THE PLACE OF USEFUL LEARNING

University of Strathclyde Engineering

EE969 Digital Signal Processing

Speech Enhancement

Nikilkumar Patel *202265899*

MSc Machine Learning and Deep Learning 2022-23

“What is speech?” A Layman’s view

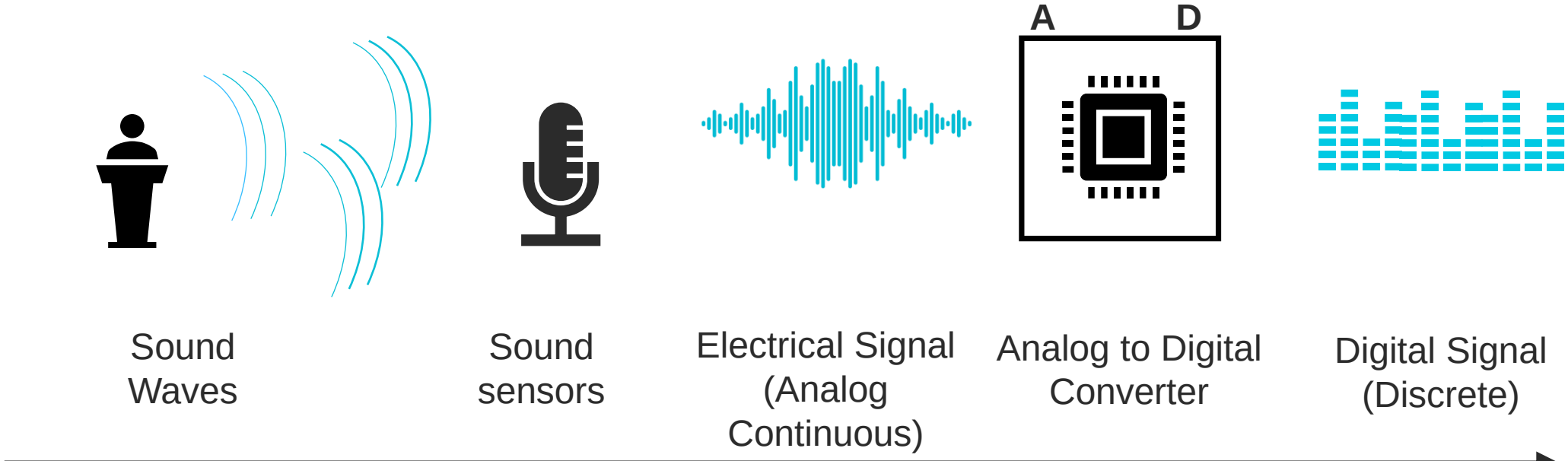
Speech sound from humans

Recorded by device- Digitization

Stored in a Digital medium

Speech ≈ Some type of Signal

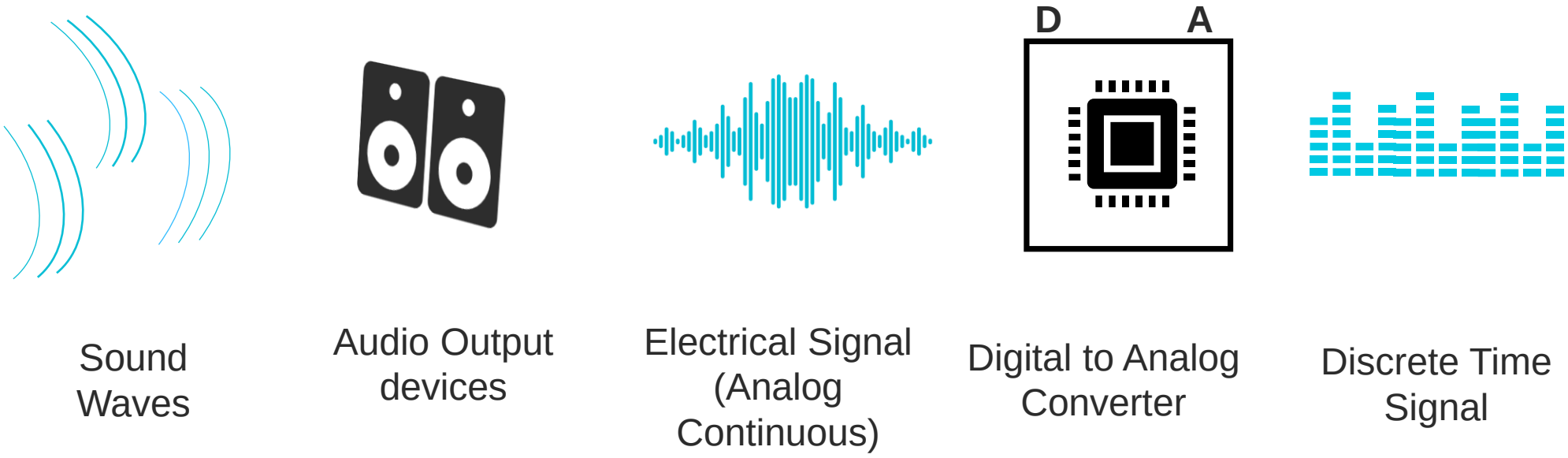
“What is speech?” An Engineer’s view



Sampling *in Digital Audio Recording* – Converting speech into computer ‘bits’

