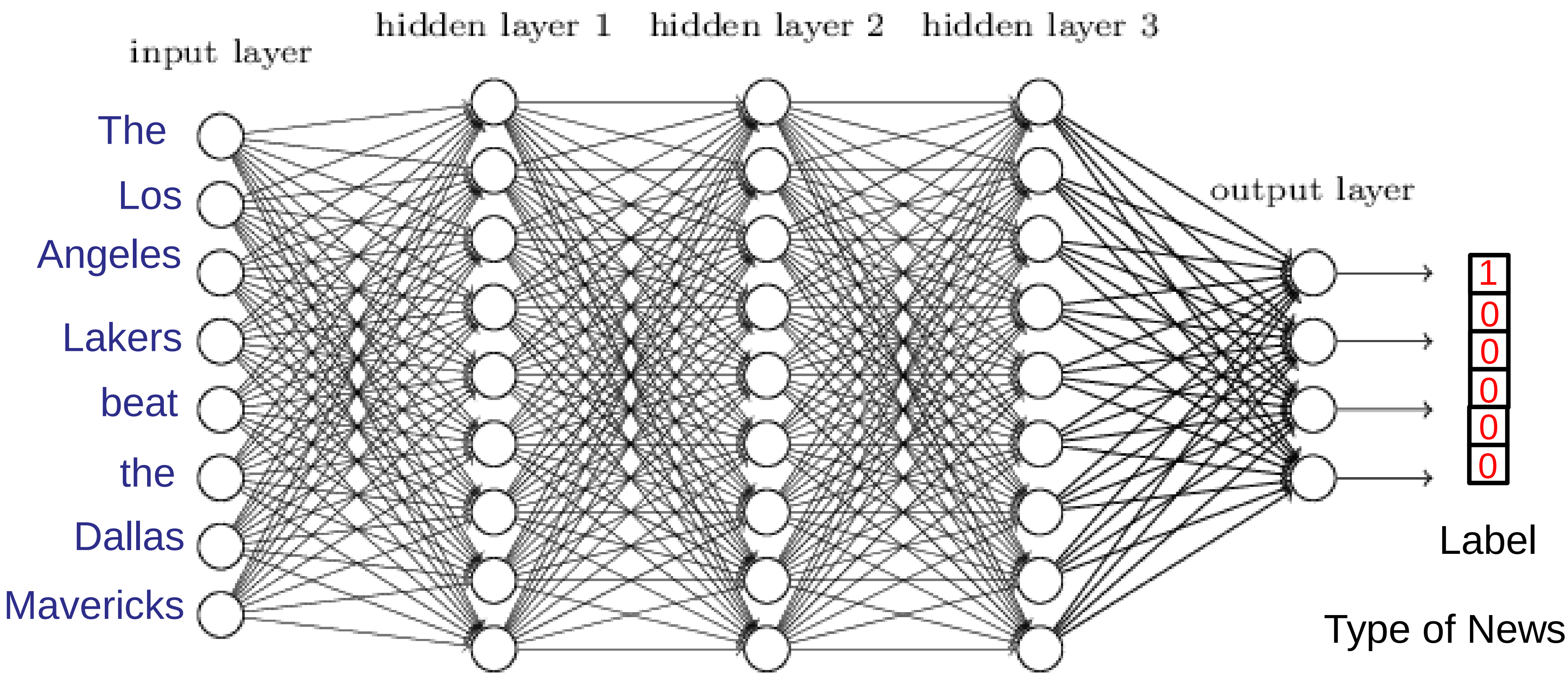
0	1	1	0	1	0
1	1	1	0	0	1
1	0	0	1	0	0
1	0	1	0	0	1
0	0	0	1	0	0
0	1	1	0	0	1

6 x 6 image



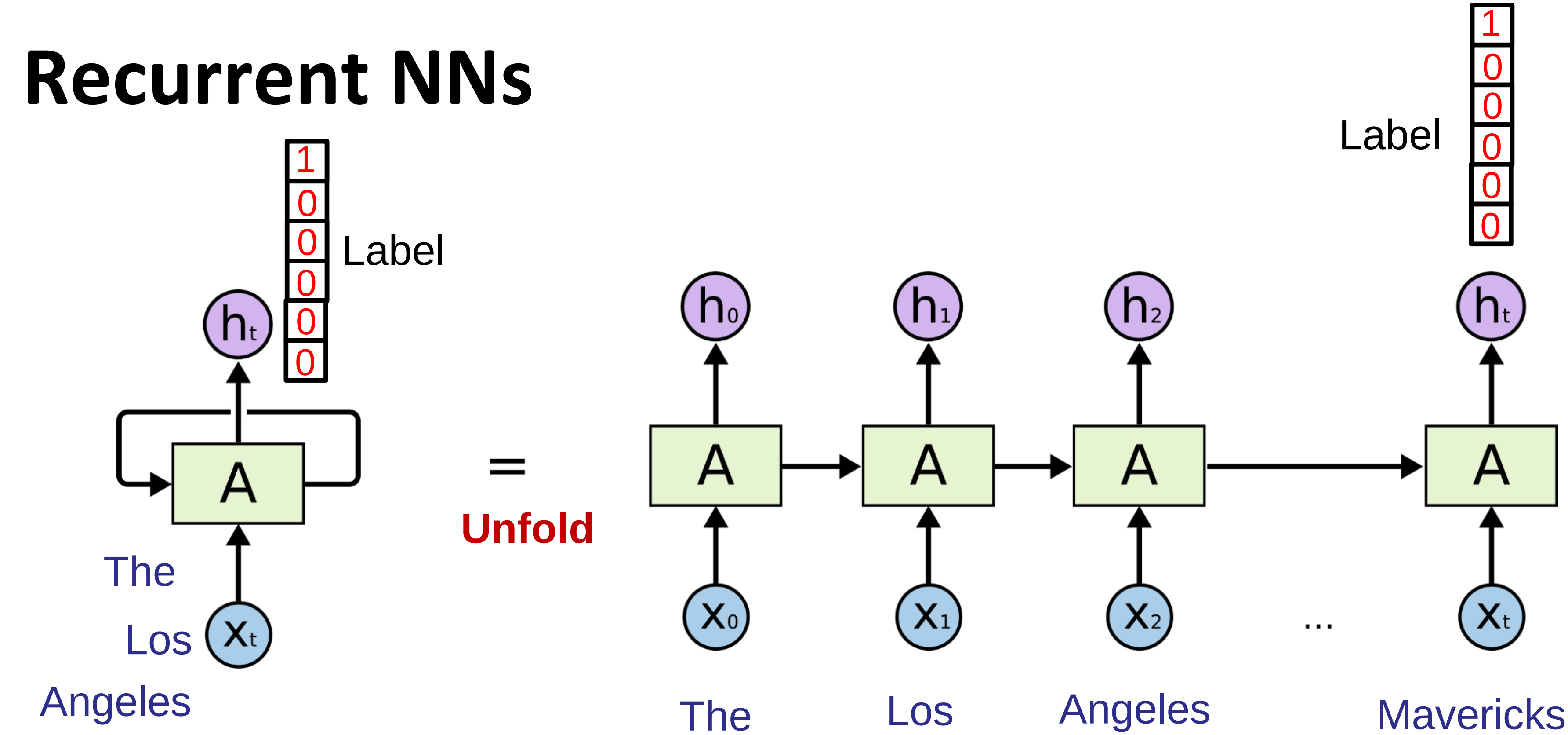
- Reducing number of connections
- Shared weights on the edges
- Max pooling further reduces the complexity

Fully Connected NNs



Waste time and computations

Recurrent NNs



Lakers

beat

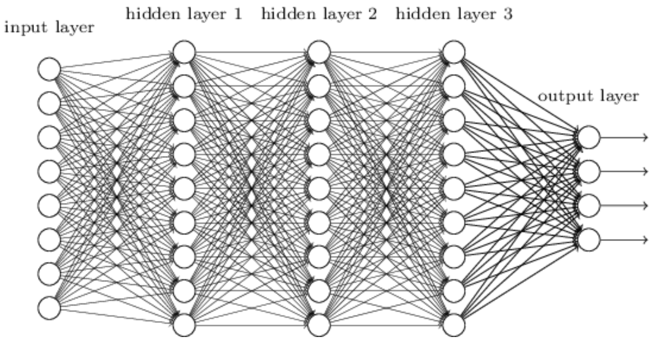
the

Dallas

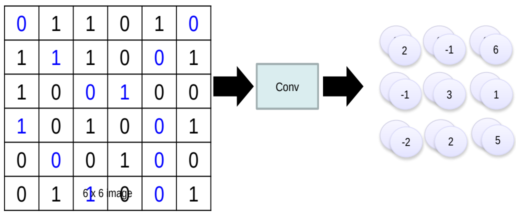
Mavericks

- Considering bi-direction of the signal→ Bidirectional RNN
- Considering history information→ Long short-term memory (LSTM)
- Simplified LSTM→ Gated recurrent unit

