A Regression Approach to Speech Enhancement Based on Deep Neural Networks

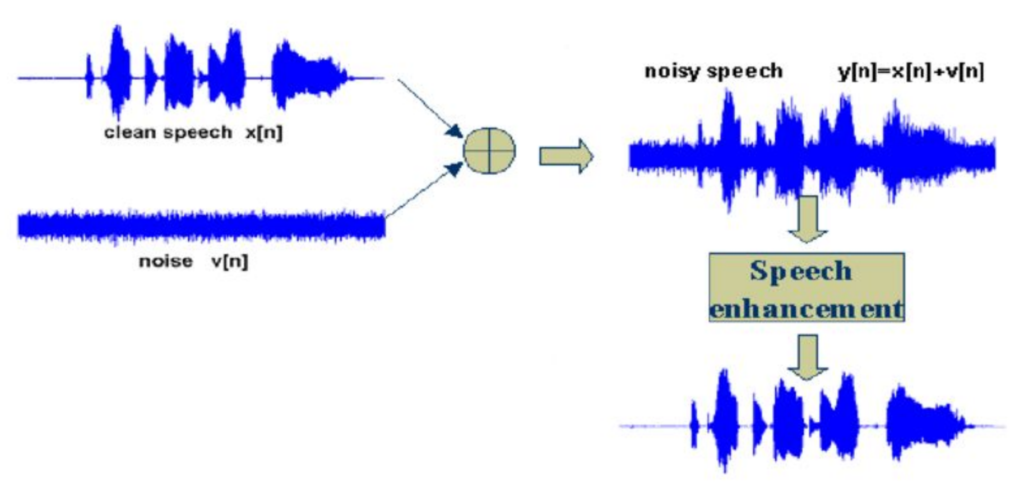
Micah Baldonado, Anushka Deshmukh, Hasan Dheyaa, William Mclain

Background Information

- Over 1.5 billion people worldwide are affected by hearing impairment, with estimates projecting an increase to 2.5 billion by 2050.
- Hearing impairment can result from various factors such as trauma, genetics, and proximity effects of other diseases.
- Those with hearing impairment struggle to perceive sounds due to a narrower dynamic range and difficulty differentiating relative intensity.
- Existing solutions like hearing aids and cochlear implants struggle to detect dynamic speech patterns and perform effectively in noisy environments.

Background Information

- Using deep neural networks to enhance speech quality and intelligibility for hearing-impaired individuals.
- The model will address background noise, reduce distortion, and adapt to users' hearing needs
- Possible integration into hearing aids and cochlear implants, to improve user outcomes.
- Aims to bridge the communication gap to for individuals with hearing impairment, enhancing their quality of life



Method Description

Input Data

- Our inputs come in pairs:

