NEURAL NETWORK APPLICATIONS IN SPEECH ENHANCEMENT

ESR9: Muhammed Shifas PV

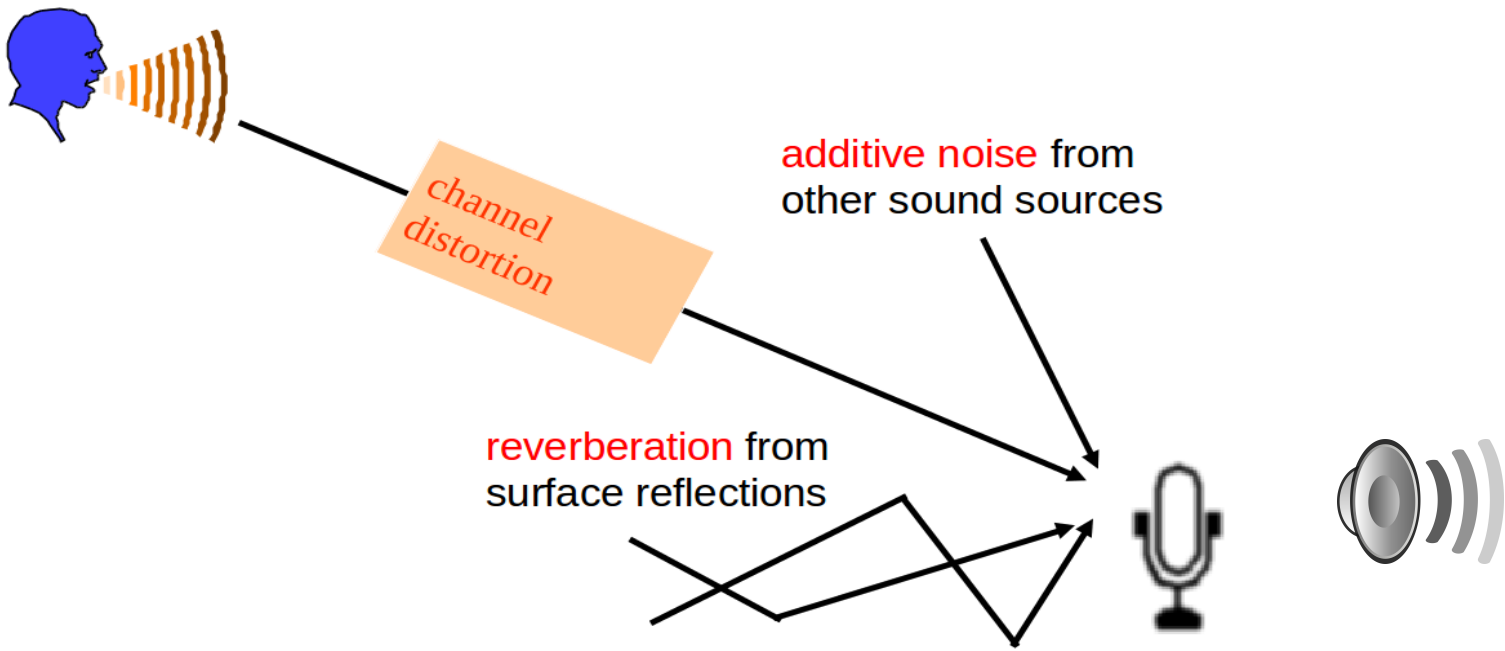
University of Crete, Dept of Computer Science
shifaspv@csd.uoc.gr

Hy578: Voice Processsing
Crete, May .13

Outline

Outline

Speech Denoising

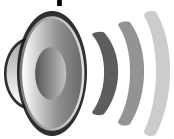


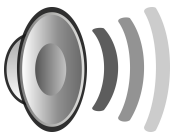
- **Speech Denoise**: A common terms used on dealing with the non-speech interference

Traditional signal processing approach

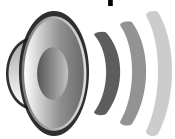
- The noise and speech in the mixture will vary over the time
- The intensity of noise variations will be lower compared to the speech
- Traditional Approach: Estimate the variations of the noise over time and subtract.
  -Spectral Substractions
  -Wiener filtering

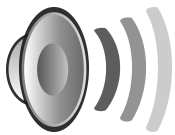
Input:





Output:





Outline

Shallow SE models for low-end devices: Secondment at Sonova (Oct-Nov 2019)