Audio Signal → Electrical Signal → Digital Signal

“What is speech?” An Engineer’s view

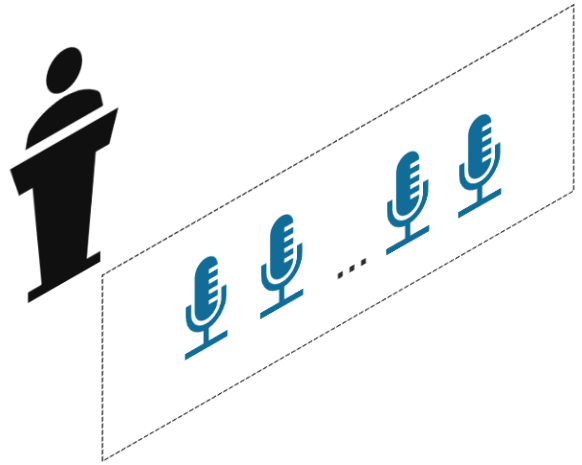


Playing back – Converting Digital signal back to sound waves

<u>Digital Signal → Electrical Signal → Audio Signal</u>

Single channel vs Multichannel Speech

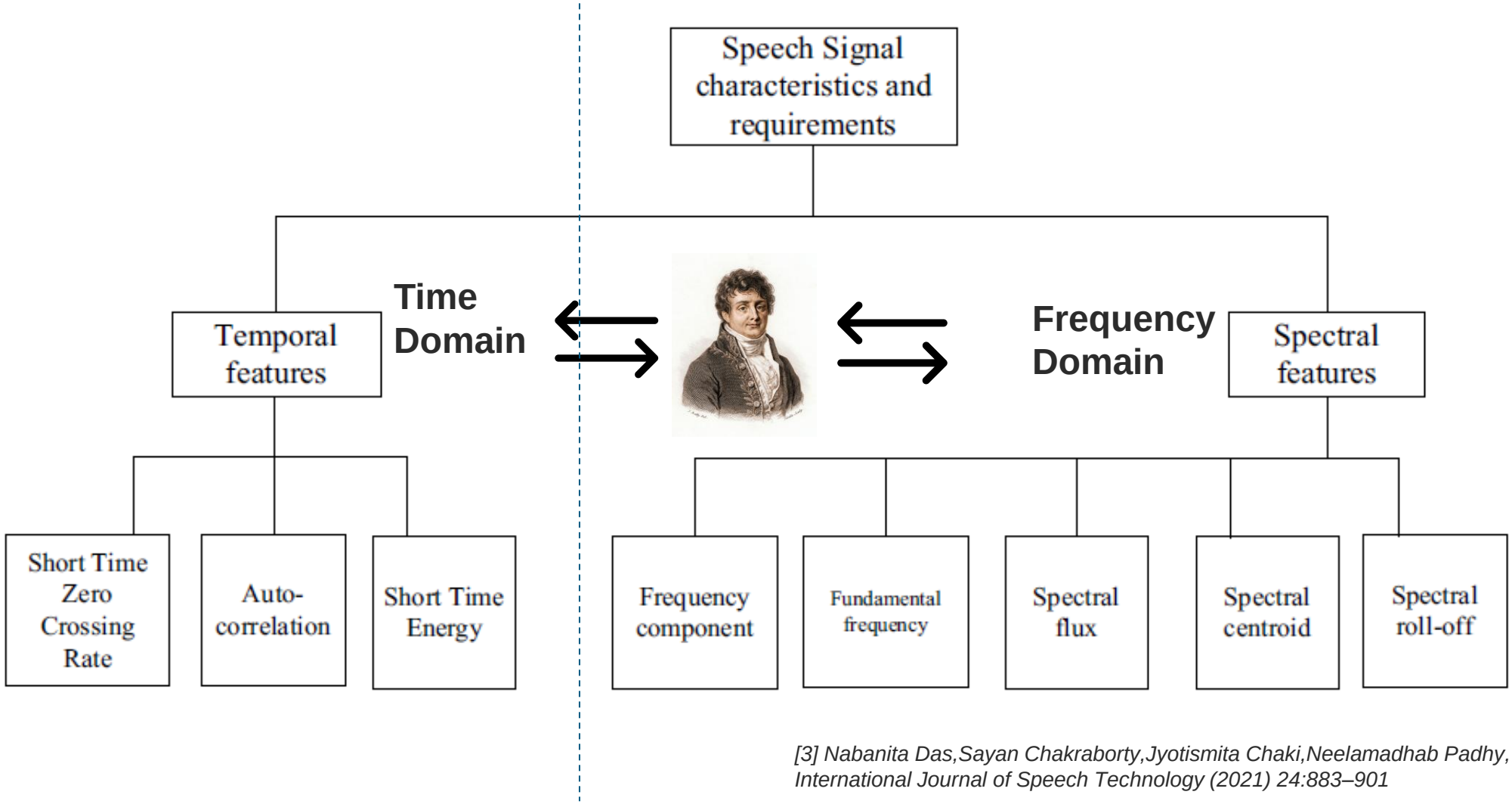
- Simpler and easy analysis and deployment
- Cost effective
- Telephony, voice assistance, voice activated devices



- Improvised speech quality, better noise reduction
- Applications where speech quality and noise reduction are critical

[1] H. S. Dabis and E. Toner, "Single and multiple channel approaches to speech enhancement," IEE Colloquium on Techniques for Speech Processing, London, UK, 1990, pp. 9/1-9/4.
[2] Loizou, P.C. (2013). Speech Enhancement: Theory and Practice, Second Edition (2nd ed.). CRC Press. https://doi.org/10.1201/b14529