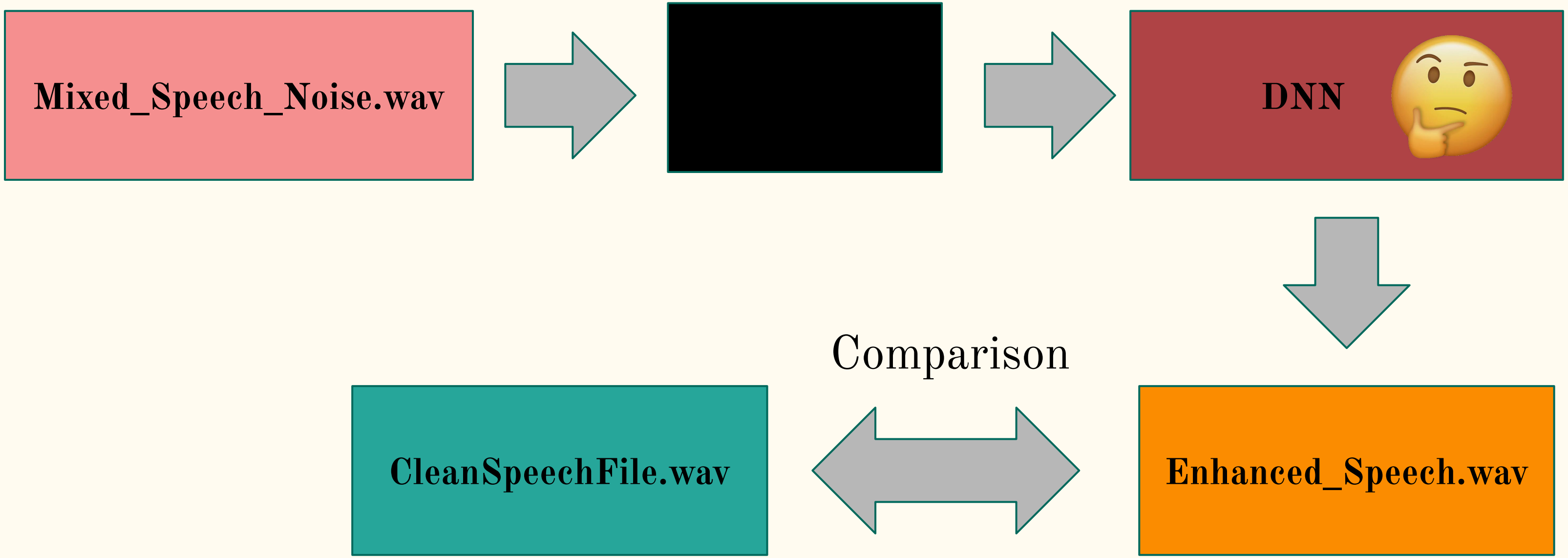
CleanSpeechFile.wav   **Noise.wav**

- We mix our input data to simulate real-world speech-enhancement scenarios

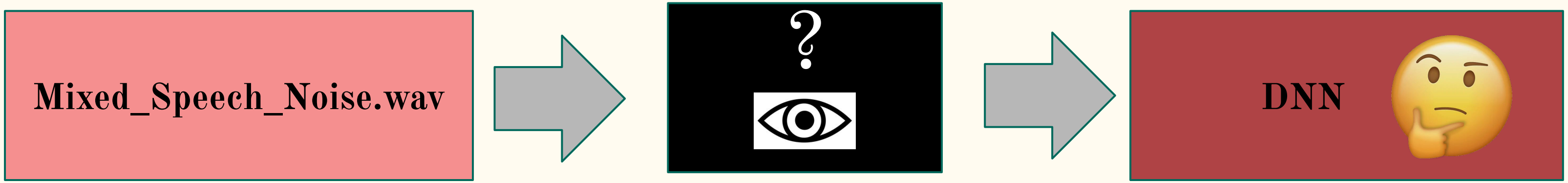
CleanSpeechFile.wav + **Noise.wav** = **Mixed_Speech_Noise.wav**

Evaluating Performance

- How we evaluate Deep Neural Network (DNN) Performance



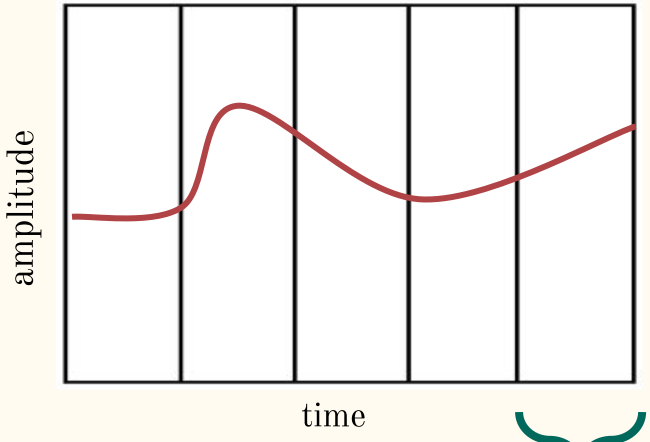
How can we get the DNN to see the data?



- Preprocessing
  - Prepare the sound data in a similar way to how humans process it

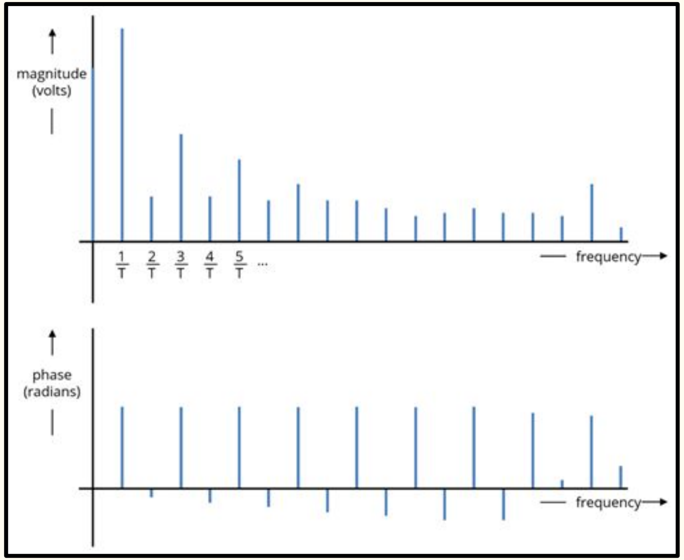
Frequency Analysis

- From each individual sample, we use a Discrete Fourier Transform (DFT) to calculate the frequency components



1 Frame

DFT

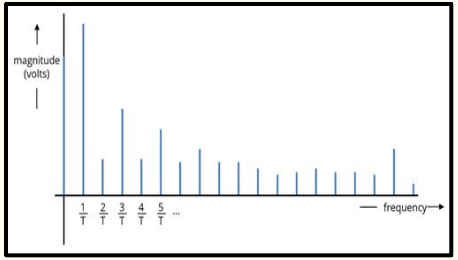


Frequency Domain (magnitude, phase)

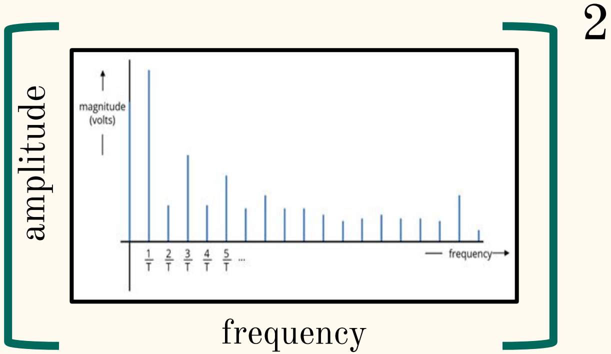
Preprocessing: Converting Frequency of Sound Files