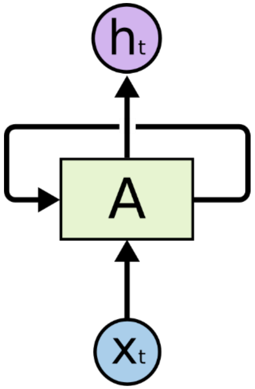
The NN Family



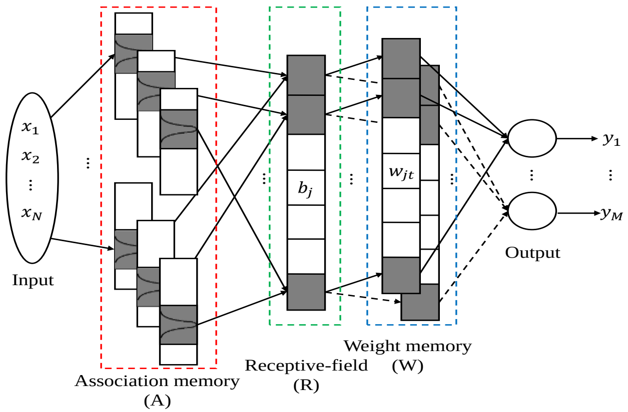
- Fully connected

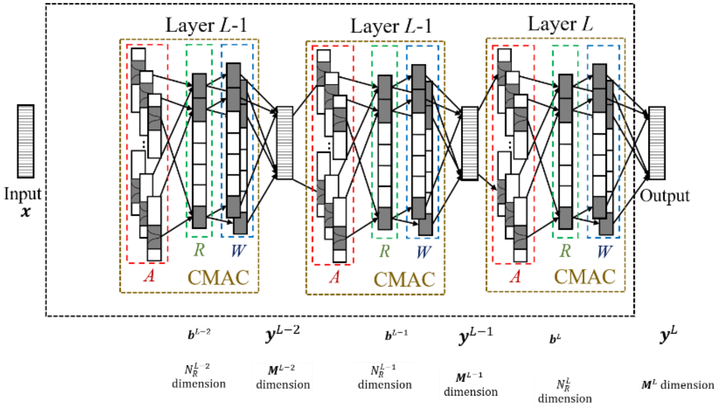


- CNN



- RNN
- Bidirectional RNN
- Long short-term memory
- Gated recurrent unit



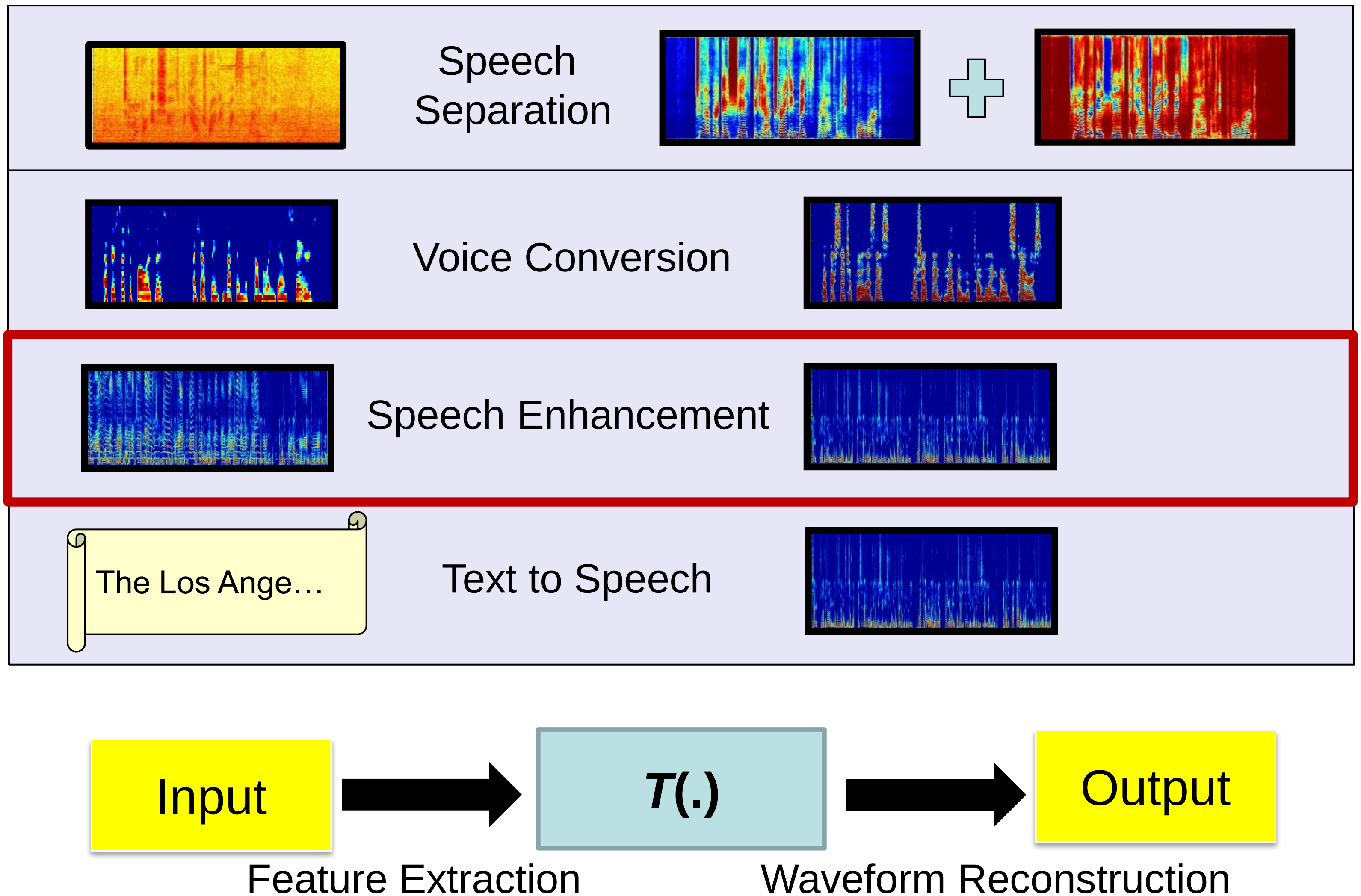


- Cerebellar Model Articulation Controller (CMAC) and deep CMAC

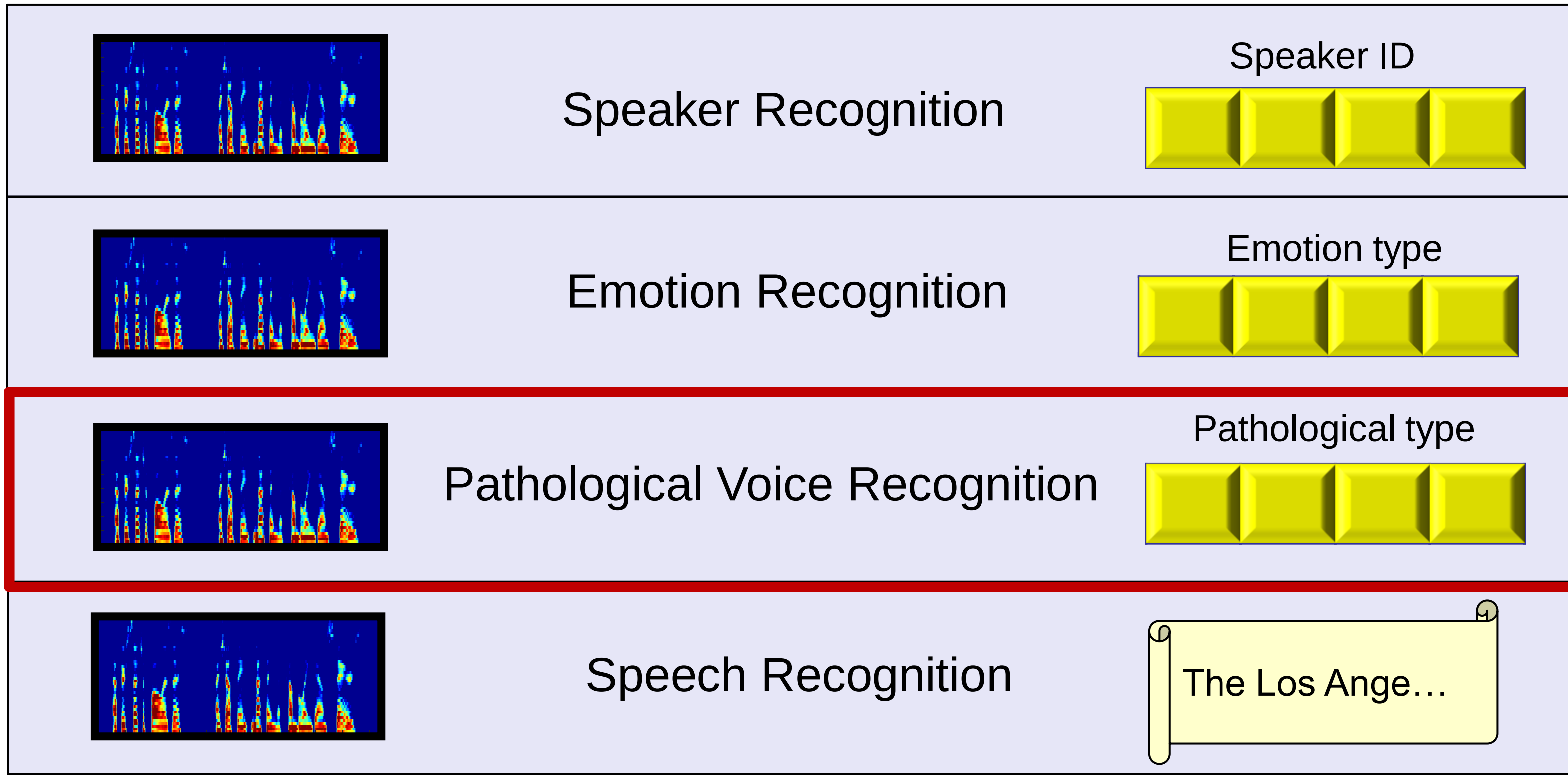
Outline

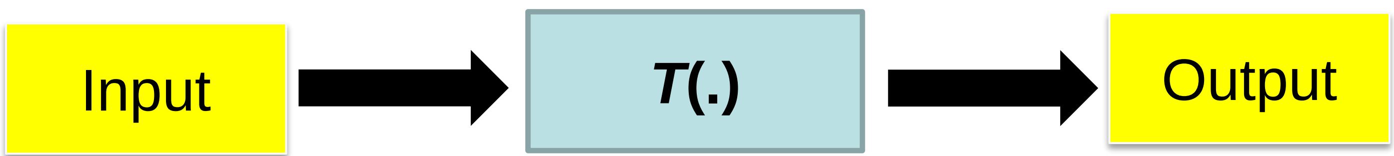
- Deep Learning
  - Artificial intelligence, machine learning, deep learning
  - Human learning versus machine learning
  - Some histories about deep learning
  - Popular deep learning models
- Speech Signal Processing
  - Two categories of tasks: recognition and generation
  - Recognition: pathological voice recognition
  - Generation: speech enhancement

Speech Generation (Regression Task)



Speech Signal Recognition (Classification Task)

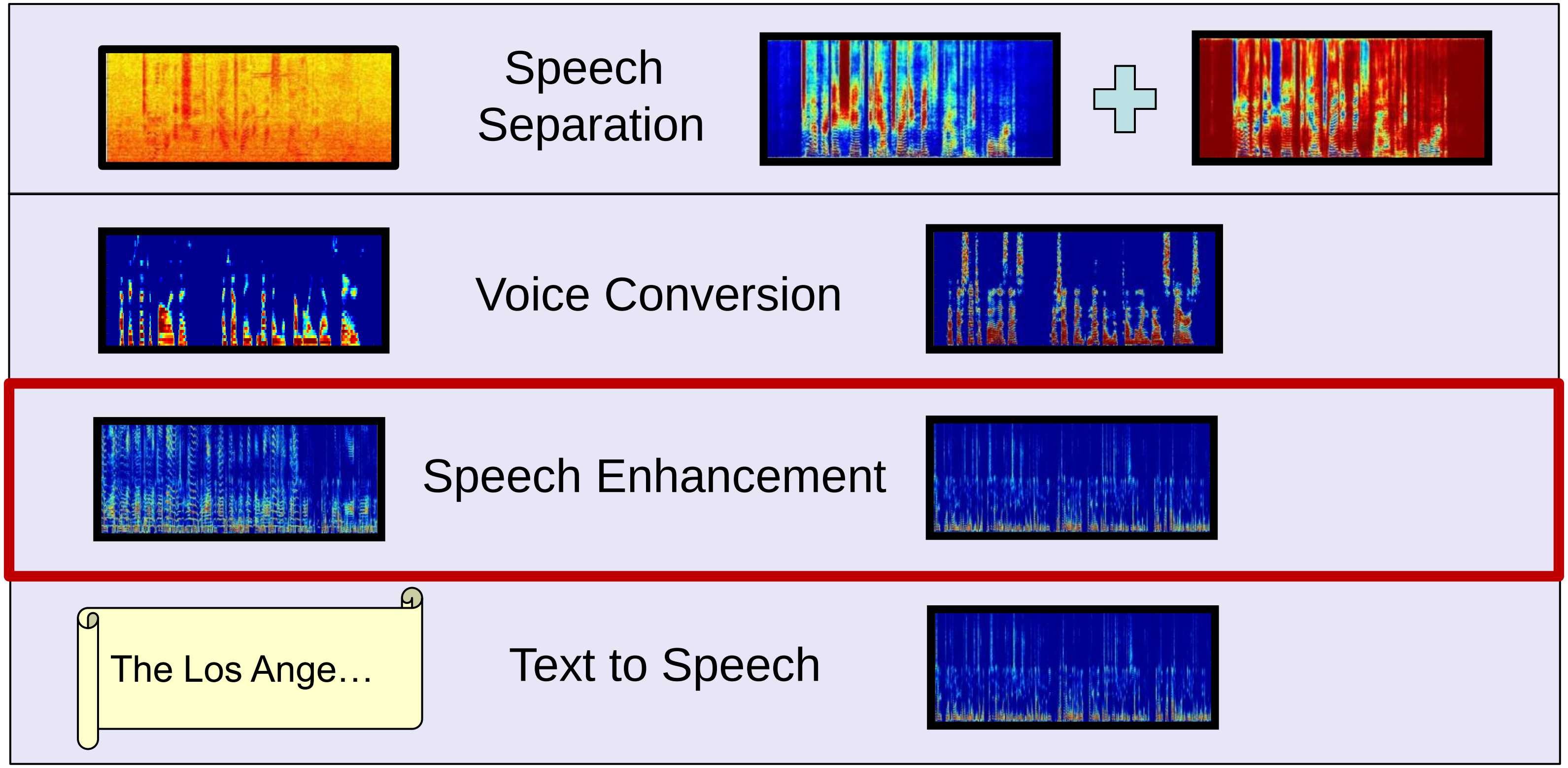


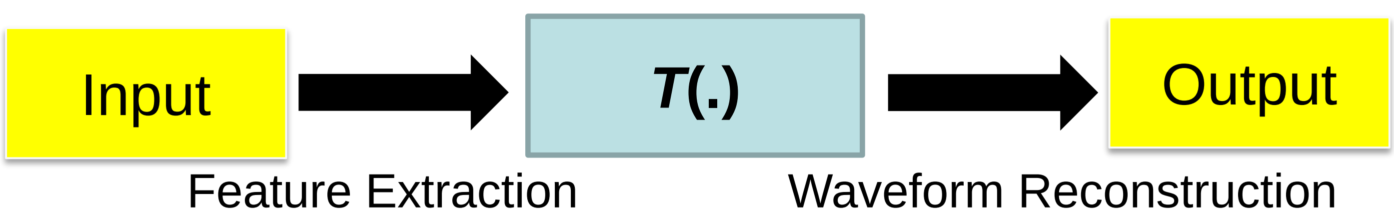


Feature Extraction

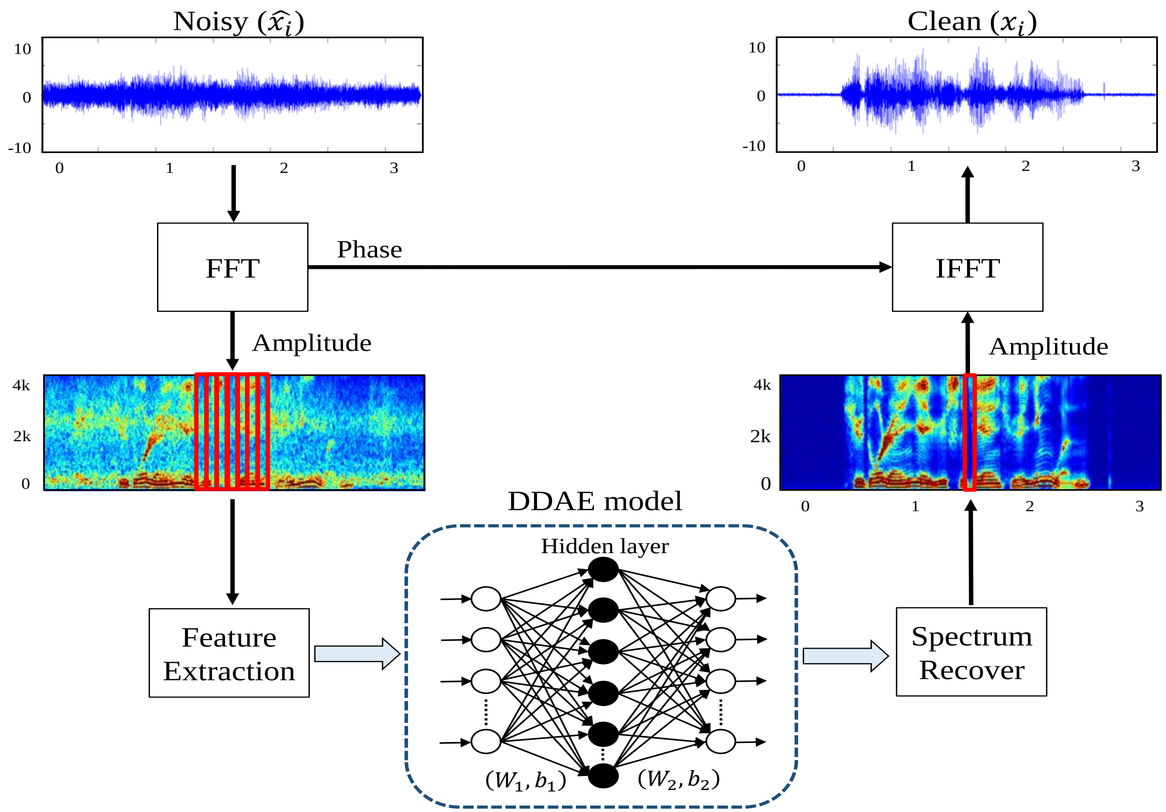
MFCC, Fbank, I-vector, Prosodic Features, and Raw-data