Signal Characteristics

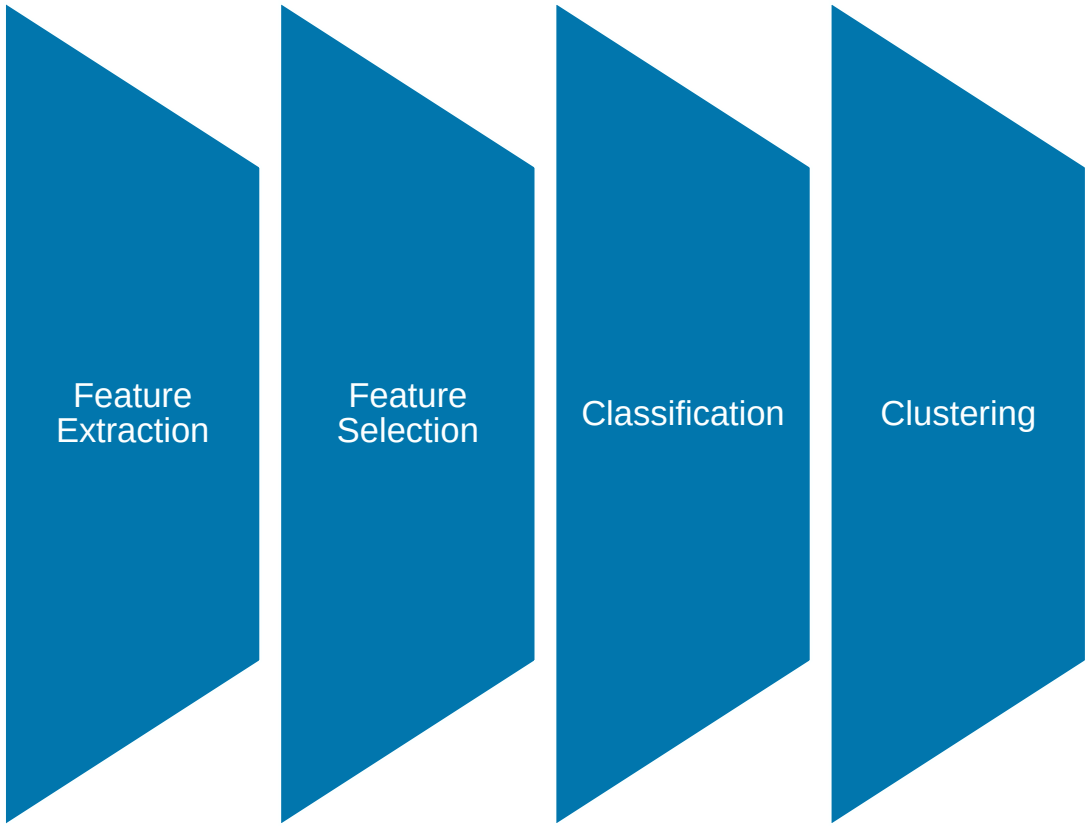


[3] Nabanita Das,Sayan Chakraborty,Jyotismita Chaki,Neelamadhab Padhy, Nilanjan Dey International Journal of Speech Technology (2021) 24:883–901

Why the enhancement?

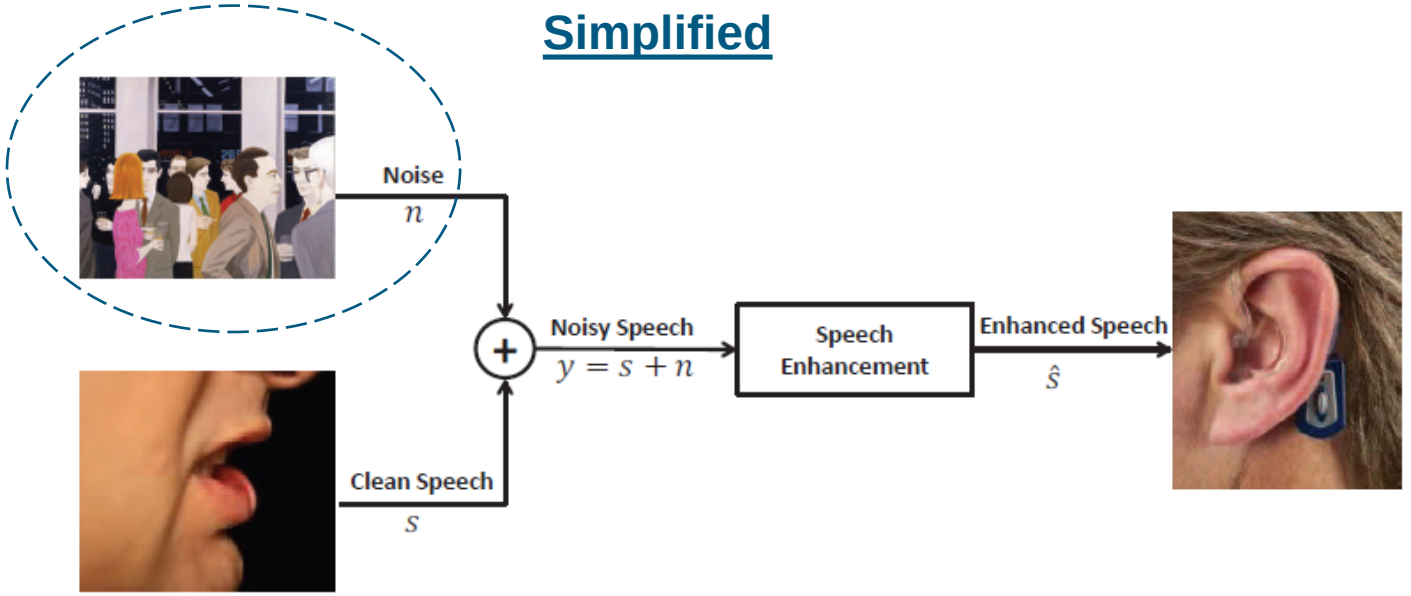
- The world is full of undesired signals.
- The culprits : Noise, Interference and Reverberation
- Quality Degradation due to these unwanted components
  - Speech Inteligibility
  - Speech Quality
  - Listening comfort