1)

amplitude

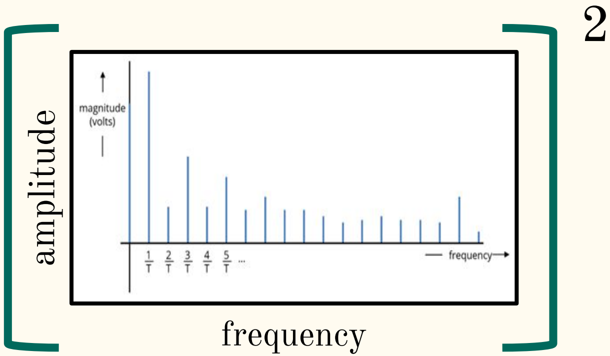


frequency

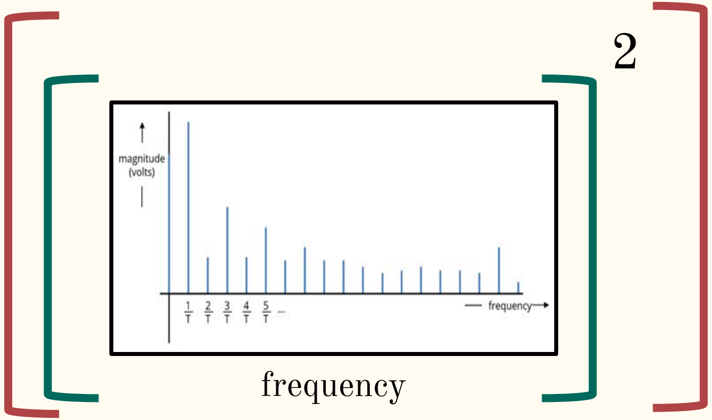


= power spectrum

2)



log



*scales amplitude proportional to human hearing

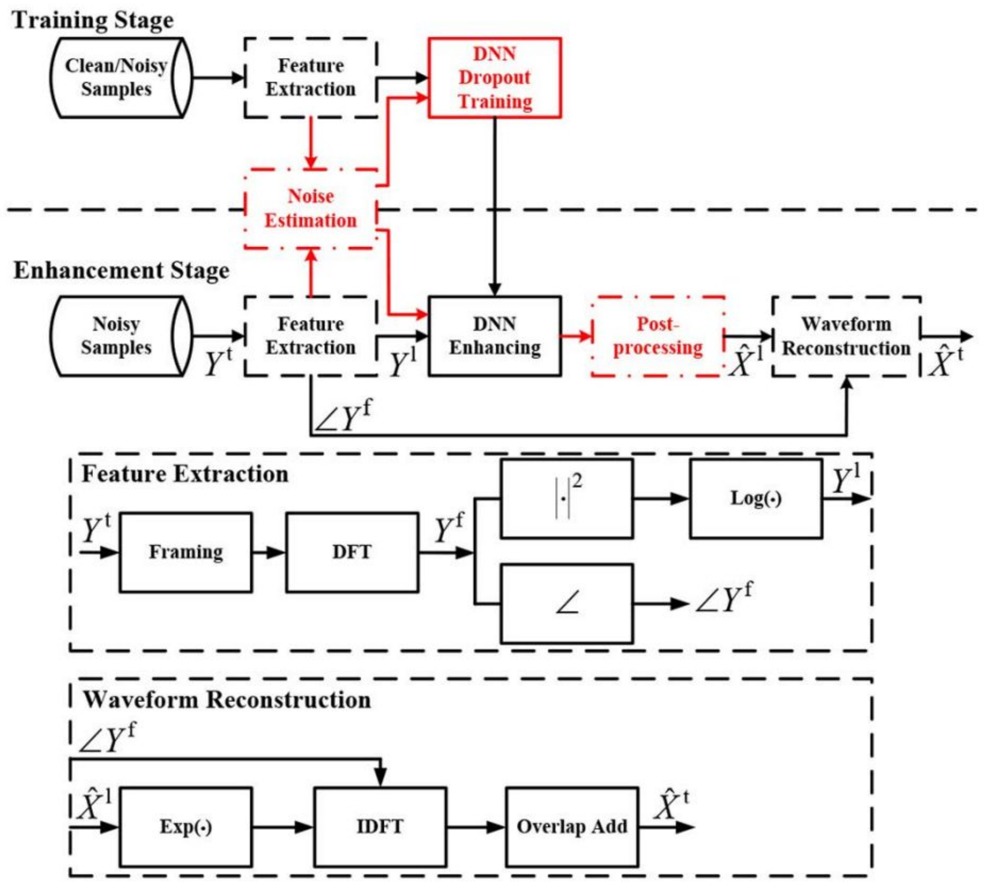


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

Architecture.

(Xu et al.)

