

# NEURAL NETWORKS APPLICATION TO SPEECH ENHANCEMENT

Muhammed Shifas PV



University of Crete, Dept of Computer Science  
shifaspv@csd.uoc.gr

CS-HY578: Speech Signal Processing, 9 May 2022

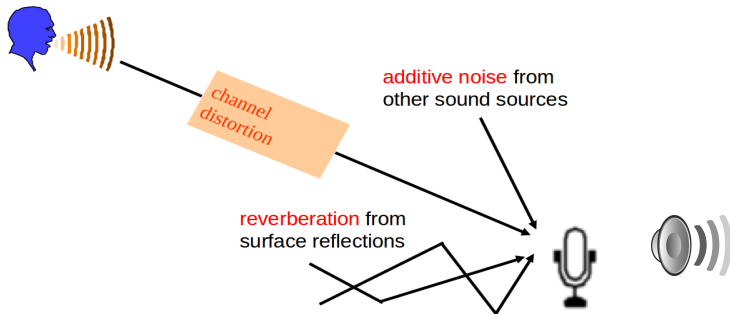
# OUTLINE

- 1 THE SPEECH DENOISING PROBLEM
- 2 NEURAL FEATURE DOMAIN MODELS
- 3 WAVEFORM DOMAIN MODELS: WAVENET AND FFTNET
- 4 CONCLUSION

# OUTLINE

- 1 THE SPEECH DENOISING PROBLEM
- 2 NEURAL FEATURE DOMAIN MODELS
- 3 WAVEFORM DOMAIN MODELS: WAVENET AND FFTNET
- 4 CONCLUSION

# SPEECH DENOISING



- **Speech Denoise:** A common terms used on dealing with the non-speech interference

# TRADITIONAL SIGNAL PROCESSING APPROACH

- The noise and speech in the mixture will vary over the time
- The intensity of noise variations will be lower compared to the speech
- Traditional Approach: Estimate the variations of the noise over time and subtract.

-Spectral Substractions

-Wiener filtering

Input:



Output:



# OUTLINE

- 1 THE SPEECH DENOISING PROBLEM
- 2 NEURAL FEATURE DOMAIN MODELS**
- 3 WAVEFORM DOMAIN MODELS: WAVENET AND FFTNET
- 4 CONCLUSION



# SHORT TIME FOURIER TRANSFORM (STFT)

**Fourier Theorem:** Any signal can be represented as sum of complex periodic signals with different harmonics.

$$x[n] = \sum_{\omega=-\infty}^{\infty} X(\omega)e^{j\omega n} \quad (1)$$

$$\text{FT}\{x[n]\} \equiv X(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \quad (2)$$

**STFT:** FT at short time instances.

$$\text{STFT}\{x[n]\} \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \quad (3)$$





# FEATURE DOMAIN MODELING

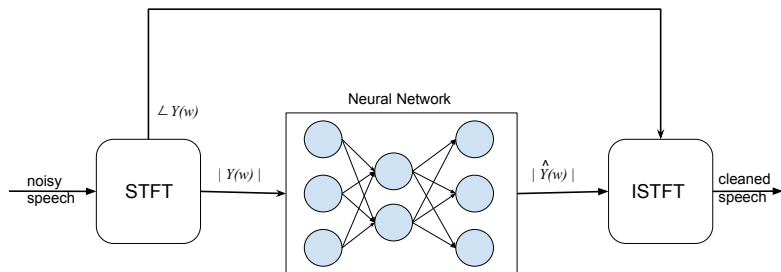
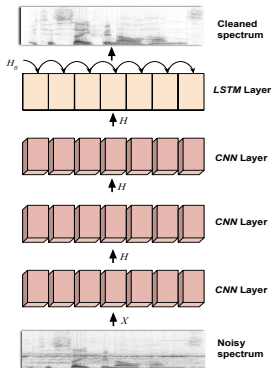


FIGURE: Feature domain SE framework

# CONVOLUTIONAL RECURRENT SE <sup>1</sup>



- Local spectral patterns are detected by convolution (CNN) layers
- Temporal dependency among frames are modelled by LSTM
- The kernel size in CNNs is customizable

FIGURE: CNN-LSTM SE architecture

<sup>1</sup>Naithani, Gaurav, et al. "Low latency sound source separation using convolutional recurrent neural networks." 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2017.

- The network has low latency; frame size processing at each instant is 5ms.
- The 160 point FFT is calculated and the magnitude of half of these points are processed: due to spectral symmetry.
- Input is the noisy magnitude spectrogram and the objective is to get the clean magnitude at output.
- The noisy phase is used for the reconstruction of the clean prediction at the output.