Feature domain modeling

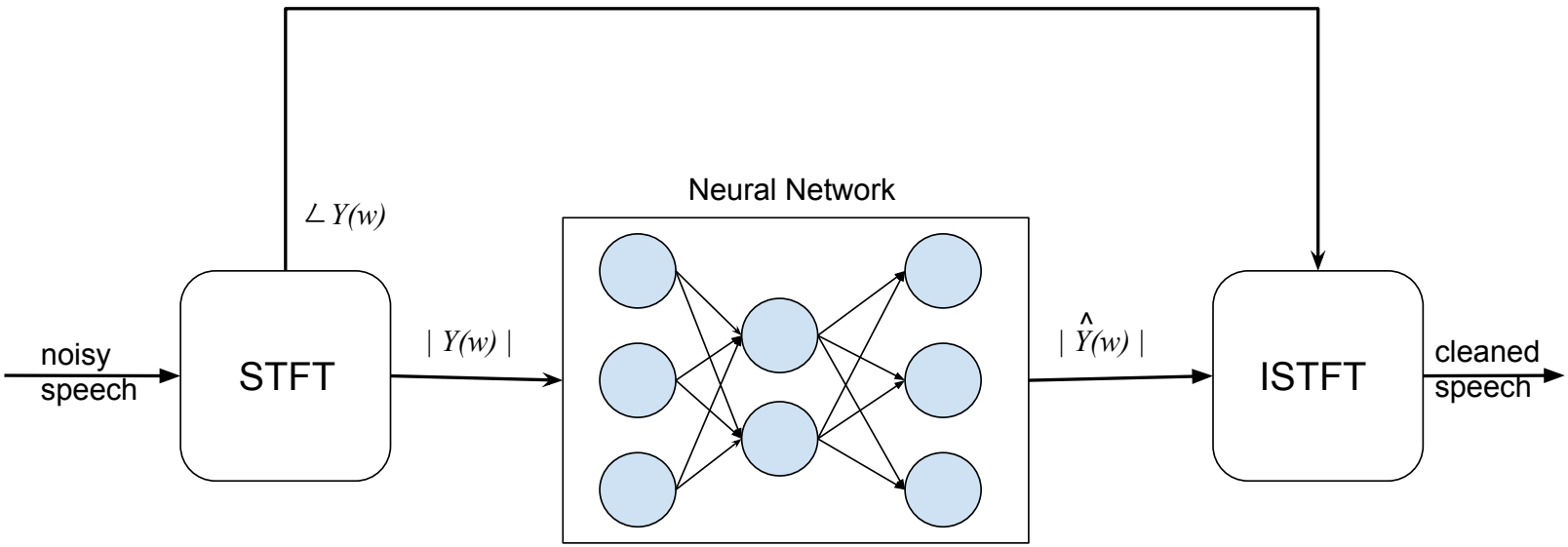
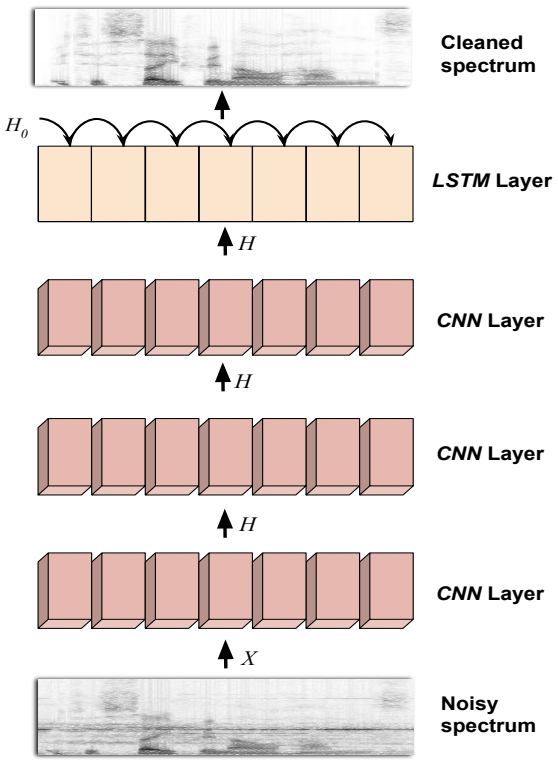


Figure: Feature domain SE framework

CONVOLUTIONAL RECURRENT SE [1]



- 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

[1] 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.