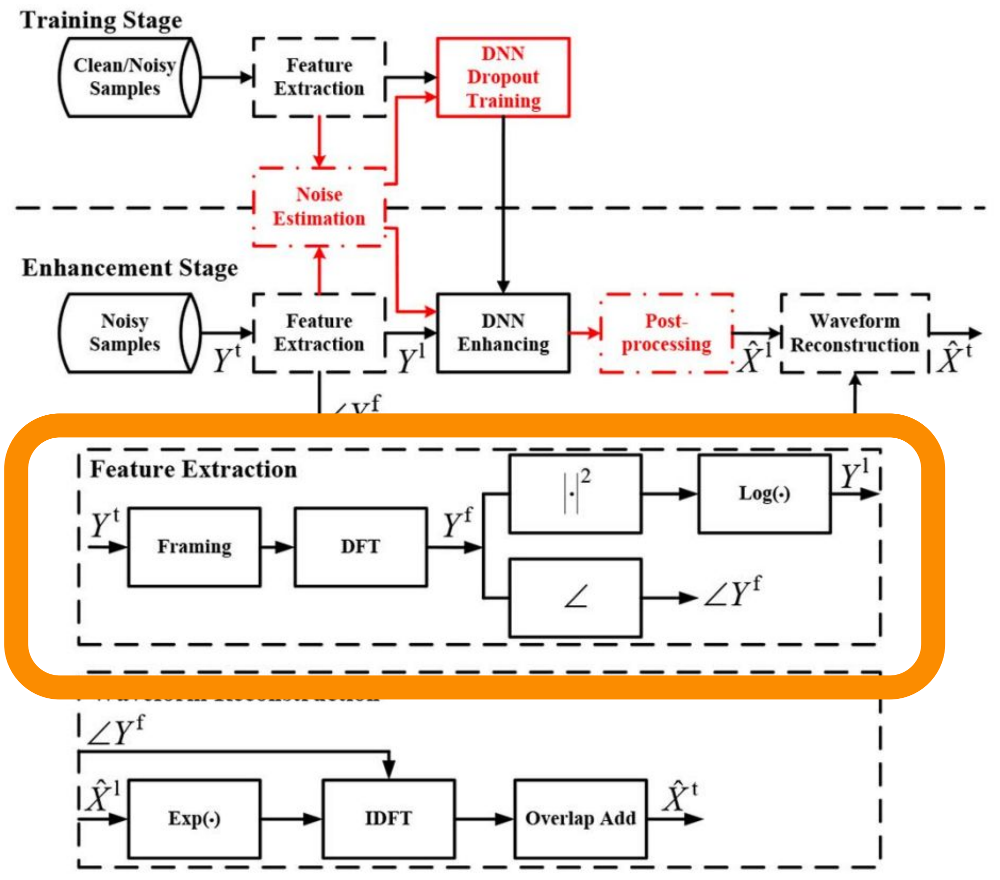


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

(Xu et al.)

Architecture:

So far this is is Really Just Feature Extraction…

- We preprocess our data via a feature extractor…
- This **allows our network to see our sound data (preprocessing)**

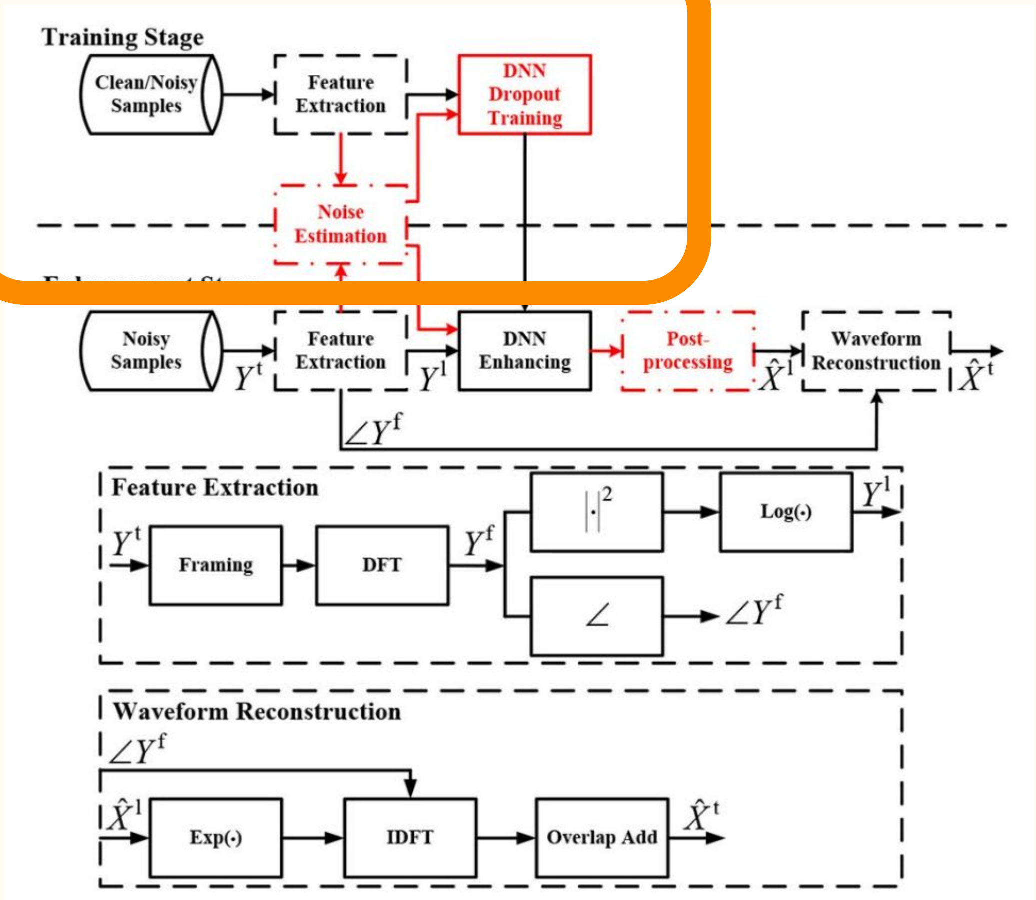


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

(Xu et al.)