Methodology

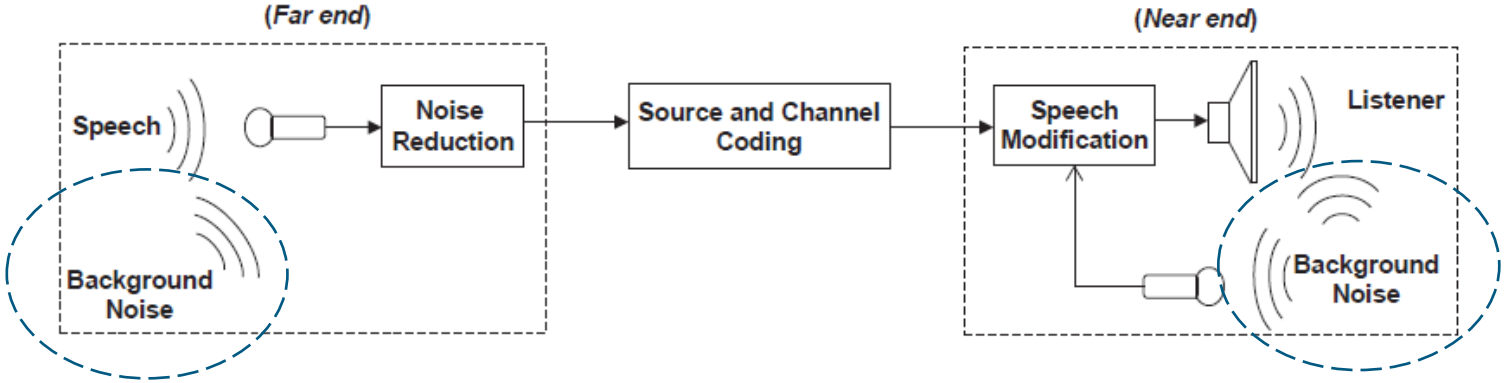


Speech Enhancement System

Simplified



Full schematics



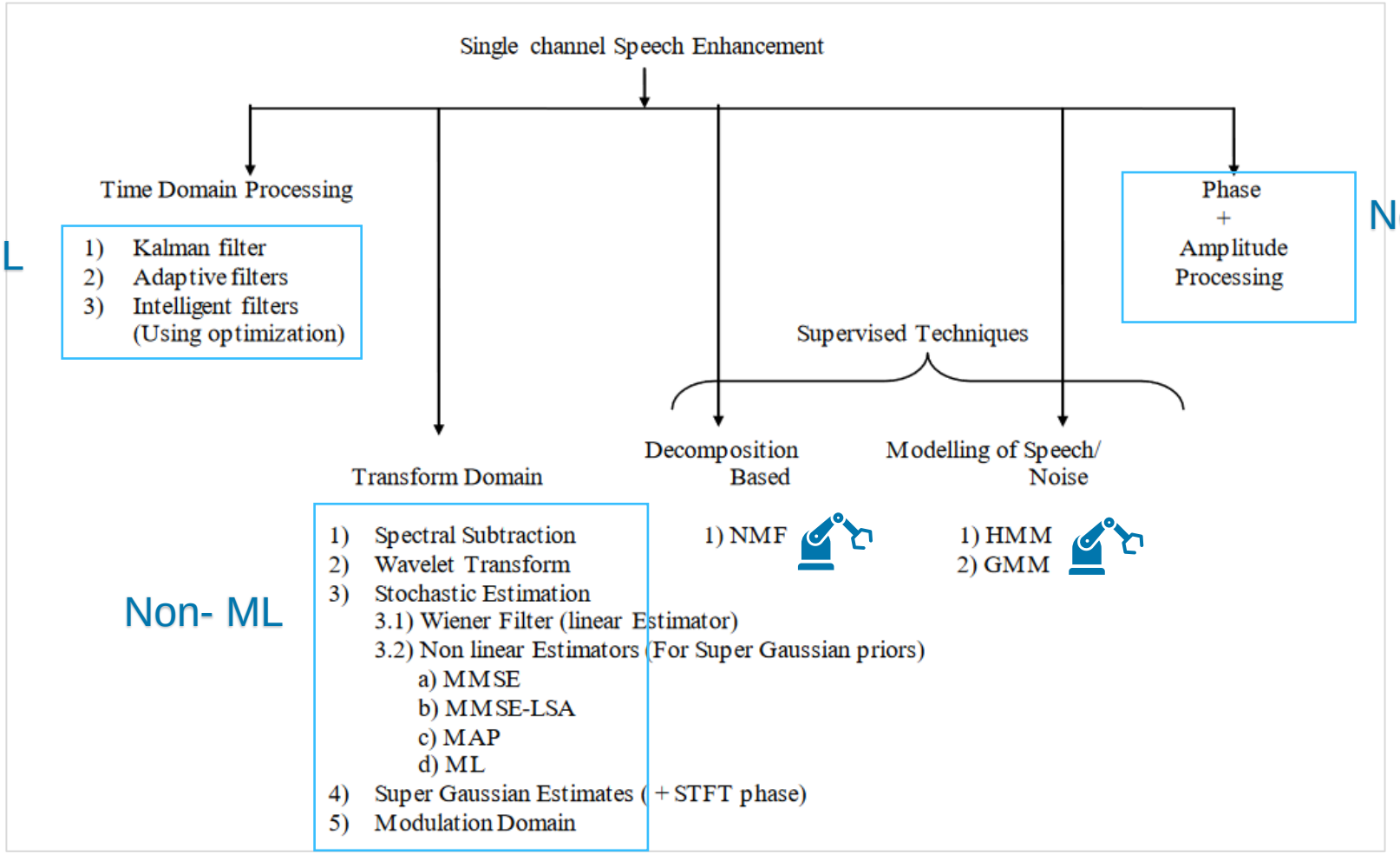
[5] Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models
Nasser Mohammadiha,
Communication Theory Laboratory
School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm

Measurement of Speech Quality

Objective Performance measures	**S**ignal to **N**oise **R**atio (SNR)	→Signal power to Noise power →The higher the better noise reduction
	Perceptual **E**valuation of **S**peech **Q**uality (PESQ)	→Comparison between perceived quality of processed signal and original signal →The higher the better speech quality
	Short-**t**ime **O**bjective **I**ntelligibility (STOI)	→Measures intelligibility by comparison between Original signal and processed signal for short time →The Higher the better Intelligibility
	Speech **T**ransmission **I**ndex (STI)	→measures the intelligibility of speech in a noisy environment by evaluating SNR, reverberation, and other factors →The higher the better the speech ineligibility
Subjective Performance Measures	**M**ean **O**pinion **S**quare(MOS)	→Manual approach to ask human to rate speech quality on scale to 1 to 5; 1 = worst, 5 = best
	Signal Distortion Scale (SIG)	→ Listener attends the quality of speech signal on scale of 1 to 5: 1 = worst, 5 = best
	Background Noise (BAK)	→ Listener attends the background noise signal on scale of 1 to 5: 1 = worst, 5 = best

[4] O. Tymchenko, O. Khamula, B. Havrysh, S. Vasiuta and A. Jagiełło, "Speech quality measurement methods and models over ip-networks," 2020, https://doi.org/10.1109/REM49740.2020.9313079

State of the Art Techniques - Traditional Approach in SE



[6] Ravi, K. K., & Subbaiah, P. V. (2016). A survey on speech enhancement methodologies. International Journal of Intelligent Systems and Applications, 8(12), 37. Retrieved from https://www.proquest.com/scholarly-journals/survey-on-speech-enhancement-methodologies/docview/1884171951/se-2

Welcome to the realm of Deep Learning

where… things can get extremely complicated and chaotic.

But it works!