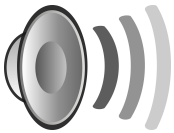
- The network has low latency; frame size processing at each instant is 5ms.
- The 160 point FFT is calculated and the magnitude of half of these points are processed: due to spectral symmetry.
- Input is the noisy magnitude spectrogram and the objective is to get the clean magnitude at output.
- The 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[2]

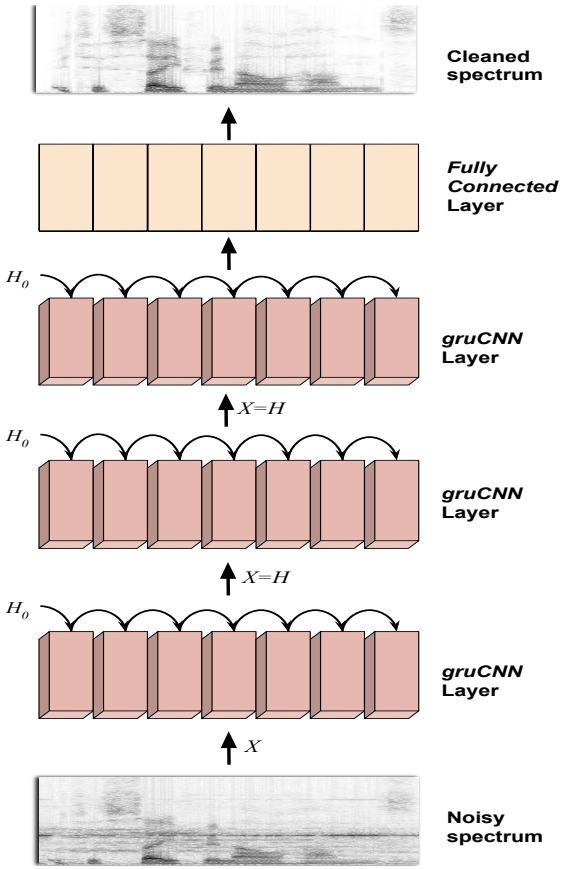


FIGURE: gruCNN SE

- 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%.