Speech Generation (Regression Task)





Deep Learning based Speech Enhancement

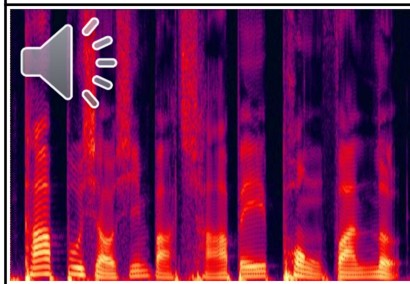
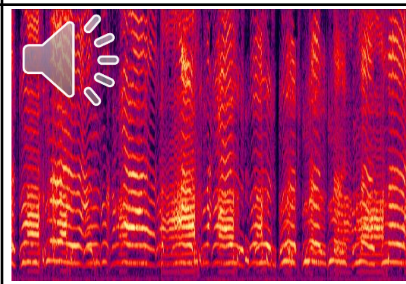
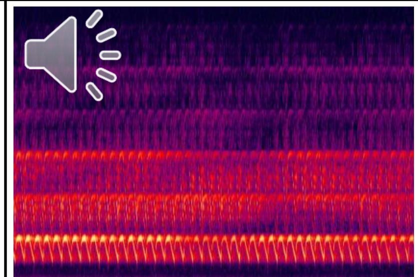
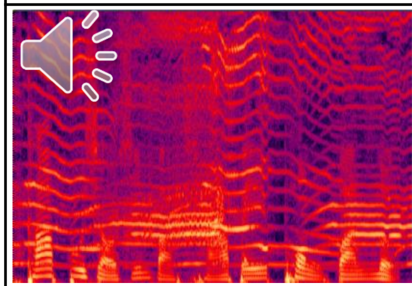
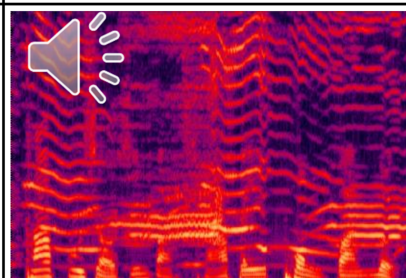
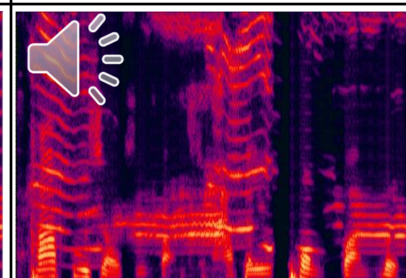
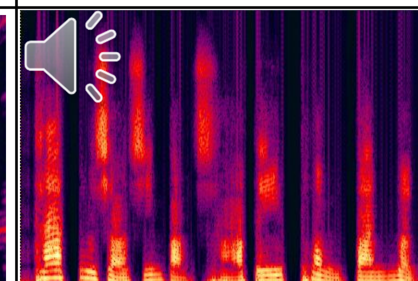
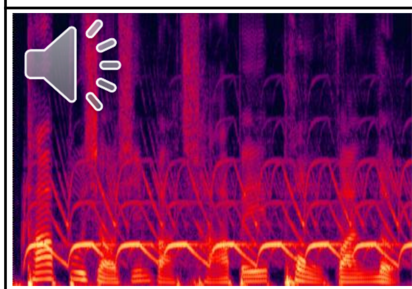
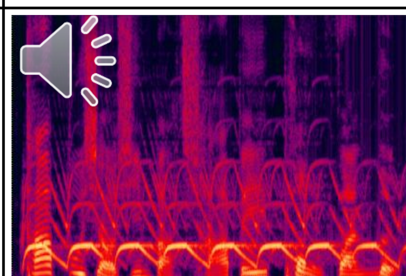
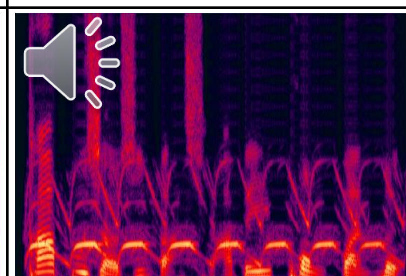
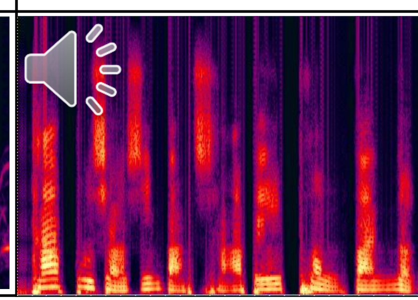


$$\theta^* = \arg\min_{\theta}\left(F(\theta) + \eta^1\|\boldsymbol{W}^1\|_{\mathrm{F}}^2 + \cdots + \eta^L\|\boldsymbol{W}^L\|_{\mathrm{F}}^2\right),$$

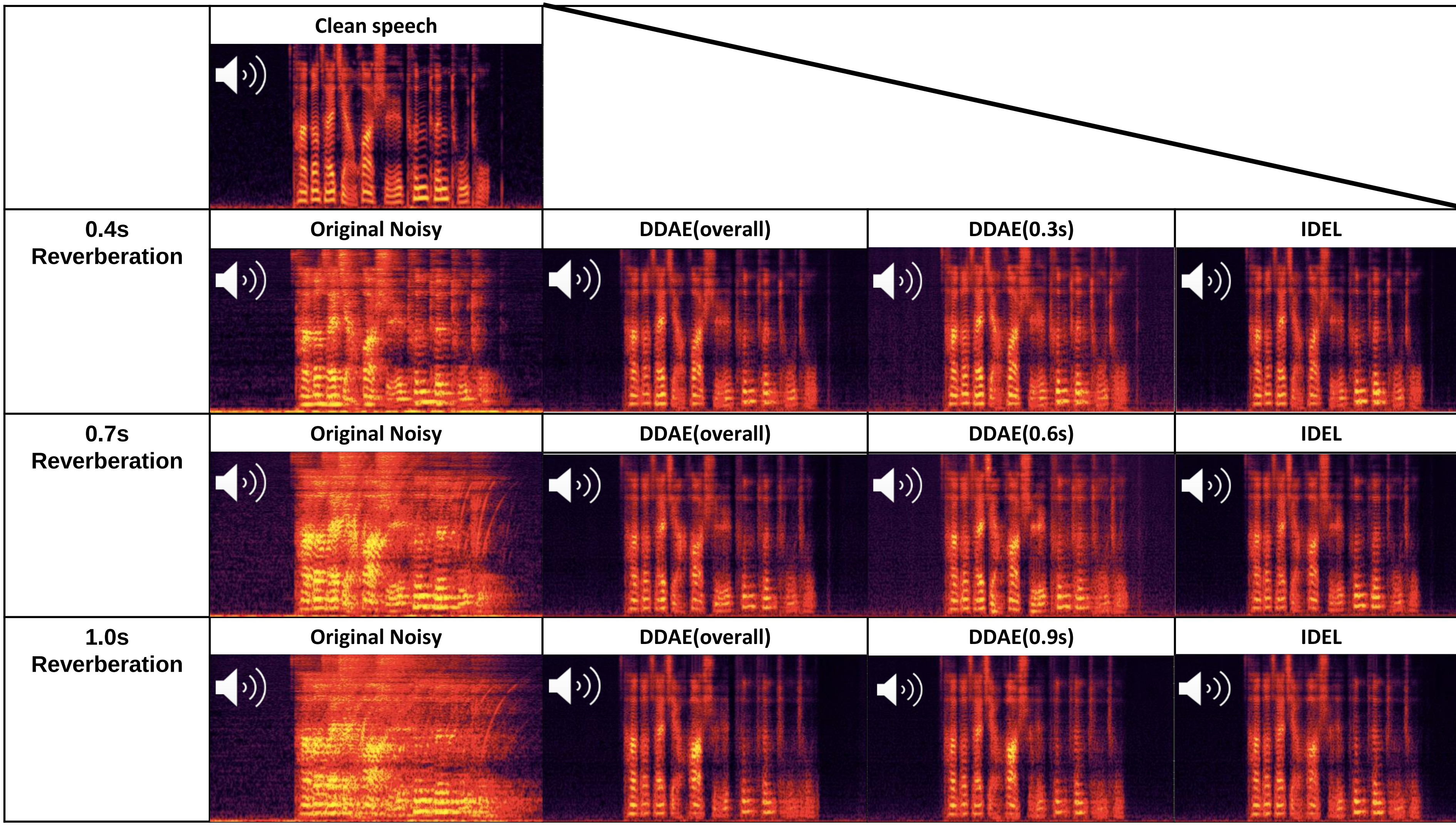
$$F(\theta) = \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{X}_m^{LPS} - \widehat{\boldsymbol{X}}_m^{LPS}\right\|_2^2,$$

- X. Lu, Y. Tsao, S. Matsuda and C. Hori, “Speech Enhancement based on Deep Denoising Autoencoder,” Interspeech 2013.

DL for Denoising

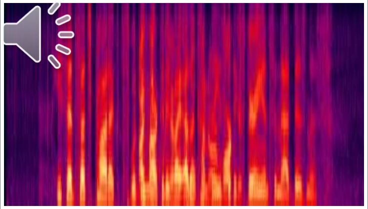
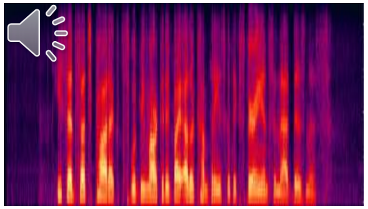
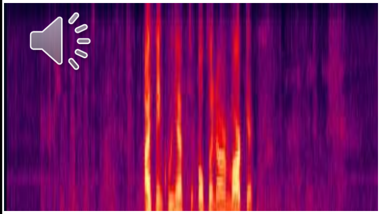
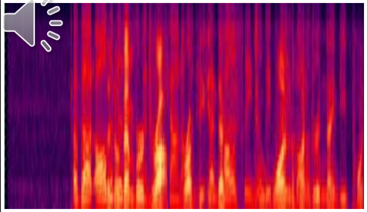
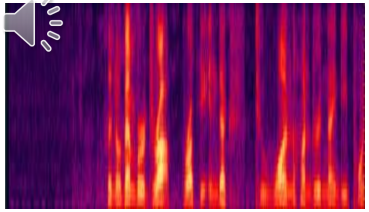
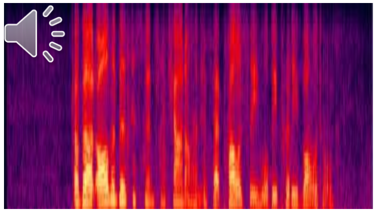
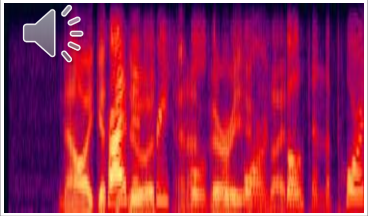
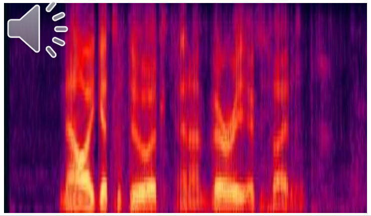
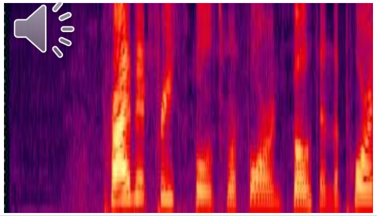
	Clean speech	Noise: 2baby Crying	Noise: Siren	
				