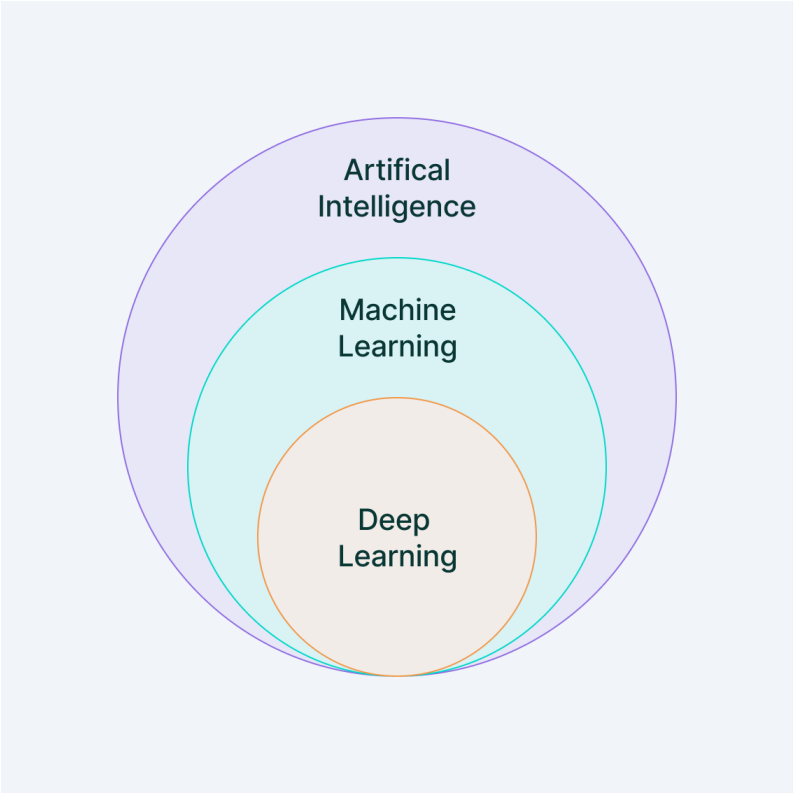
- can detect cancer more accurately than do doctors.
- has beaten the world champion in the game of Go
- can figure out appropriate tax policies when traditional economic models would be too complex to solve.
- Can code for you and generate images, sound, videos and even PPTs.

But….

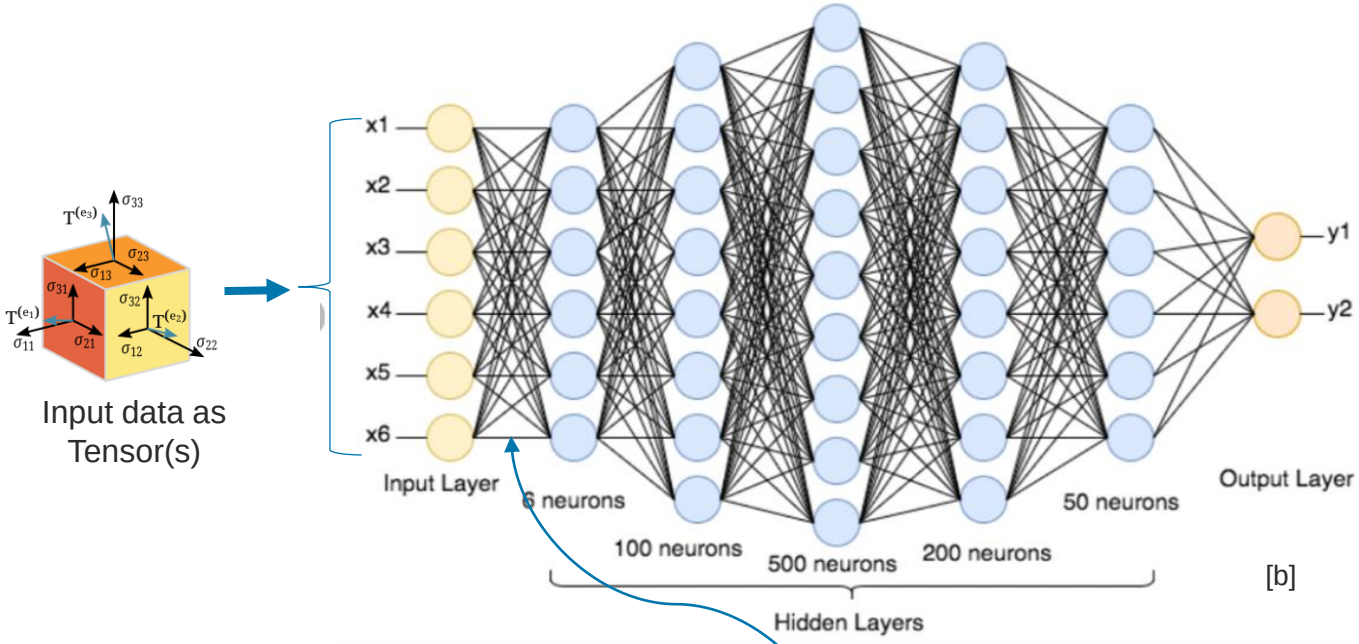
- Has a huge computational cost to train them
- Has not so much explainability (now it has actually!)

Istockphoto
[7] https://spectrum.ieee.org/deep-learning-sustainability

An extremely brief introduction to Deep Learning



[a]



[b]

Deep Learning ≈ Neural Networks with multiple “layers”

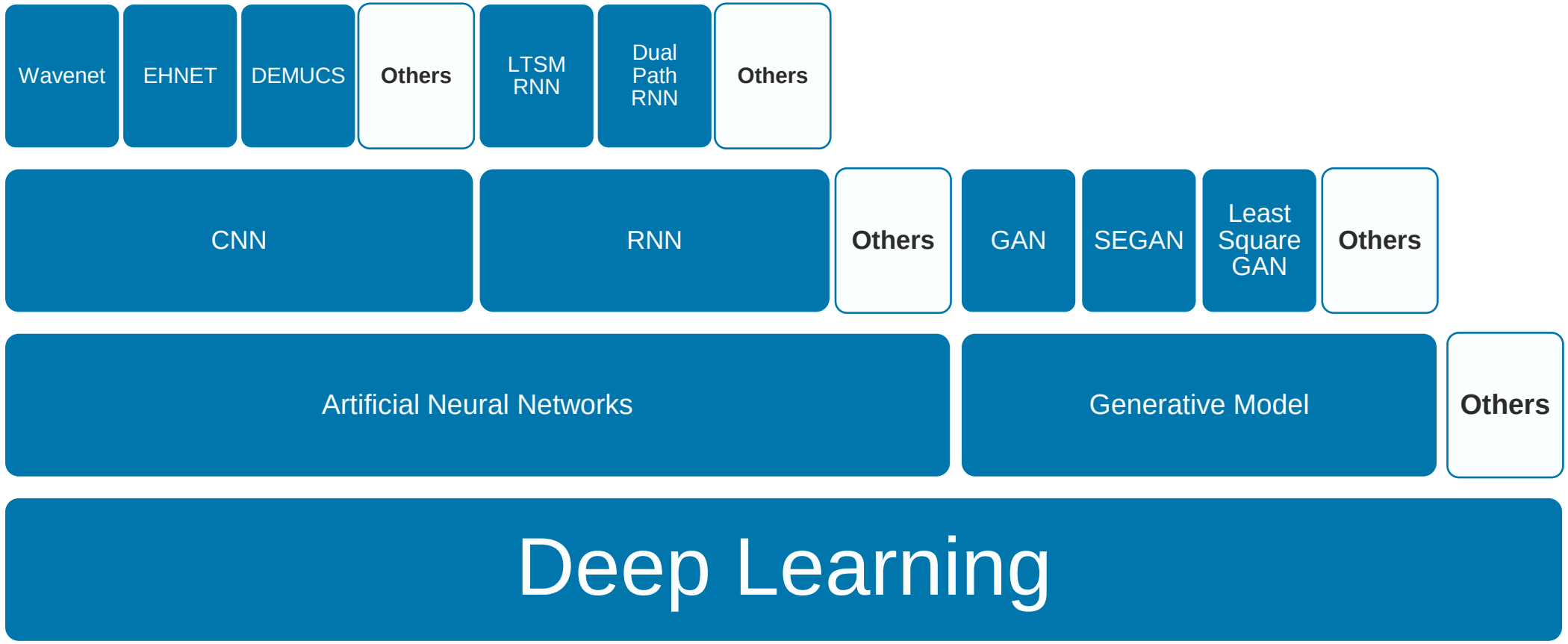
Each connection in each layer is having its “importance” called weight