Architecture:

- Training Stage

After feature extraction, DNN dropout training **refines its weights/biases** based on supervised learning from the feature extraction.

The weights and biases are tweaked to fully distinguish speech from noise

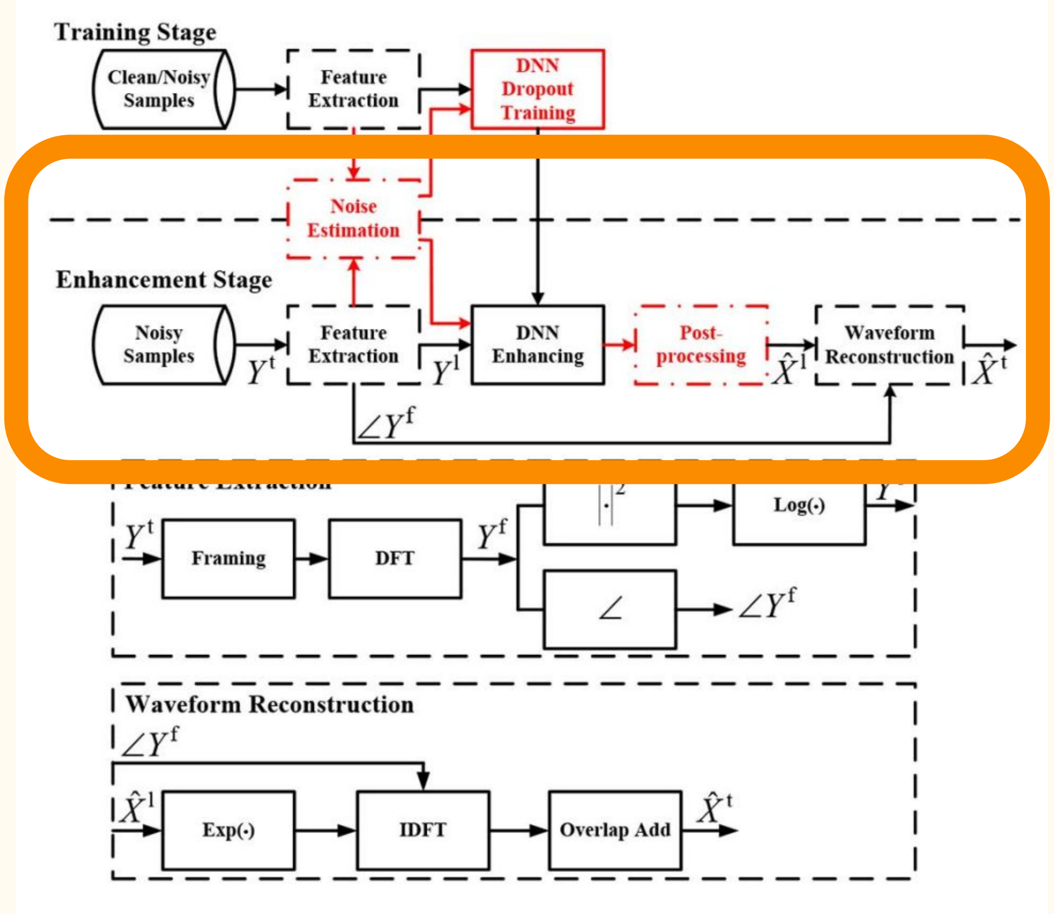


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

Architecture:

- Enhancement Stage

With tweaked weights and biases, the DNN Enhancing section **isolates** the clean speech from the noise sound files **in the frequency domain**

(Xu et al.)