2baby Crying	Original Noisy	MMSE (Trandtional-1)	KLT (Trandtional-2)	DDAE
				
Siren	Original Noisy	MMSE (Trandtional-1)	KLT (Trandtional-2)	DDAE
				

DL for De-reverbaration

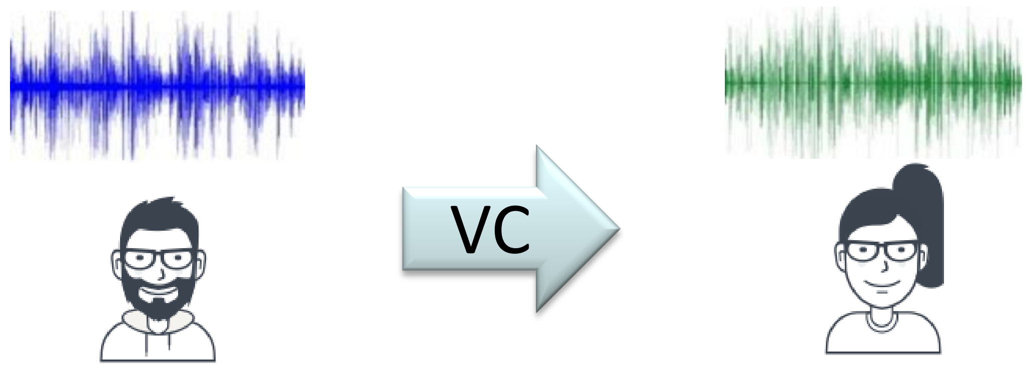


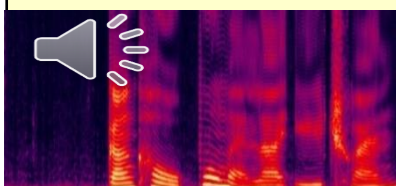
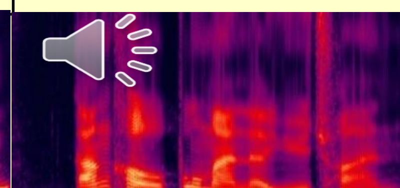
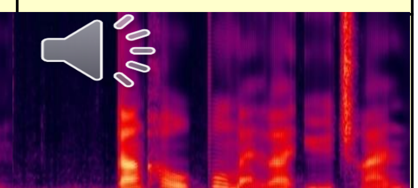
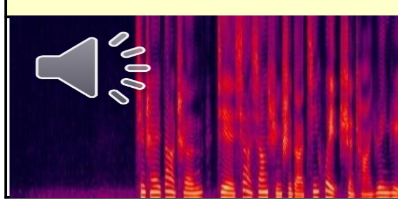
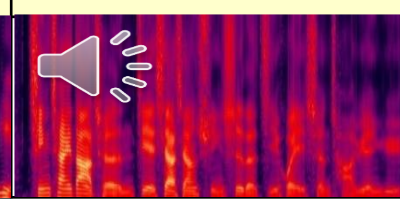
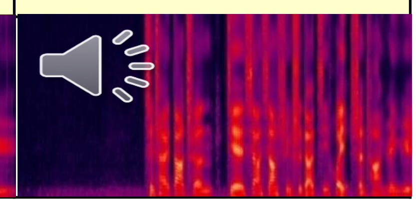
Source Separation



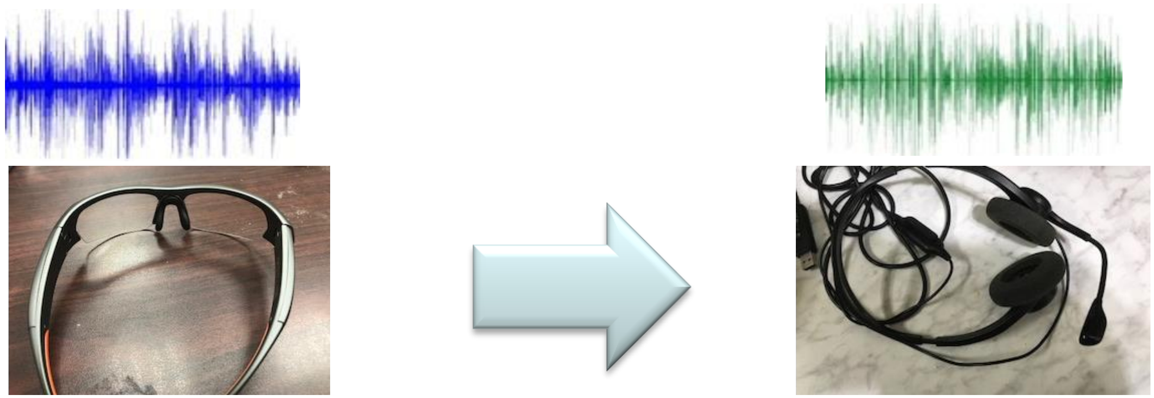
Sample	Mix	Speaker1	Speaker 2
1			
2			
3			

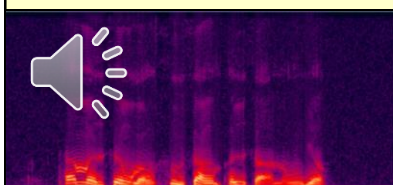
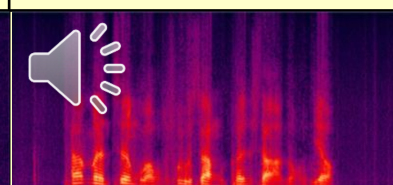
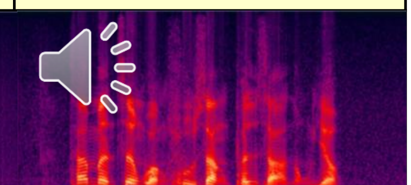
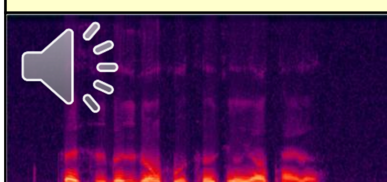
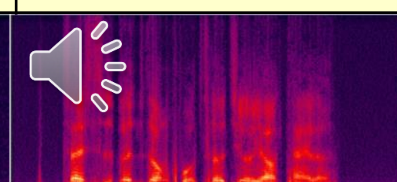
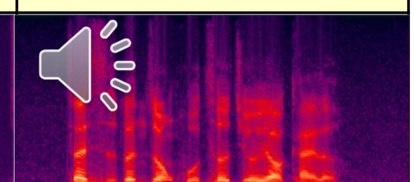
DL for Voice Conversion