# MODEL LAYER DETAILS

TABLE: Number of parameters at each layer

Layer	Kernal size	Params
Convolution	[3X3]	1X[3X3]X256
Convolution	[3X3]	256X[3X3]X256
Convolution	[3X3]	256X[3X3]X256
FC-LSTM	[80X256]	[80X256]X256X 11
FC-Layer	[256X81]	[256X81]
Total		12 Million

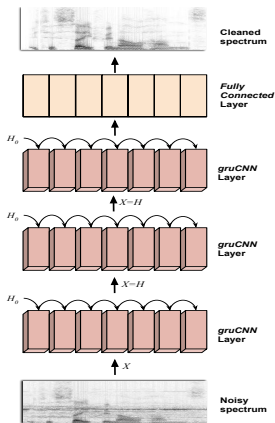
Input:



Output:



# RECURRENT CONVOLUTION SPEECH ENHANCEMENT<sup>2</sup>



- Temporal dependencies over time is being modelled at feature extraction stage
- with gruCNN cell dependencies in local patches of the spectrum can be detected
- While the model complexity is reduced up to 60%.

FIGURE: gruCNN SE

<sup>2</sup>PV. Muhammed Shifas, Santelli, C., and Stylianou, Y. (2019) Towards a Neural-Based Single Channel Speech Enhancement Model for Hearing-Aids. In ICA 2019 Proceedings, pp. 5745-5748. DOI: 10.18154/RWTH-CONV-239594

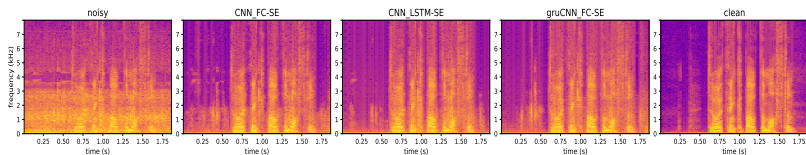
# MODEL LAYER DETAILS

TABLE: Model Parameter count

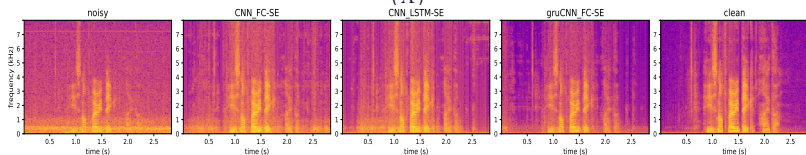
Layer	Kernal size	Params
Convolution	[3X3]	1X[3X3]X256
gruCNN	[3X3]	3X[3X3]X256
gruCNN	[3X3]	3X[3X3]X256
FC-Layer	[256X81]	[256X81]
Total		4 Million



# ENHANCEMENT IN NOISES WITH DIFFERENT SPECTRAL DISTRIBUTION



(A)



(B)

FIGURE: Model enhancement in two different noise types.

# OBJECTIVE EVALUATION SCORE

TABLE: Objective measures comparing the performance

Noise level	Metric	Noisy	CNN_FC-SE	CNN_LSTM-SE	gruCNN_FC-SE
2.5 dB	PESQ	1.20	1.41	1.51	<b>1.57</b>
	STOI	0.68	0.71	0.72	<b>0.74</b>
	COVL	1.58	1.96	2.15	<b>2.22</b>
	SSNR	- 3.63	2.39	3.20	<b>3.94</b>
12.5 dB	PESQ	1.49	1.87	2.01	<b>2.08</b>
	STOI	0.77	0.78	0.79	<b>0.80</b>
	COVL	2.11	2.59	2.74	<b>2.83</b>
	SSNR	3.24	7.61	7.85	<b>8.96</b>
22.5 dB	PESQ	2.27	2.47	2.58	<b>2.66</b>
	STOI	0.85	0.83	0.84	<b>0.85</b>
	COVL	3.05	3.20	3.30	<b>3.41</b>
	SSNR	12.26	11.21	11.14	<b>12.83</b>

# SUBJECTIVE RATING

- 0: very bad quality with very annoying artifacts
- 1: bad quality annoying artifacts
- 3: medium quality with artifacts
- 4: good quality with little artifacts
- 5: very good quality with no artifacts

TABLE: Mean opinion score (MOS) with standard error

Metric	Noisy	CNN_FC-SE	CNN_LSTM-SE	gruCNN_FC-SE	Clean
MOS	2.01±0.97	2.75±0.92	2.77±0.89	<b>3.16±0.92</b>	4.86±0.42

# ENHANCED SAMPLES

[https://www.csd.uoc.gr/~shifaspv/IEEE\\_Letter-demo](https://www.csd.uoc.gr/~shifaspv/IEEE_Letter-demo)

# OUTLINE

- 1 THE SPEECH DENOISING PROBLEM
- 2 NEURAL FEATURE DOMAIN MODELS
- 3 WAVEFORM DOMAIN MODELS: WAVENET AND FFTNET**
- 4 CONCLUSION

# SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

# SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

# SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>



## SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

## SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

## SAMPLE DOMAIN MODELS

- Neural models are developed to operate on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Residual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoder (TTS)<sup>a</sup>
- It models the dependency of a sample at  $t$  on the  $r$  previous samples as:

$$f(y_t | x_{t-1}, \dots, x_{t-r}) \quad (4)$$

- The models differ on how this dependency is achieved

<sup>a</sup><https://deepmind.com/blog/wavenet-generative-model-raw-audio/>





## MODEL DETAILS

- Causal architecture: the current sample does not depend on the future samples.
- Target is the clean samples corresponding to the noisy input signal.