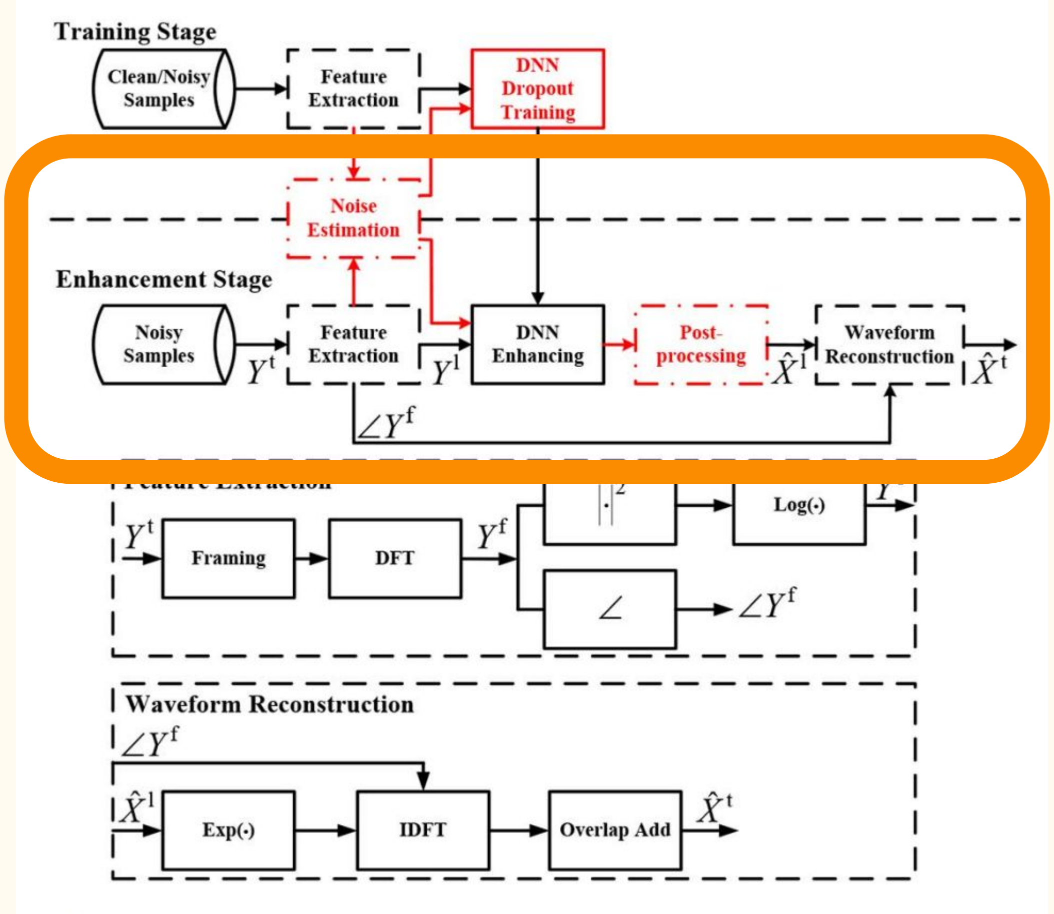


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

Architecture:

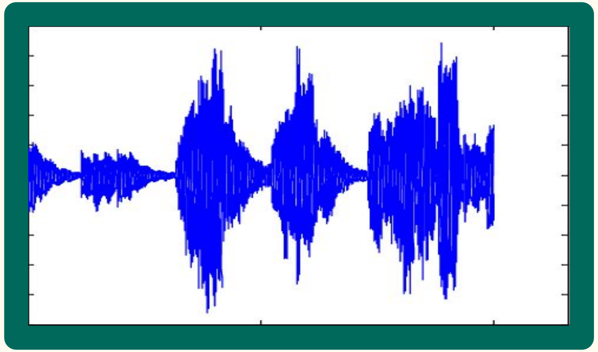
- Enhancement Stage

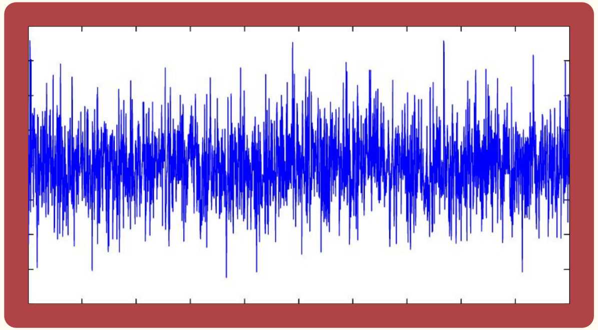
After post-processing, the signal is converted from the frequency domain to the time domain

End Result: Enhanced Sound File

(Xu et al.)

Input: Dataset descriptions





Speech data:

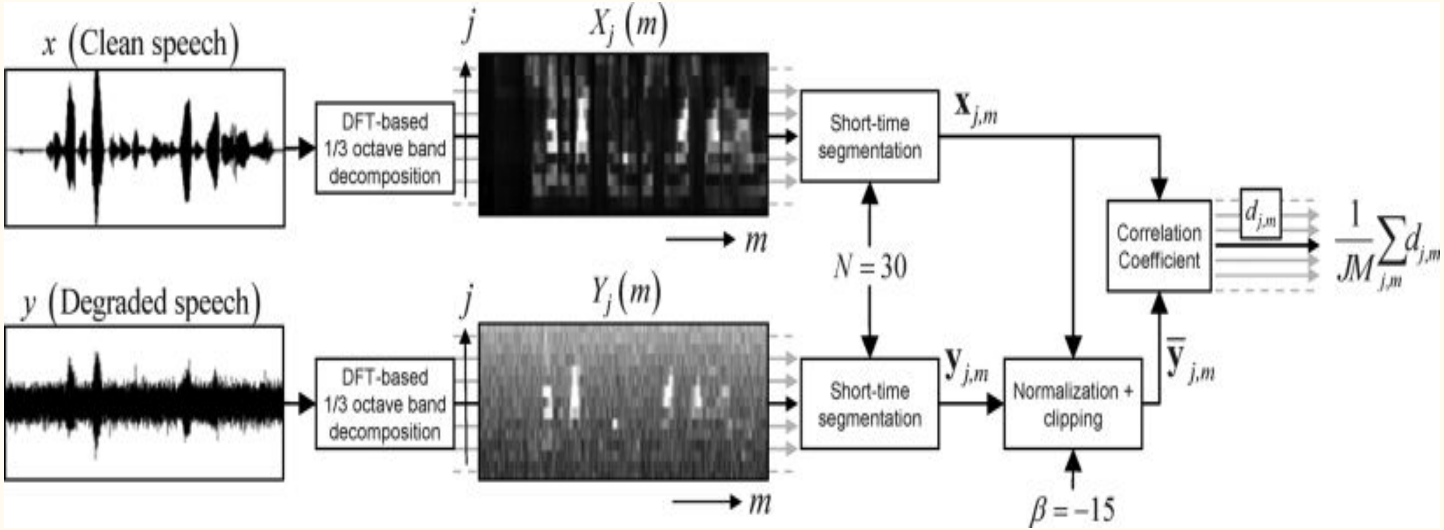
- ARU Speech Corpus (University of Liverpool)
- This corpus comprises single channel recordings of IEEE (Harvard) sentences (IEEE, 1969) spoken by twelve adult native British English speakers in anechoic conditions.
  - https://datacat.liverpool.ac.uk/681/