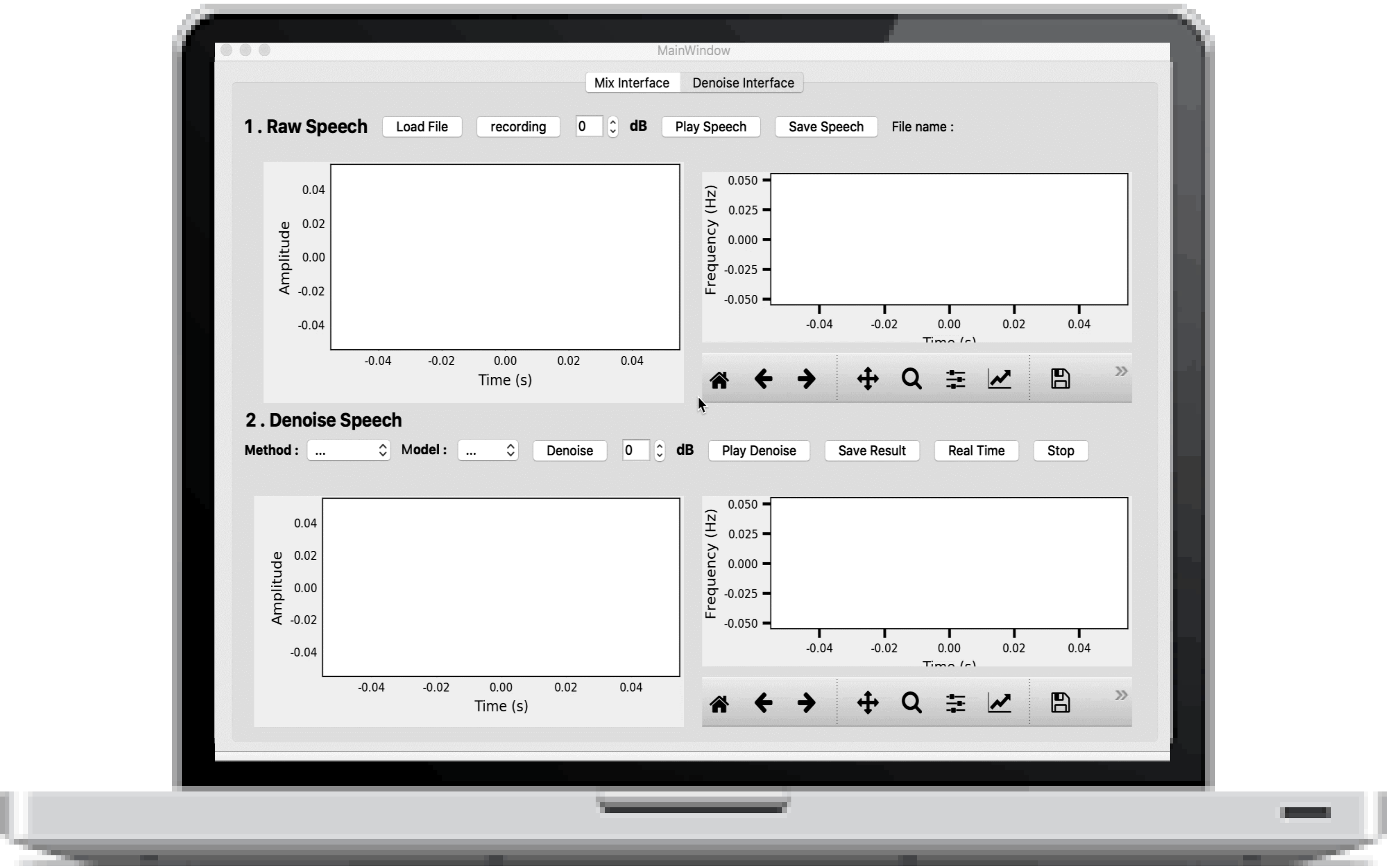
Source	Target	ANN
		
Source	**Target**	**ANN**
		

DL for Channel Compensation

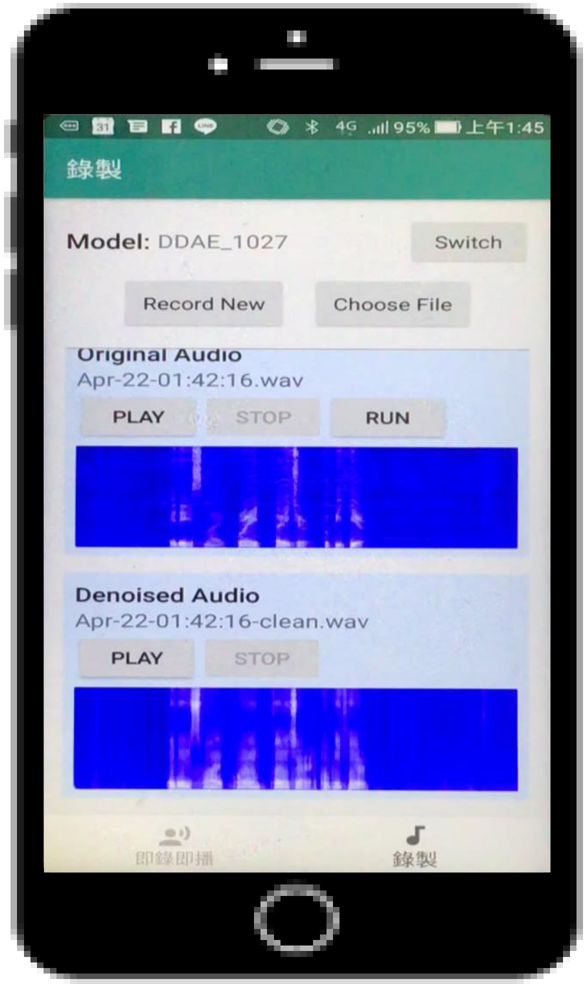


Bone 1	B_1->MSI	Ref MSI
		
Bone 2	**B_2->MSI**	**Ref MSI**
		

Audio Denoising System on PC



Audio Denoising System on Smartphone

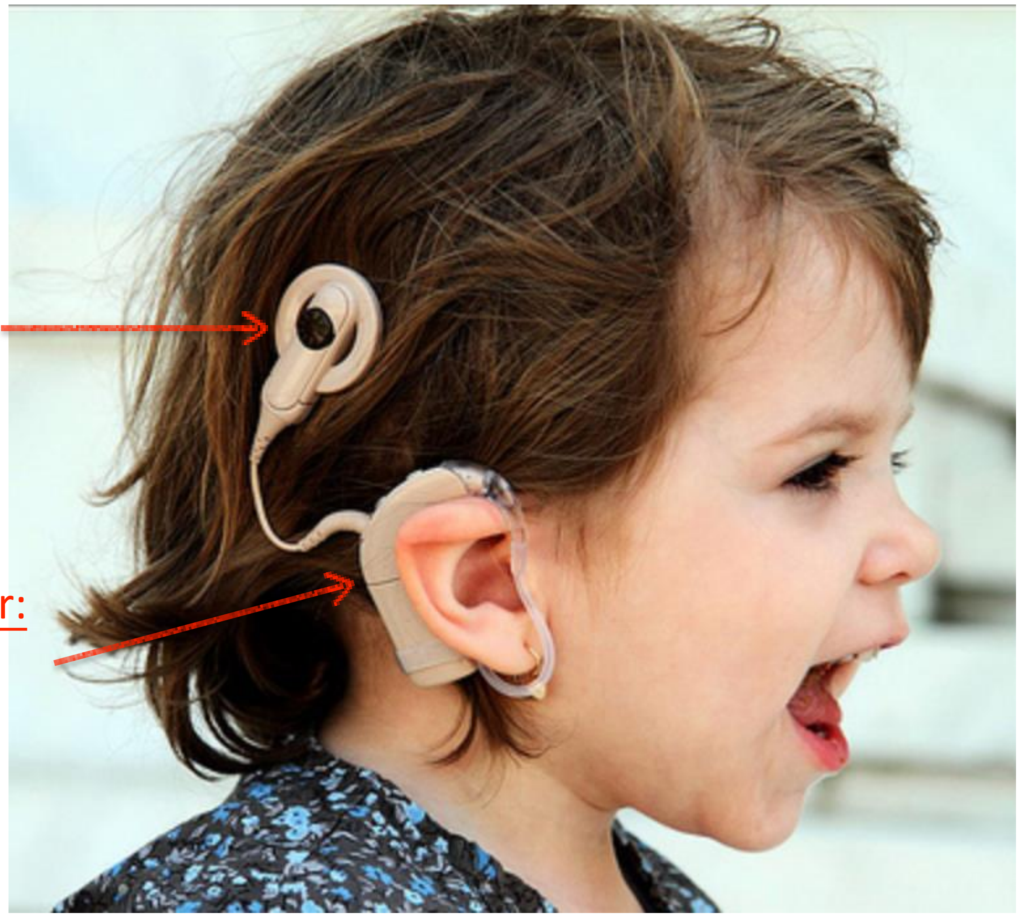


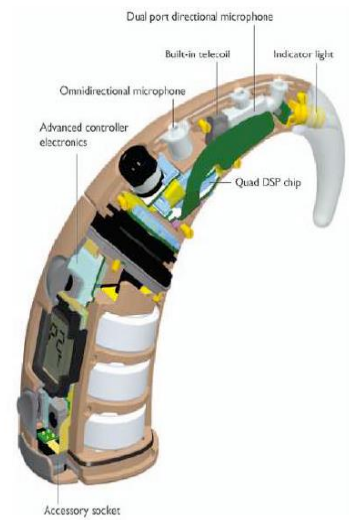
CI Device

Transmitter coil

<ins>Speech processor:</ins>

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





聞

聽

大學曰：心不在焉，聽而不聞

聞 (intelligibility) 比聽 (perception) 重要

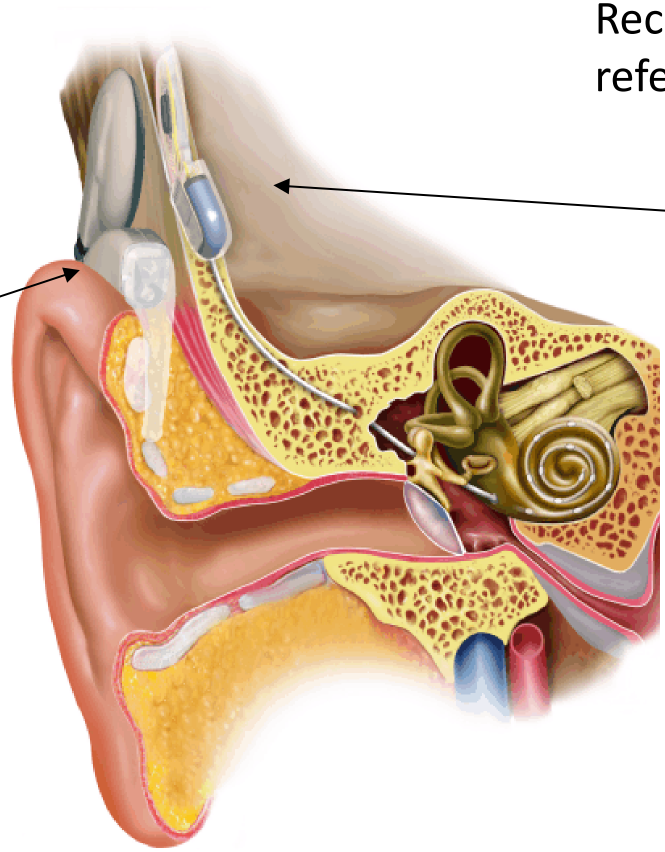
Cochlear Implant (CI)

- Surgically implanted device that **electrically stimulates** surviving auditory nerve fibers to provide sound for those with **severe hearing loss.**
- Over **200,000** users worldwide.
- FDA approved in 1985, now approved for children as young as **12 months.**

CI Device

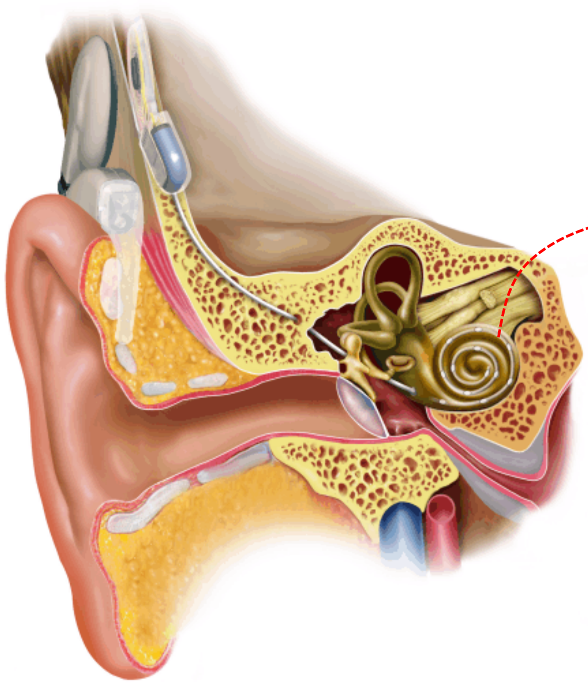
Emitter

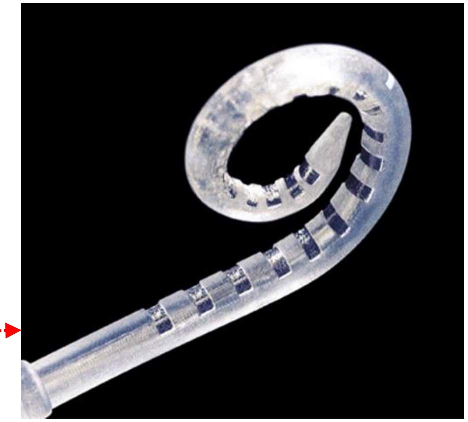
Speech Processor



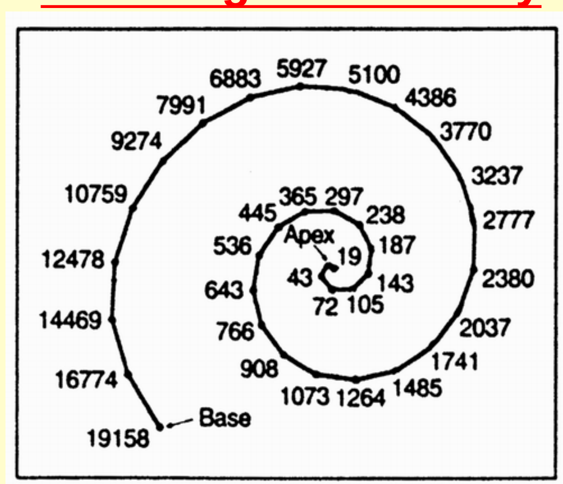
Receiver with stimulating reference electrode and

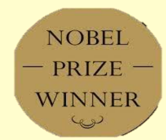
CI Device





Traveling wave theory

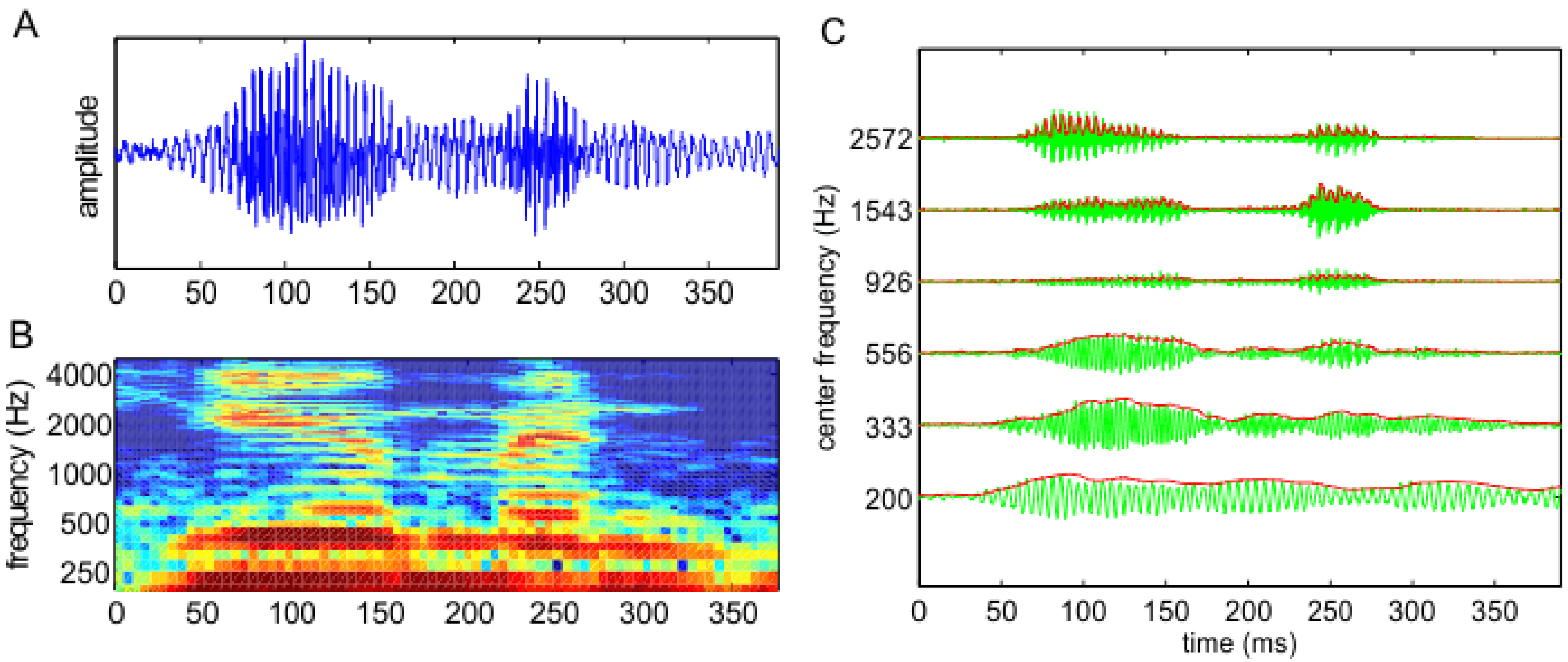




1961

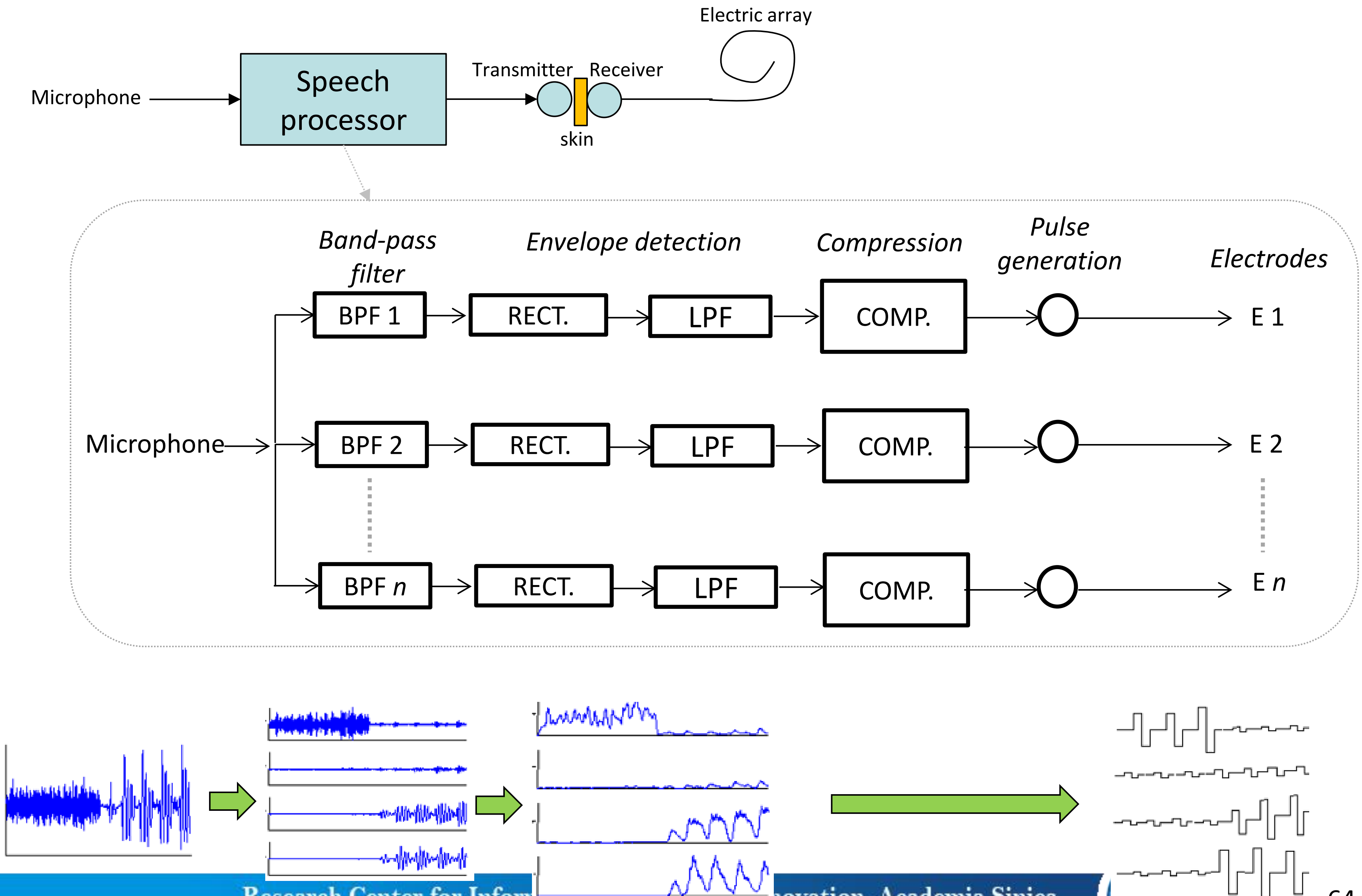
Von Békésy, Georg (1960). Experiments in hearing. Ed. Ernest Glen Wever. Vol. 8. New York: McGraw-Hill.