[2] 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

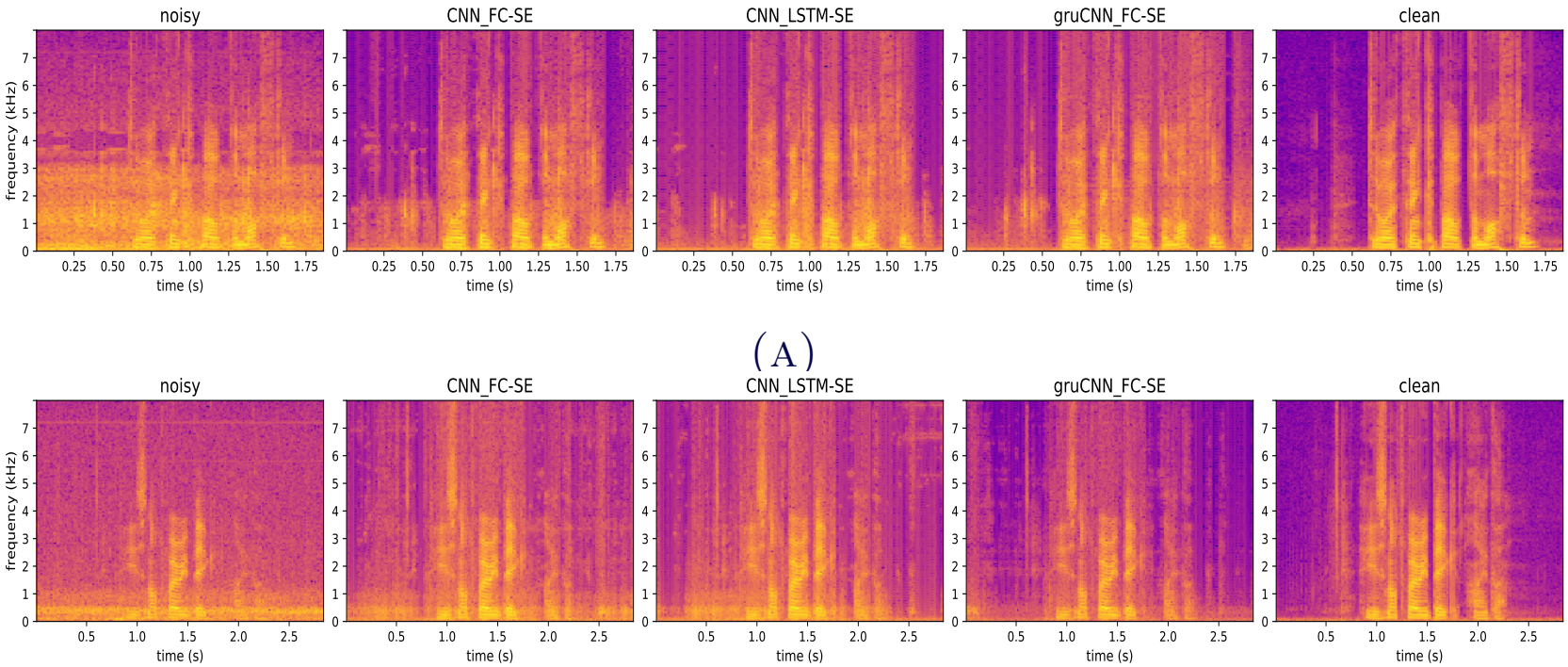


Figure: Model enhancement in two different noise types.

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	**1.57**
	STOI	0.68	0.71	0.72	**0.74**
	COVL	1.58	1.96	2.15	**2.22**
	SSNR	- 3.63	2.39	3.20	**3.94**
12.5 dB	PESQ	1.49	1.87	2.01	**2.08**
	STOI	0.77	0.78	0.79	**0.80**
	COVL	2.11	2.59	2.74	**2.83**
	SSNR	3.24	7.61	7.85	**8.96**
22.5 dB	PESQ	2.27	2.47	2.58	**2.66**
	STOI	0.85	0.83	0.84	**0.85**
	COVL	3.05	3.20	3.30	**3.41**
	SSNR	12.26	11.21	11.14	**12.83**

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	**3.16±0.92**	4.86±0.42

Enhanced samples

https://www.csd.uoc.gr/~shifaspv/IEEE_Letter-demo

Outline

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{1}$$

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{1}$$

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{1}$$

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

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

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{1}$$

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Sample Domain models

- Neural models are developed to operates on sample domain.
- It was difficult initially due to the constraints like gradient vanishing, and gradient burst.
- Recently, the Resideual Network reported to overcome the vanishing gradient
- The WaveNet and FFTNet are two sample domain models proposed as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{1}$$