Each neuron input is dot product of input and the respective weights

Each neuron output is an activation function (binary, sigmoid, ReLU etc.)

[8] https://www.v7labs.com/blog/deep-learning-guide
[9] https://levity.ai/blog/difference-machine-learning-deep-learning
[10] Deep Learning for Ligand-Based Virtual Screening in Drug Discovery - Scientific Figure on ResearchGate

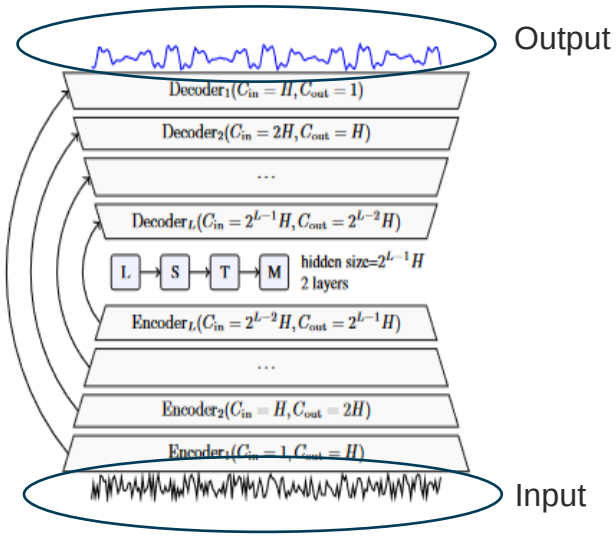
State of the Art Techniques – Deep Learning Approach in SE



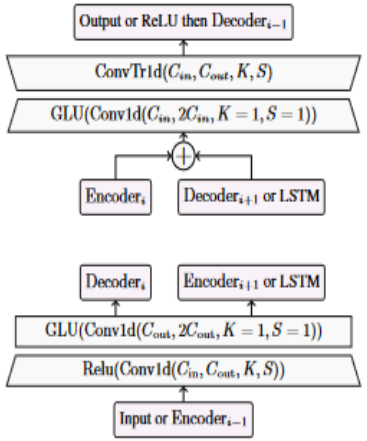
[11]Deep Learning for Speech Enhancement: A Study on WaveNet, GANs and General CNN RNN Architectures: OSCAR XING LUO

Deep Learning in action – CNN based SE

DEMUCS Architecture (Facebook AI Research - FAIR)



(a) *Causal* DEMUCS *with the noisy speech as input on the bottom and the clean speech as output on the top. Arrows represents U-Net skip connections. H controls the number of channels in the model and L its depth.*

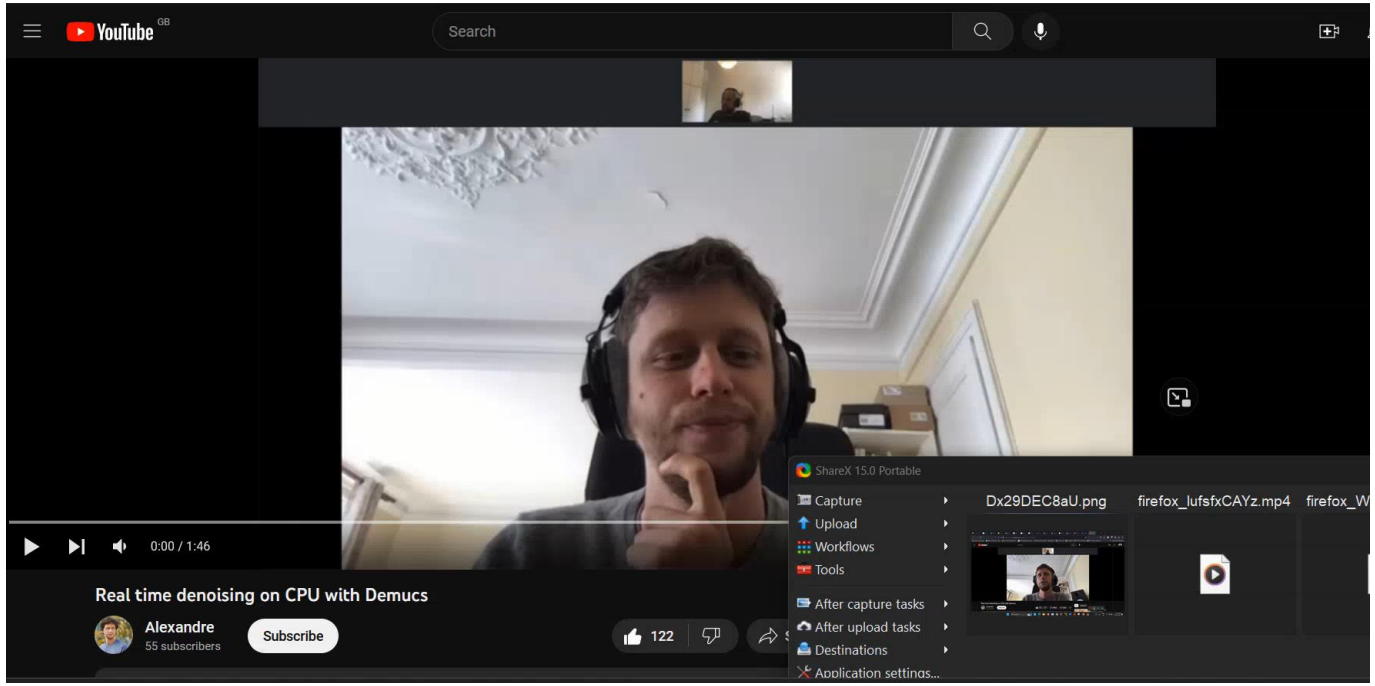


(b) *View of each encoder (bottom) and decoder layer (top). Arrows are connections to other parts of the model.* C_{in} *(resp.* C_{out}*) is the number of input channels (resp. output), K the kernel size and S the stride.*