Bandpass and Envelope Extraction



(A) Waveform of the word "human" spoken by a native American speaker. (B) Spectrogram of the same word. (C) Green lines: Output of a set of six bandpass filters in response to the same word. The filter spacing and bandwidth in this example are two-thirds of an octave.

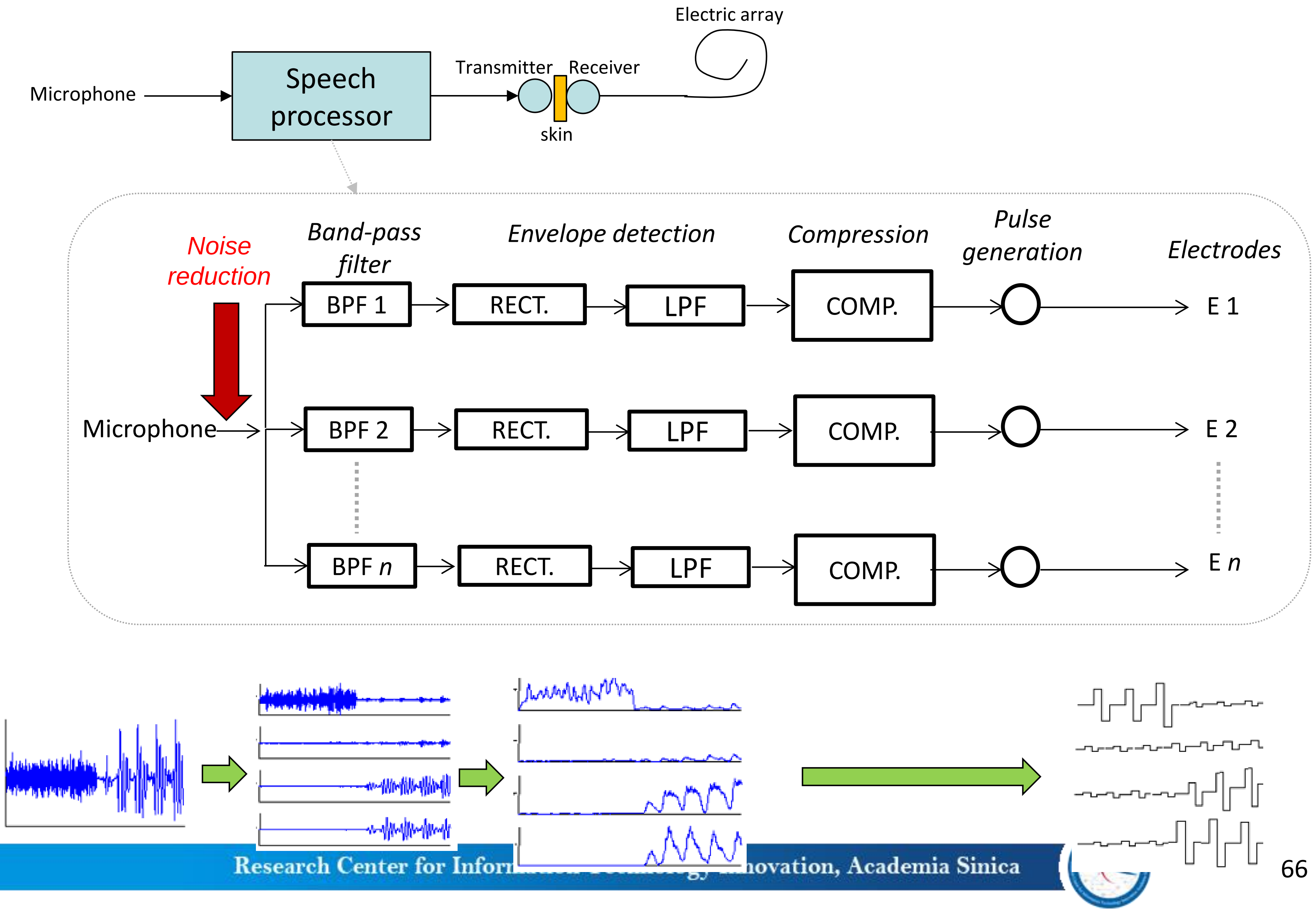
Signal Processing of CI



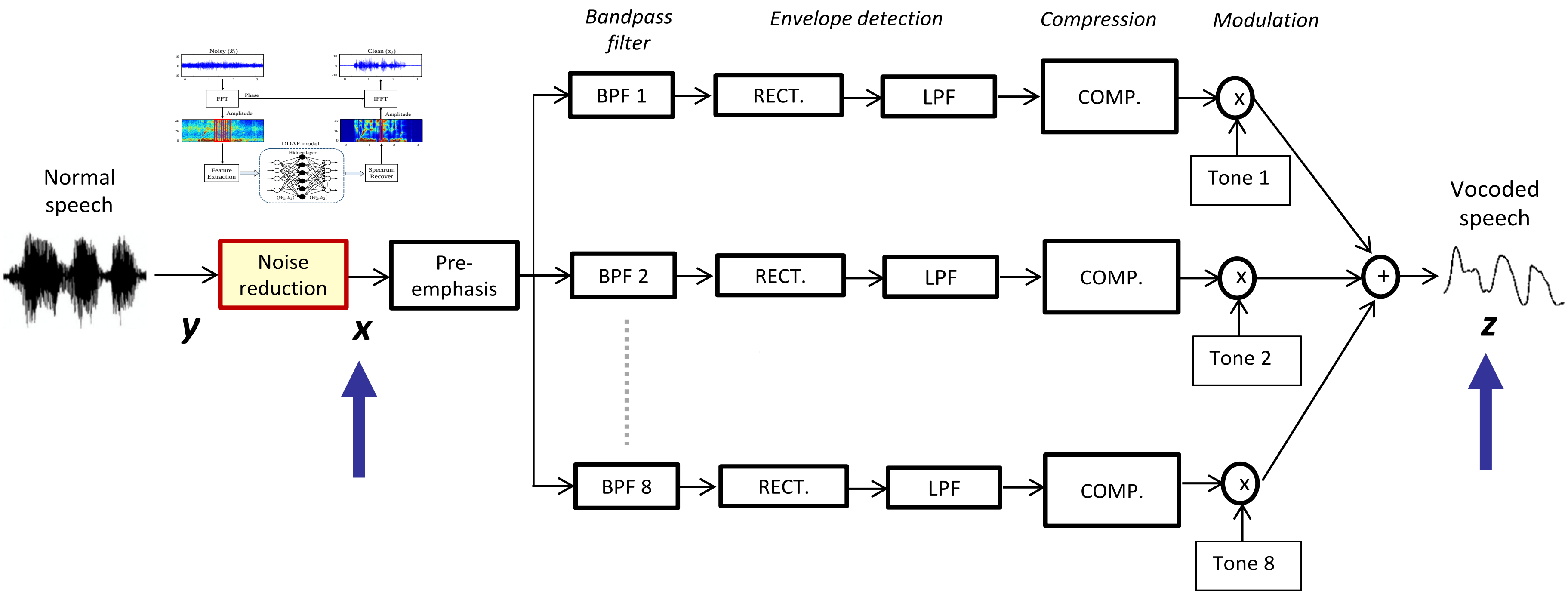
A Critical Issue of CI

- 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.