Noise data:

- Room Impulse Response and Noise Database
  - https://www.openslr.org/28/

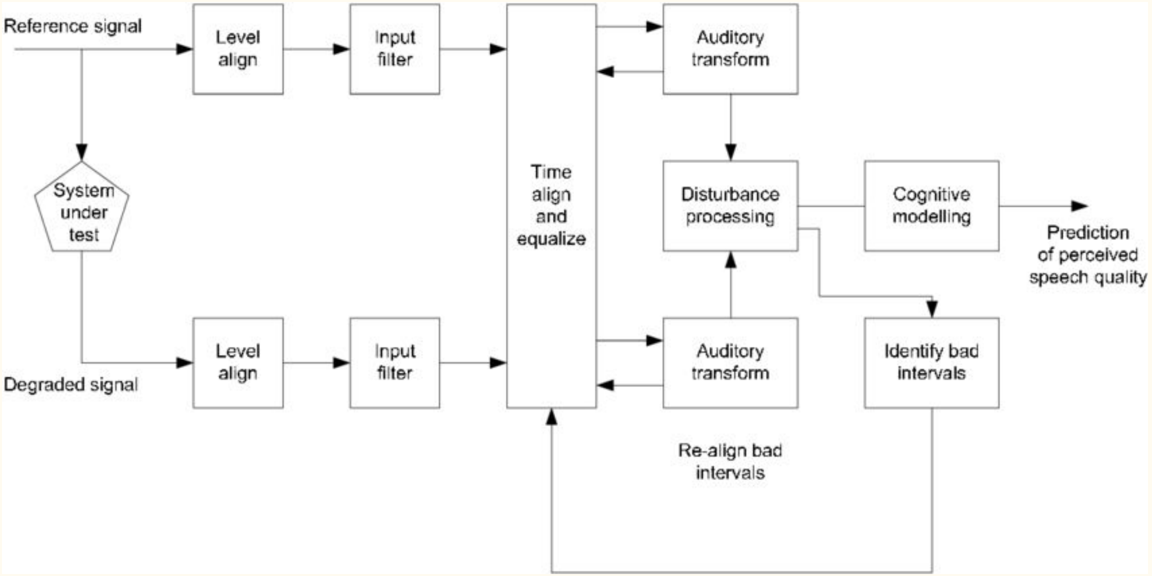
Objective Metric: STOI

- Short-time objective intelligibility measure
- Obtained by comparing the enhanced speech with the clean reference speech
- Human speech intelligibility score ranging from 0 to 1



Objective Metric: PESQ

- Perceptual evaluation of speech quality
- Calculated by comparing the enhanced speech with the clean reference speech, and it ranges from —0.5 to 4.5.



Shortcomings and advantages of the current study

Results: Objective Metric

	Noisy	logMMSE (State of the Art)	DNN
-5dB	STOI: 0.372 PESQ : 1.102	STOI: 0.286 PESQ: 1.169	STOI: 0.136 PESQ: 1.143
0dB	STOI: 0.350 PESQ : 1.081	STOI: 0.280 PESQ: 1.072	STOI: 0.130 PESQ: 1.154

Shortcomings

- It is crucial to have a large training set to learn the structure and map function between noisy and clean speech features
- The designed model faces the issue of oversmoothing, over filtering the mixed files resulting in a noisy output
- The model is very time consuming when it comes to training and running based on the architecte
- The model is only trained on the English language. Adding different languages would increase its diversity
- The dataset used utilized 72 sentences. Phonetically balanced is defined as having analysis of 100,000 words in newsprint.
- Range of Noise: We only ran the model for 0db and -5db noise

Advantages

- The model can handle large set of data
- Proposed DNN framework can handle non-stationary noises in real-world
- Pretty robust to noise, given the limited noise types used in the paper
- Includes Dropout Training to battle overfitting in DNN
  - Dropout ensures that no neuron ends up relying too much on other neurons and learns something meaningful instead
- Preproprocessing is similar to how humans decode sound (frequency-based decoding)

Future ideas and improvement suggestions

- Improving the Dataset
  - Increasing the size of our dataset would reduce our margin of error and emphasize any patterns
  - Expansion of our model for speech data from varied dialects could allow for more diverse application
- Improving the Model
  - Adopting a Gammatone filterbank would allow for better simulation model of human cochlea
  - Implementation of Multi-Resolution CochleaGram (MRCG) captures local and contextual information
  - Dynamic noise adaptation scheme would improve tracking of non-stationary noises
- Industry Application
  - Entertainment: Filtering out unwanted noise in films
  - Medical: Integration into hearing aids and cochlear implants to enhance noise reduction

Their Conclusion

- The DNN model performed better than the LogMMSE
- More acoustic information results reduced discontinuities in the enhanced speech output
- Multi-condition noise can lead to good generalization capability in unseen noise applications
- The model is effective for processing data in different languages across varying recording environments

Our Conclusion

- Our implementation of the paper's model showed that the DNN model did not perform better than the LogMMSE when implemented on our dataset
- Strange results having lower STOI and PESQ metrics for both denoising models compared to the noisy data
- This could be due to overprocessing of the noise or due to our smaller dataset
  - ~3 hours of noise vs 650 hours in the paper
  - Limited training set size due to minimal computational power

Any Questions?