- The models differ on how this dependency is achieved

[a] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

The Wavenet Architecture

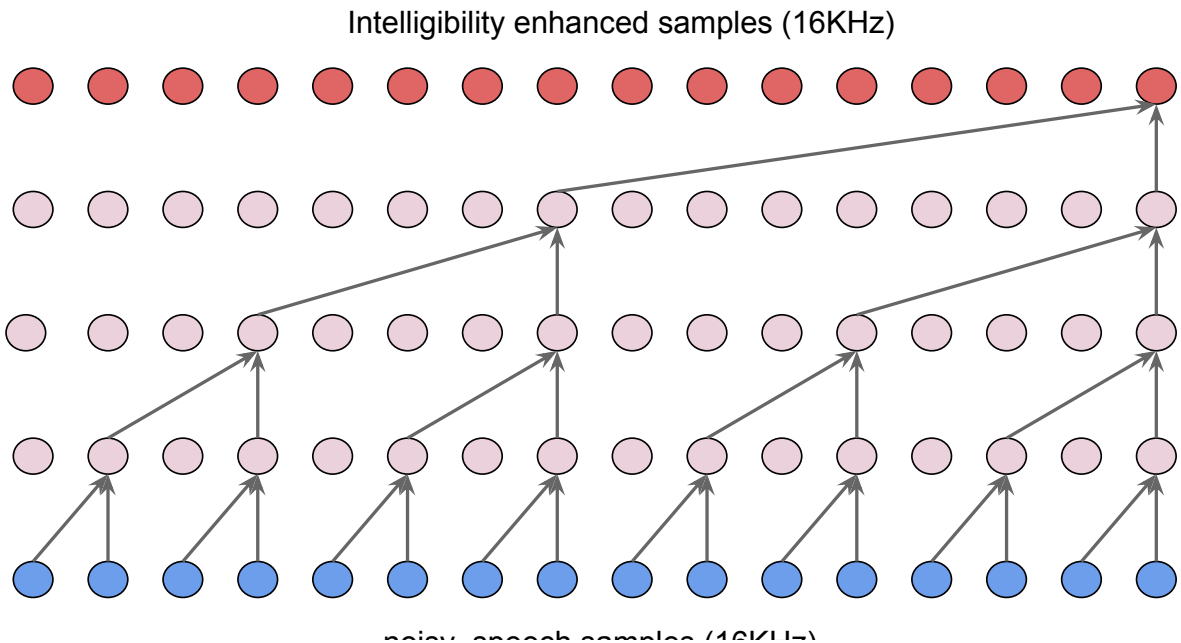


Figure: Causal Wavenet architecture

The FFTNet architecture

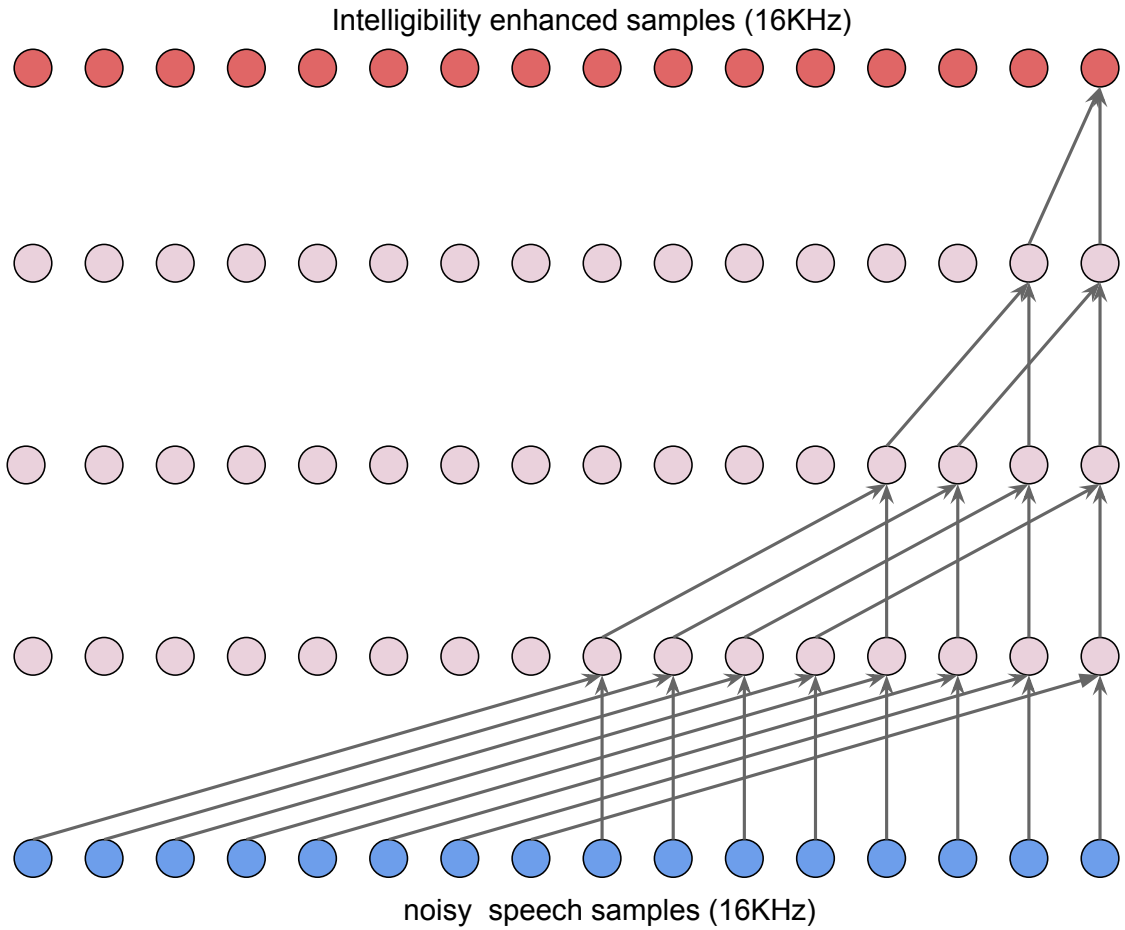


Figure: Causal FFTNet architecture

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[3]
- 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

[3] 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	**2.54**
LSD	1.48	1.17	1.22	**1.04**
STOI	0.28	**0.87**	0.86	**0.87**

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

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 wavefrom 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 wavefrom 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 wavefrom 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 wavefrom 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 wavefrom domain model: WaveNet and FFTNet

Thank for your attention!!