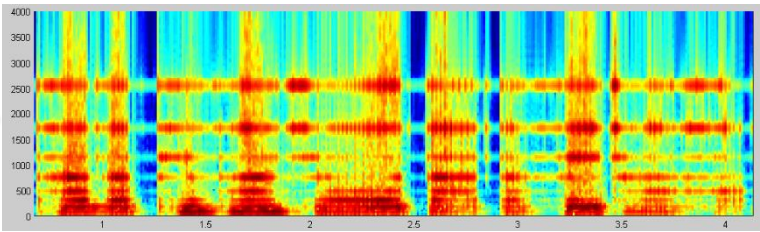
Signal Processing of CI



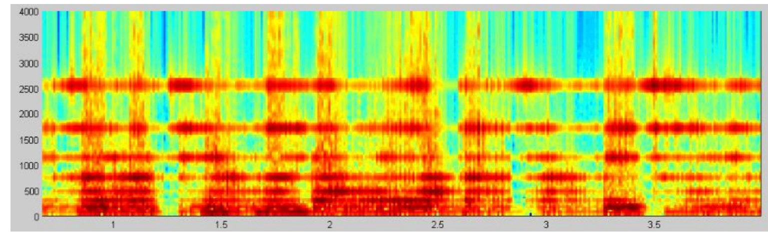
DL-based Noise Reduction on CI



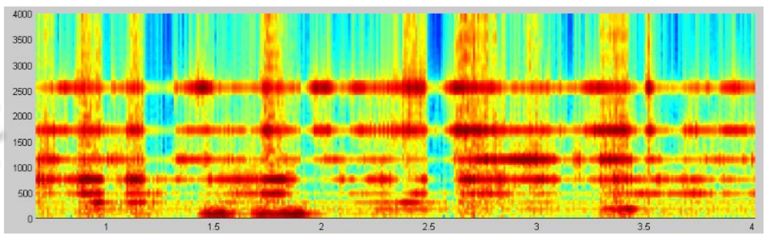
Vocoded Speech



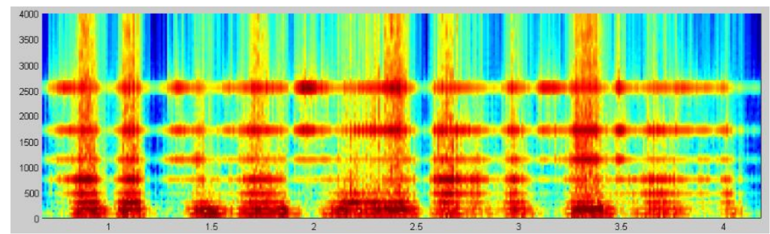
Clean



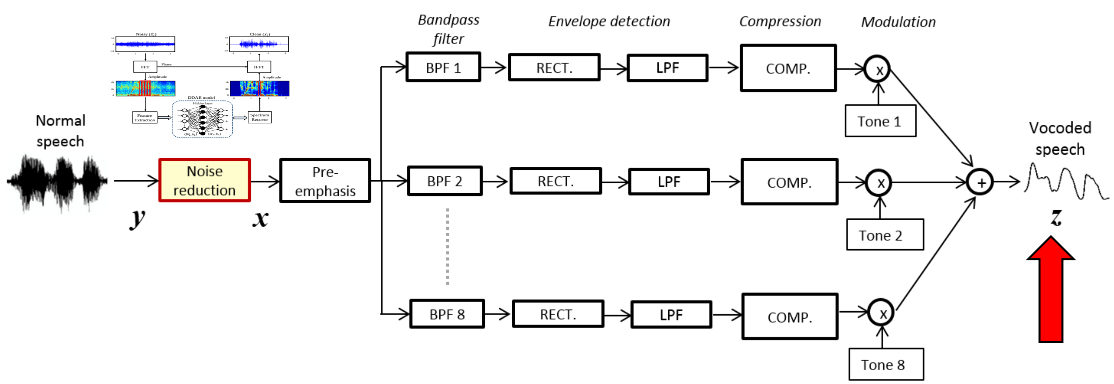
2T Noise 0dB



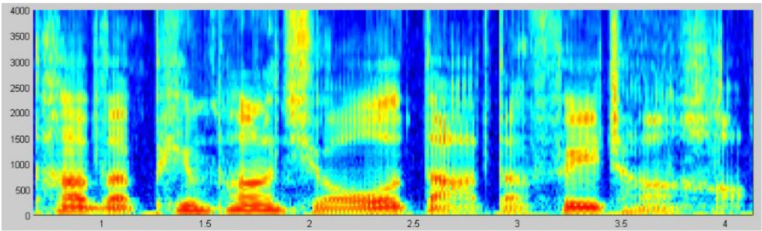
MMSE



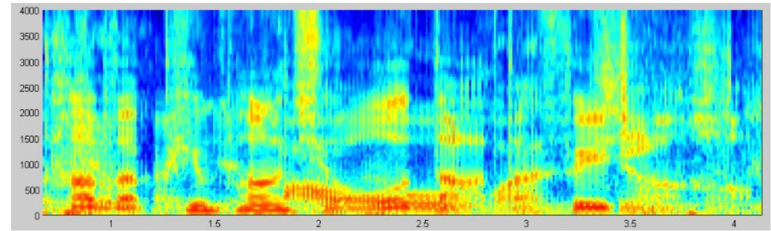
DDAE



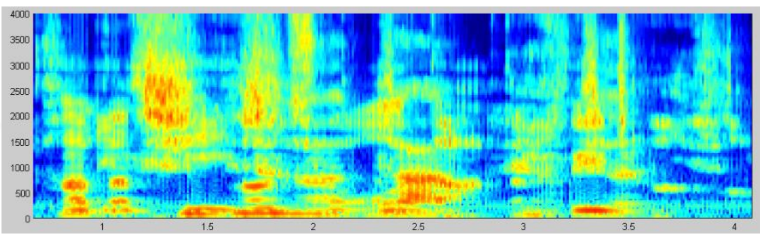
Normal Speech



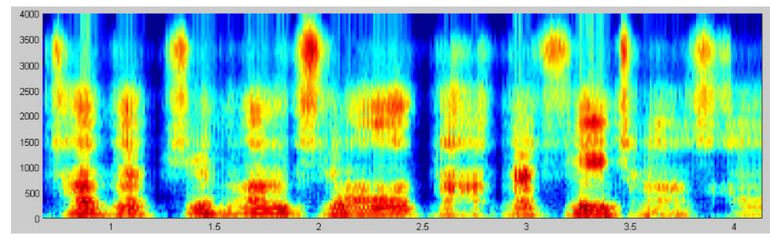
Clean



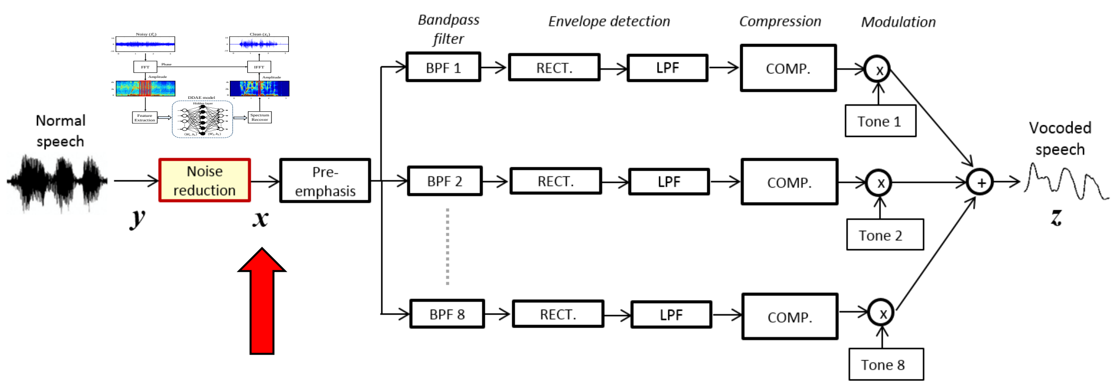
Babble Noise 0dB



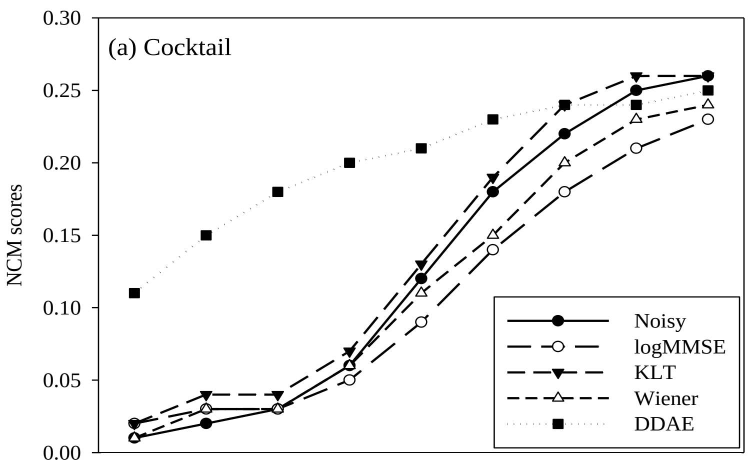
MMSE



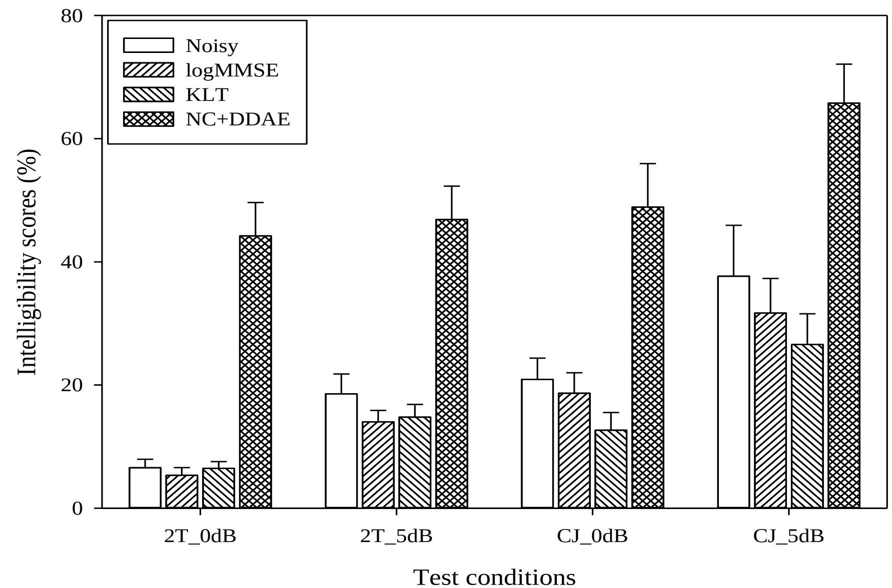
DDAE



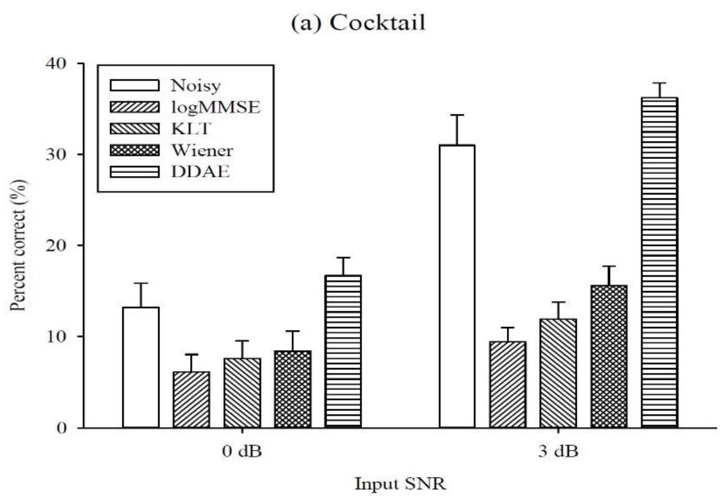
Evaluation Results (Simulations and Subject Tests)



Objective evaluation (NCM)



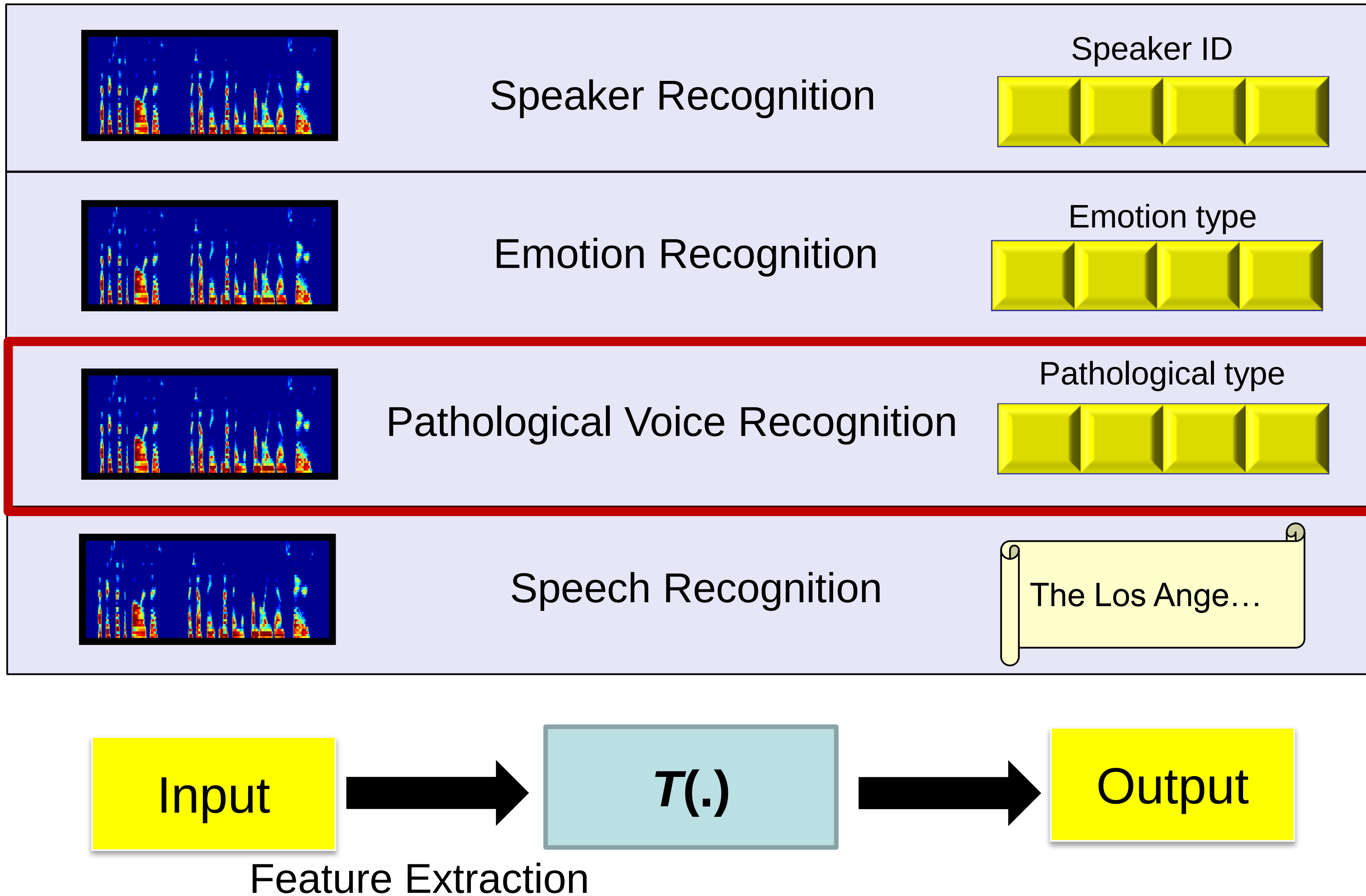
Clinical trial: 9 CI subjects.



Vocoder results: 10 normal hearing subjects.

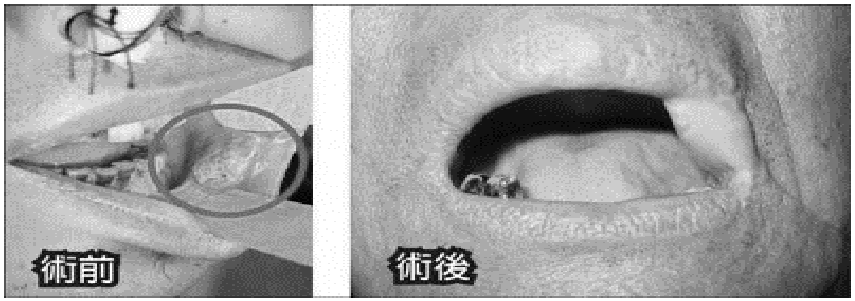
- Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation," IEEE Transactions on Biomedical Engineering.
- Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, K.-C. Chen, Y.-H. Chen,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

Speech Signal Recognition (Classification Task)

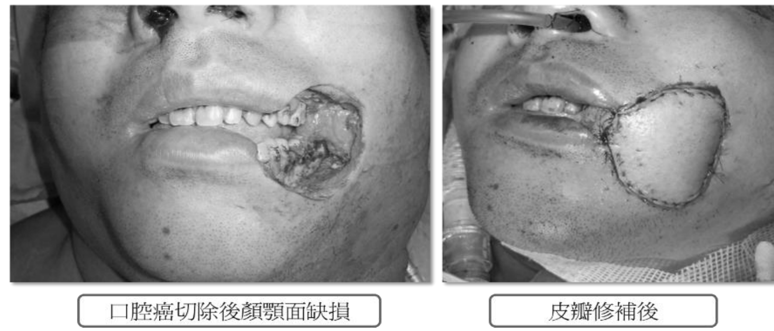


Pathological Voice Detection and Processing

Oral cancer (top five cancer for male in Taiwan).

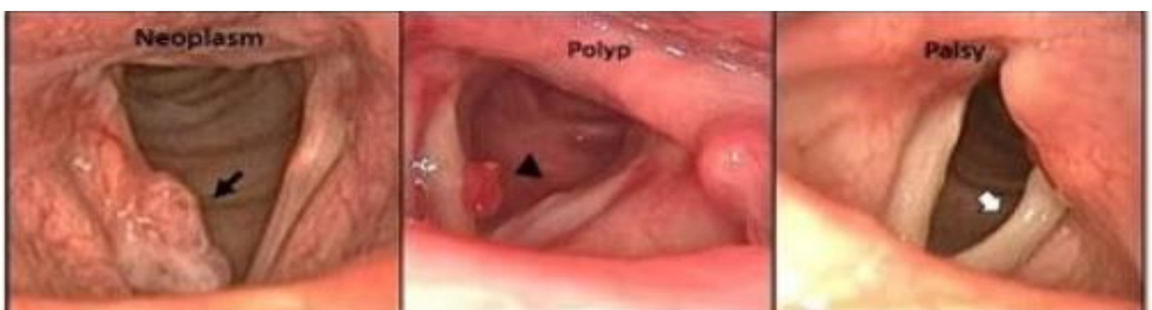


摘自自由時報



臺北榮民總醫院 (口腔醫學部)