Thank you!

Additional Figures

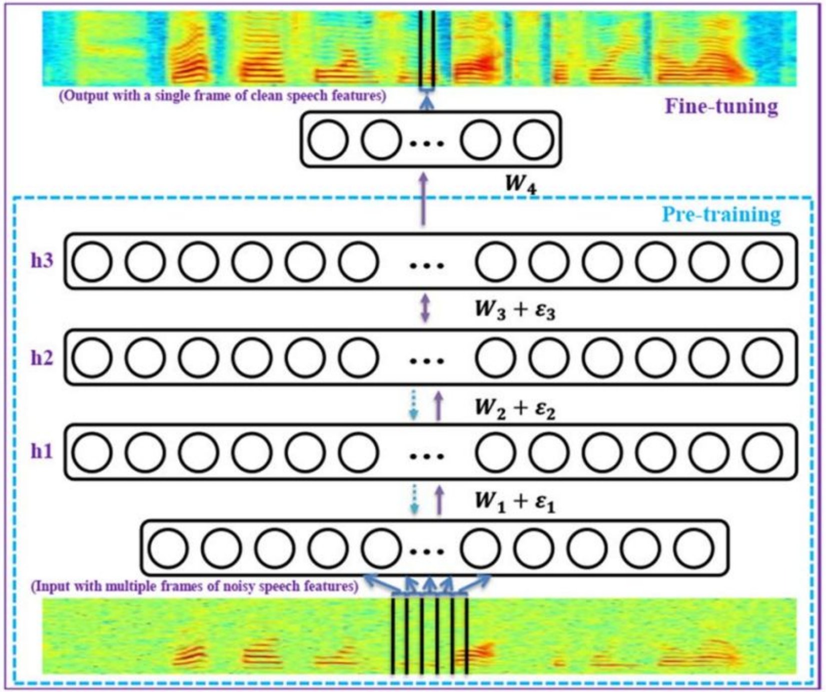
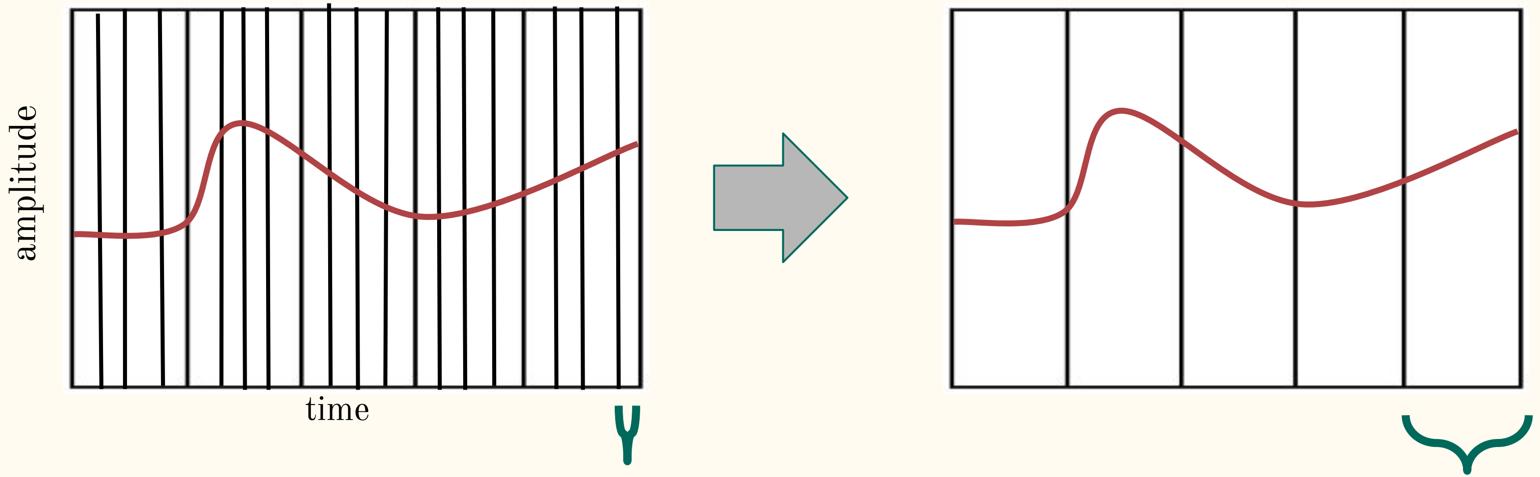


Fig. 2. Illustration of the basic DNN training procedure.

The Human Limit: Temporal Resolution

- Humans cannot detect temporal differences in sound below 3ms (hearinghealthmatters.org)
- The importance of using **frames** to process sound



Sampling rate chunk <<<<<3ms

EACH FRAME>=3ms

(32ms frames in the paper, 20-40ms is normal)