### THE LOSS FUNCTION

- Loss function: Mean Absolute Error (time domain):

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} |y_t^{(k)} - \hat{y}_t^{(k)}|$$

## MODEL DETAILS

- Causal architecture: the current sample does not depend on the future samples.
- Target is the clean samples corresponding to the noisy input signal.

### THE LOSS FUNCTION

- Loss function: Mean Absolute Error (time domain):

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} |y_t^{(k)} - \hat{y}_t^{(k)}|$$

## MODEL DETAILS

- Causal architecture: the current sample does not depend on the future samples.
- Target is the clean samples corresponding to the noisy input signal.

### THE LOSS FUNCTION

- Loss function: Mean Absolute Error (time domain):

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} |y_t^{(k)} - \hat{y}_t^{(k)}|$$



## MODEL DETAILS

- Causal architecture: the current sample does not depend on the future samples.
- Target is the clean samples corresponding to the noisy input signal.

### THE LOSS FUNCTION

- Loss function: Mean Absolute Error (time domain):

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} |y_t^{(k)} - \hat{y}_t^{(k)}|$$

# DATA

- Noisy and clean files has been selected from NSDTSEA dataset<sup>3</sup>
- It consists of 20 native speakers speaking 400 different sentences
- Noisy set composed of 20 different environmental noises mixed with clean speech with different SNR points

---

<sup>3</sup>Valentini-Botinhao, Cassia. "Noisy speech database for training speech enhancement algorithms and TTS models." (2017)

# OBJECTIVE EVALUATION

Metric	Noisy	SEGAN	SE-WaveNet	SE-FFTNet
PESQ	1.96	2.24	2.23	<b>2.54</b>
LSD	1.48	1.17	1.22	<b>1.04</b>
STOI	0.28	<b>0.87</b>	0.86	<b>0.87</b>

# THE PROCESSED SAMPLES

<https://www.csd.uoc.gr/~shifaspv/IS2019-demo.html>

# REFERENCES

- Padinjaru Veettil, M.S., Santelli, C., and Stylianou, Y. (2019) Towards a Neural-Based Single Channel Speech Enhancement Model for Hearing-Aids. In ICA 2019 Proceedings, pp. 5745-5748. DOI: 10.18154/RWTH-CONV-239594
- Muhammed Shifas, P. V., Adiga, N., Tsiaras, V., and Stylianou, Y. (2019). A Non-Causal FFTNet Architecture for Speech Enhancement. Proc. Interspeech 2019, p. 1826-1830. DOI: 10.21437/Interspeech.2019-2622.

# OUTLINE

- 1 THE SPEECH DENOISING PROBLEM
- 2 NEURAL FEATURE DOMAIN MODELS
- 3 WAVEFORM DOMAIN MODELS: WAVENET AND FFTNET
- 4 CONCLUSION

# CONCLUSION

- Discussed in details the basic neural architectures
- Talked about feature domain models for speech enhancement:  
Aiming to computational constraint applications
- We have seen how to model the recurrency inside the feature extraction block.
- It has the potential to be implemented in the DSP processor for hearing aid.
- Discussed about the advanced waveform domain model:  
WaveNet and FFTNet

# CONCLUSION

- Discussed in details the basic neural architectures
- Talked about feature domain models for speech enhancement:  
Aiming to computational constraint applications
- We have seen how to model the recurrency inside the feature extraction block.
- It has the potential to be implemented in the DSP processor for hearing aid.
- Discussed about the advanced waveform domain model:  
WaveNet and FFTNet



# CONCLUSION

- Discussed in details the basic neural architectures
- Talked about feature domain models for speech enhancement:  
Aiming to computational constraint applications
- We have seen how to model the recurrency inside the feature extraction block.
- It has the potential to be implemented in the DSP processor for hearing aid.
- Discussed about the advanced waveform domain model:  
WaveNet and FFTNet

# CONCLUSION

- Discussed in details the basic neural architectures
- Talked about feature domain models for speech enhancement:  
Aiming to computational constraint applications
- We have seen how to model the recurrency inside the feature extraction block.
- It has the potential to be implemented in the DSP processor for hearing aid.
- Discussed about the advanced waveform domain model:  
WaveNet and FFTNet

# CONCLUSION

- Discussed in details the basic neural architectures
- Talked about feature domain models for speech enhancement:  
Aiming to computational constraint applications
- We have seen how to model the recurrency inside the feature extraction block.
- It has the potential to be implemented in the DSP processor for hearing aid.
- Discussed about the advanced waveform domain model:  
WaveNet and FFTNet

# Thanks for your attention