Disordered voice



吶

文言版《説文解字》：訥，言難也。

吶語症:主要是跟說話有關的神經或肌肉缺損，造成肌肉動作失調，言語清晰度降低，進而影響溝通。

Detection of Pathological Voice based on Acoustic Signals

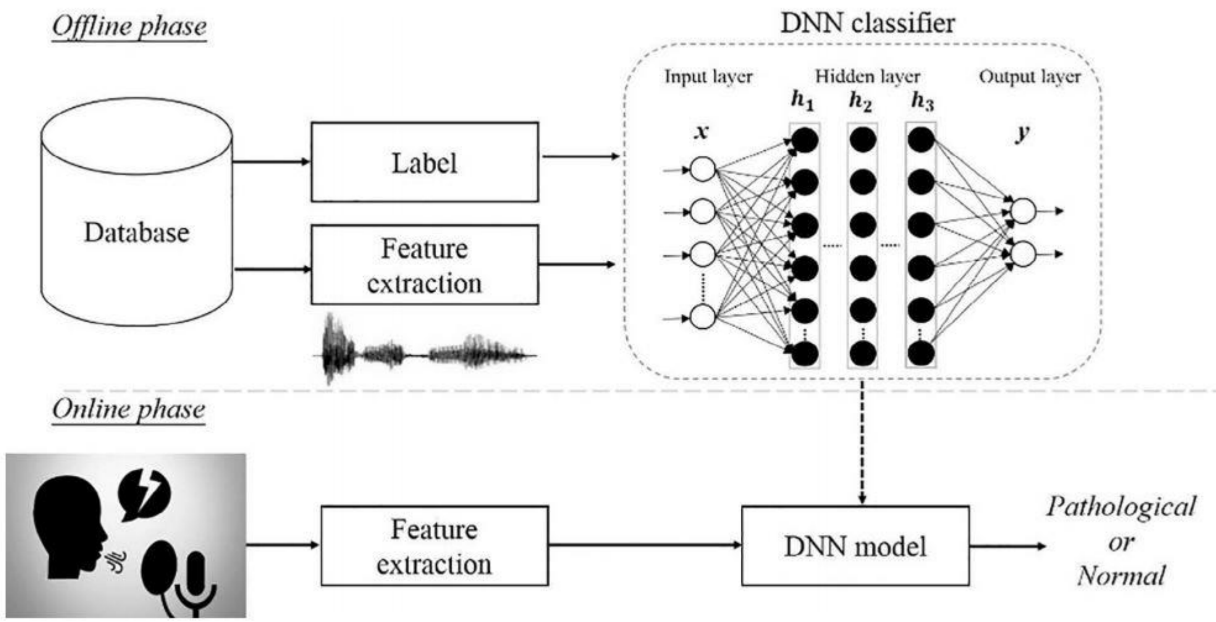
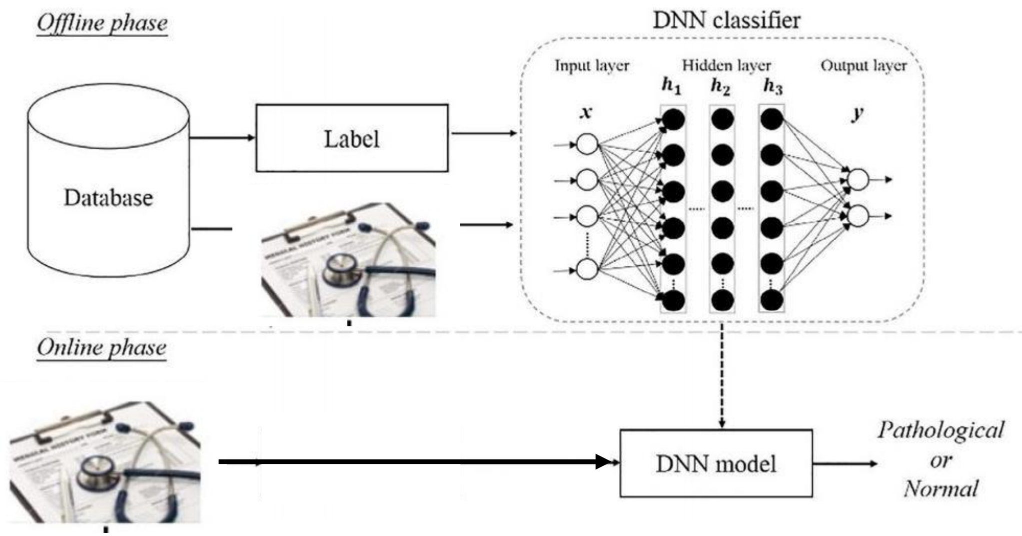


TABLE 5.
Detection of Pathological Voice Samples in the MEEI Voice Disorder Database

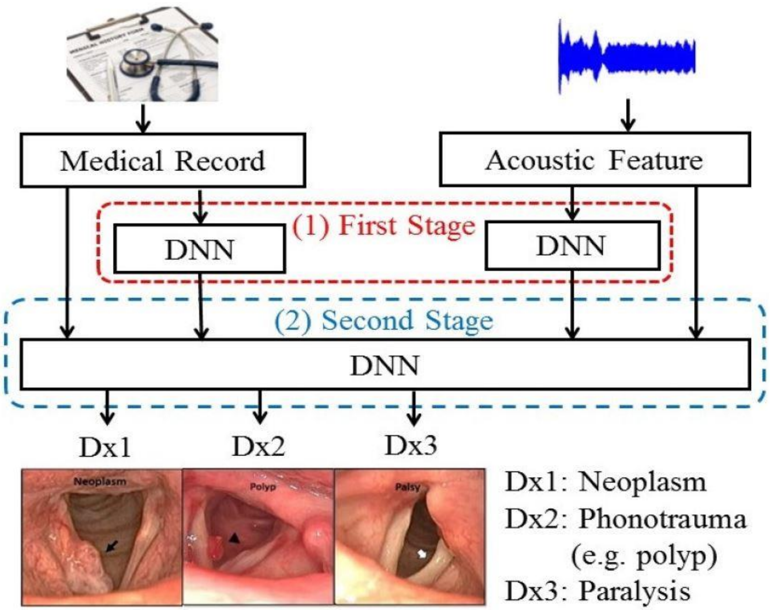
	SVM	GMM	DNN
	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation	Accuracy ± Standard Deviation
MFCC	98.28 ± 2.36%	98.26 ± 1.80%	99.14 ± 1.92%
MFCC + delta	93.04 ± 2.74%	90.24 ± 4.18%	94.26 ± 2.25%
MFCC(N) + delta	87.40 ± 1.92%	90.20 ± 3.83%	90.52 ± 2.00%

Detection of Pathological Voice based on Demographic and Symptomatic Features



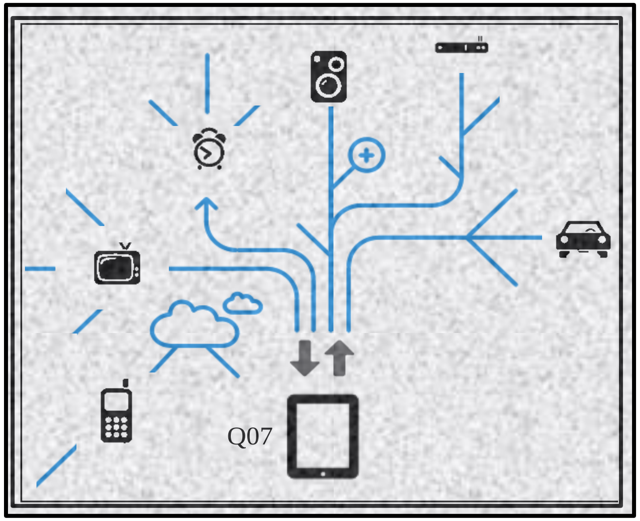
	Neoplasm	Phonotrauma	Palsy	Overall
		Demographics		
SVM	73.0%±10.4%	72.8%±2.34%	67.1%±6.99%	71.0%±3.87%
ANN	80.0%±7.07%	71.2%±2.74%	62.6%±4.33%	71.3%±1.85%
		Symptoms		
SVM	60.0%±12.3%	61.4%±6.07%	73.6%±6.20%	65.0%±3.07%
ANN	56.0%±5.48%	68.4%±5.25%	72.3%±5.40%	65.6%±2.29%
		Demographics + symptoms		
SVM	82.0%±7.58%	79.5%±1.45%	83.9%±7.90%	81.8%±4.87%
ANN	83.0%±7.58%	79.4%±1.83%	86.5%±2.70%	83.0%±1.58%

Detection of Pathological Voice based on Combining Voice and Medical Records

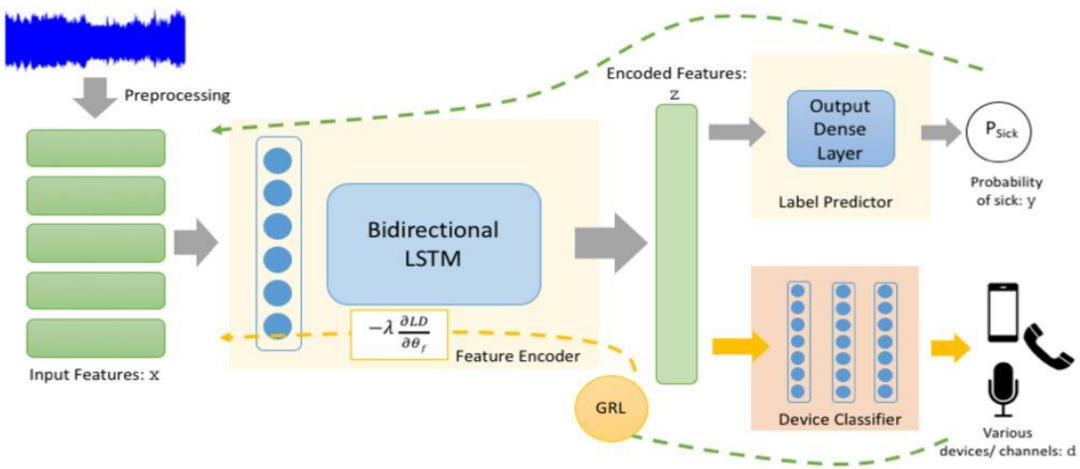


	Neoplasm	Phonotrauma	Vocal Palsy	Accuracy	UAR
		Sensitivity (Recall)			
Acoustic signals	63.00±17.89 (%)	95.36±4.39 (%)	34.40±20.12 (%)	76.94±6.71 (%)	64.25±11.04 (%)
Medical record	59.00±11.40 (%)	91.54±3.67 (%)	70.40±2.19 (%)	81.56±1.25 (%)	73.65±3.49 (%)
TSD	79.00±14.75 (%)	95.36±3.03 (%)	70.40±10.43 (%)	87.26±2.23 (%)	81.59±5.94 (%)

Remote Monitoring on Pathological Voice



Additive noise
+
Convolutive noise (reverberation)



Domain adversarial training

IEEE BigData 2019

The FEHH Voice Data Challenge
IEEE BigData 2019
Los Angeles, CA, USA

Home
Data & Regulation
Document
Leaderboard
Organizers
Registration & Contact Us.

FEMH Voice Data Challenge 2019

Welcome

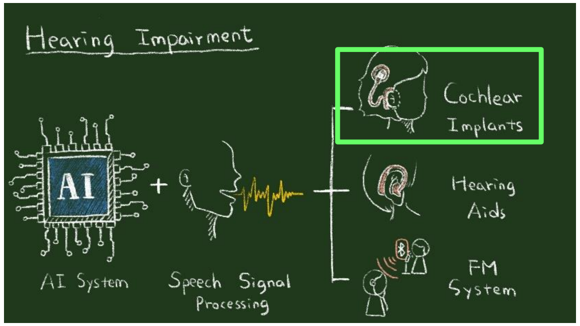
Computerized detection of voice disorders has attracted considerable academic and clinical interest in the hope of providing an effective screening method for voice diseases before endoscopic confirmation. The goal is to detect pathological voice and classify four disordered categories. Different from last year, the task of this year includes both acoustic waveforms and medical records. Therefore, multimodal algorithms should be useful. We believe that the task is more challenging and interesting than last year. We will award the top three teams with medals and cash prizes. We sincerely welcome your participation.

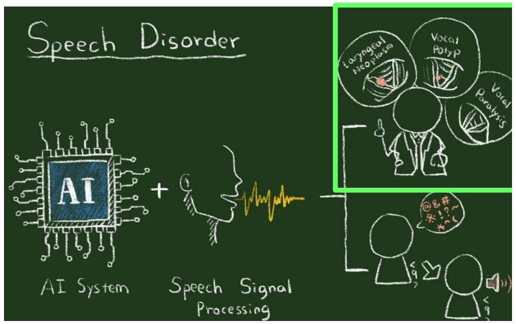
This competition builds on the experience of previous research work and all source codes are available here (Document).

Latest Updates

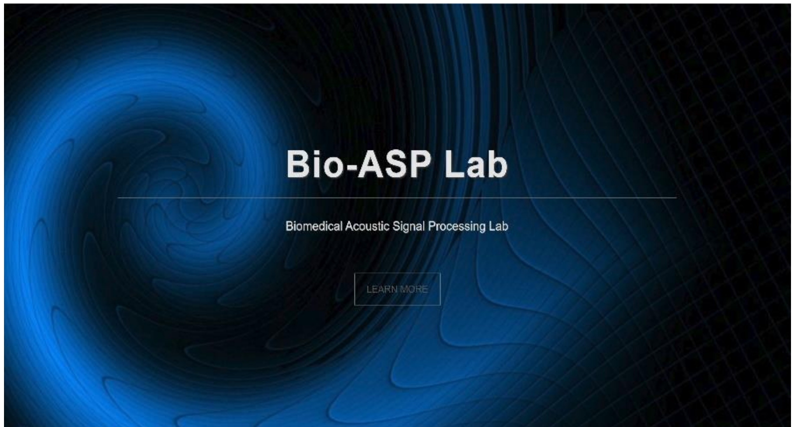
Conclusion

- Deep Learning
  - Artificial intelligence, machine learning, deep learning
  - Human learning versus machine learning
  - Some histories about deep learning
  - Popular deep learning models
- Speech Signal Processing
  - Two categories of tasks: recognition and generation
  - Generation: speech enhancement
  - Recognition: pathological voice recognition





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