- Developed for single channel speech/audio raw input
- Based on modified UNET (CNN developed for Biomedical Imaging) with skip connections
- Model parameters are initialized from ImageNet
- Uses multilayer CNN with encoders and decoders with ReLU activations with stride size of 16
- Uses LTSM on encoder's latent up sampled output and decoder yields down sampled estimations from GLU Activation

[12] Real Time Speech Enhancement in the Waveform Domain Alexandre Defossez, Gabriel Synnaeve, Yossi Adi | https://doi.org/10.48550/arXiv.2006.12847

Demo – Live Speech Enhancement using DEMUCS



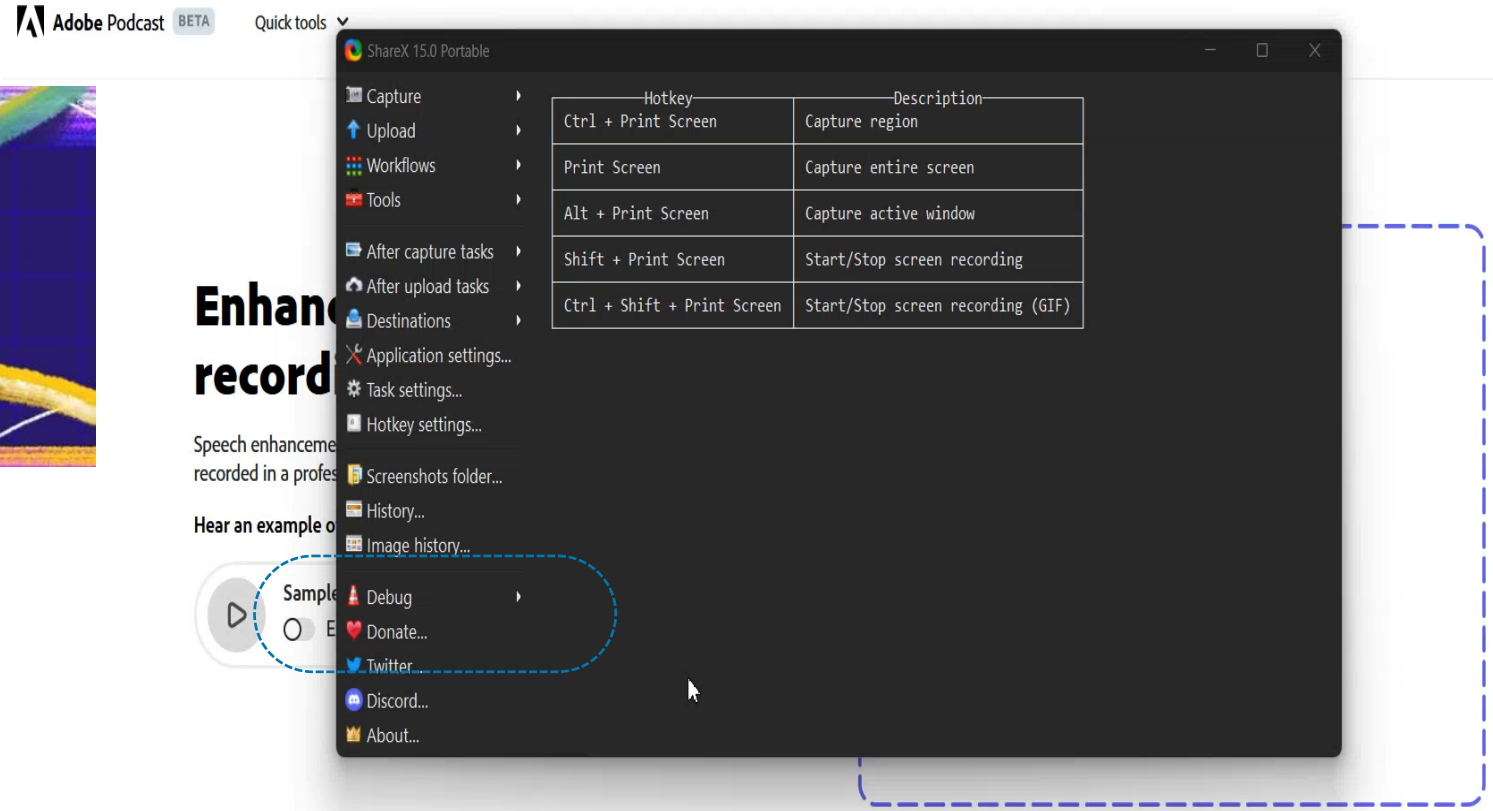
- It works extremely well in real time with random noise i.e., the 1st scenario and high volume of noise i.e., 2nd scenario when the noise volume is turned up.
- It works even if there is a high intensity nonspeech signal exists in the background. i.e., 3rd scenario when there was a hammer sound with high pitch.
- This is low-cost digital implementation that runs faster than real-time on a single laptop CPU core on open-source license.
- This also can be used to denoising the recorded speech audio raw files.

[13] https://github.com/facebookresearch/denoiser
[14] Demo source: https://www.youtube.com/@adefossez

Demo – Proprietary AI Powered Audio Processing on Cloud

[19] https://labs.adobe.com/projects/shasta/

[20] https://podcast.adobe.com



Performance Evaluation

	PESQ	STOI (%)	pred. CSIG	pred. CBAK	pred. COVL	MOS SIG	MOS BAK	MOS OVL	Causal
Noisy	1.97	91.5	3.35	2.44	2.63	4.08	3.29	3.48	-
SEGAN [7]	2.16	-	3.48	2.94	2.80	-	-	-	No
Wave U-Net [20]	2.40	-	3.52	3.24	2.96	-	-	-	No
SEGAN-D [8]	2.39	-	3.46	3.11	3.50	-	-	-	No
MMSE-GAN [21]	2.53	93	3.80	3.12	3.14	-	-	-	No
MetricGAN [22]	2.86	-	3.99	3.18	3.42	-	-	-	No
DeepMMSE [23]	2.95	94	**4.28**	**3.46**	**3.64**	-	-	-	No
DEMUCS (H=64, S=2 ,U=2)	**3.07**	**95**	**4.31**	3.4	**3.63**	4.02	3.55	3.63	No
Wiener	2.22	93	3.23	2.68	2.67	-	-	-	Yes
DeepMMSE [23]	2.77	93	4.14	**3.32**	**3.46**	4.11	**3.69**	**3.67**	Yes
DEMUCS (H=48,S=4, U=4)	**2.93**	**95**	**4.22**	3.25	**3.52**	4.08	3.59	3.40	Yes
DEMUCS (H=64,S=4, U=4)	2.91	**95**	**4.20**	3.26	**3.51**	4.03	**3.69**	3.39	Yes
DEMUCS (H=64,S=4, U=4) + dry=0.05	2.88	**95**	4.14	3.21	**3.54**	4.10	3.58	**3.72**	Yes
DEMUCS (H=64,S=4, U=4) + dry=0.1	2.81	**95**	4.07	3.10	3.42	**4.18**	3.45	3.60	Yes

Measures Used

1) **PESQ**: Wideband 04 to 4.5dB

2) **STOI**: 0% to 100%

3) **CSIG**: SIG Scale: 1 to 5

4) **CBAK**: BAK Scale : 1 to 5

5) **OVL**: Overall scale: 1 to 5

[16] https://doi.org/10.48550/arXiv.2006.12847

Rationale – Why Deep Learning is effective?

- **Speech-Noise Non-liner relationship**: Neural Networks are the best tools to model any non-liner relationship
- **Universal Function Approximation**: Neural networks are function approximators which means they can estimate any complex function give the enough neurons and layers
- **Robustness**: Conventional models rely on statistical modelling, where Neural Networks can be tuned to any design criteria as they are data-driven
- **Flexibility**: Neural Networks takes advantage of “Transfer Learning” where one architecture can serve as basis for a new hybrid or improvised architecture. For example: DEMUCS is made from UNET
- **Advancement in Computational power**: More and more parameters can be added to increase/ fine tune the complexity of neural networks to achieve any arbitrary accuracy. Takes time to train but, much more effective and robust results
- **The research community**: Due to its ubiquitousness, Deep Learning field is backed by many rapid and recent ground-breaking developments with their open source implementations

[17] https://www.researchgate.net/publication/354252395_Speech_Enhancement_Using_Deep_Learning_Methods_A_Review
[18] Yiting Wang and Zhenhua Wei 2020 J. Phys.: Conf. Ser. **1650** 032163 | https://doi.org/ 10.1088/1742-6596/1650/3/032163

Areas of Application

- **Telecommunications**: To improve the quality of voice communication in telephone systems, mobile phones, and video conferencing systems
- **Hearing aids**: To help people with hearing impairments hear speech more clearly
- **Voice assistants**: Such as Amazon Alexa, Google Assistant, and Apple Siri to improve speech recognition accuracy in noisy environments
- **Automatic Speech Recognition (ASR)**: Can improve the accuracy of ASR systems by reducing noise and reverberation in the speech signal
- **Audio and video recording**: To improve the quality of recorded speech
- **Speech therapy**: To improve the clarity and intelligibility of speech for individuals with speech disorders.
- **Forensic investigations**: To improve the quality of speech recordings that are used as evidence.
- **Military and law enforcement**: To improve the clarity of communications in noisy environments.
- And many more…..

References

[1] H. S. Dabis and E. Toner, "Single and multiple channel approaches to speech enhancement," IEE Colloquium on Techniques for Speech Processing, London, UK, 1990, pp. 9/1-9/4.

[2] Loizou, P.C. (2013). Speech Enhancement: Theory and Practice, Second Edition (2nd ed.). CRC Press. https://doi.org/10.1201/b14529

[3] Nabanita Das,Sayan Chakraborty,Jyotismita Chaki,Neelamadhab Padhy, Nilanjan Dey International Journal of Speech Technology (2021) 24:883–901

[4] O. Tymchenko, O. Khamula, B. Havrysh, S. Vasiuta and A. Jagiełło, "Speech quality measurement methods and models over ip-networks," 2020, doi: https://doi.org/10.1109/REM49740.2020.9313079

[5] Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models
Nasser Mohammadiha,Communication Theory Laboratory, School of Electrical Engineering,KTH Royal Institute of Technology, Stockholm

[6] Ravi, K. K., & Subbaiah, P. V. (2016). A survey on speech enhancement methodologies. International Journal of Intelligent Systems and Applications, 8(12), 37. https://www.proquest.com/scholarly-journals/survey-on-speech-enhancement-methodologies/docview/1884171951/se-2

[7] https://spectrum.ieee.org/deep-learning-sustainability

[8] https://www.v7labs.com/blog/deep-learning-guide

[9] https://levity.ai/blog/difference-machine-learning-deep-learning

[10] Deep Learning for Ligand-Based Virtual Screening in Drug Discovery - Scientific Figure on ResearchGate

[11]Deep Learning for Speech Enhancement: A Study on WaveNet, GANs and General CNN RNN Architectures: OSCAR XING LUO

[12] Real Time Speech Enhancement in the Waveform Domain Alexandre Defossez, Gabriel Synnaeve, Yossi Adi | https://doi.org/10.48550/arXiv.2006.12847

[13] https://github.com/facebookresearch/denoiser

[14] Demo source: https://www.youtube.com/@adefossez

[15] WEIGHTED SPEECH DISTORTION LOSSES FOR NEURAL-NETWORK-BASED REAL-TIME SPEECH ENHANCEMENT Yangyang Xia1, Sebastian Braun2, Chandan K. A. Reddy2, Harishchandra Dubey2, Ross Cutler2, Ivan Tashev2 | 1 Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA | 2 Microsoft Corporation, Redmond, WA, USA

[16]https://github.com/microsoft/MS-SNSD

[17] https://doi.org/10.48550/arXiv.2006.12847

[18]https://www.researchgate.net/publication/354252395_Speech_Enhancement_Using_Deep_Learning_Methods_A_Review

*[19] Yiting Wang and Zhenhua Wei 2020 J. Phys.: Conf. Ser. **1650** 032163| https://doi.org/ 10.1088/1742-6596/1650/3/032163*

[20] https://labs.adobe.com/projects/shasta/

[21] https://podcast.adobe.com

University of
Strathclyde
Engineering