NEURAL NETWORK APPLICATIONS IN SPEECH ENHANCEMENT

ESR9: Muhammed Shifas PV

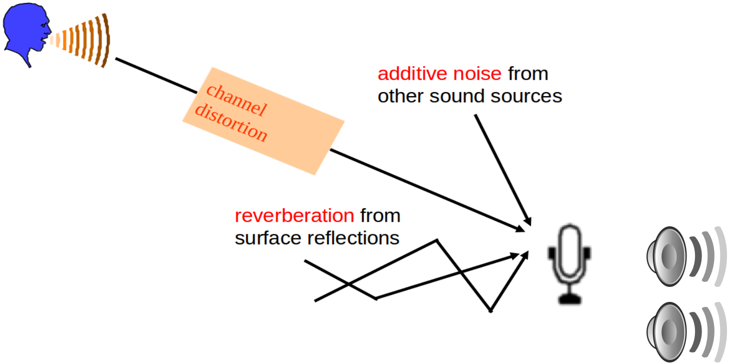
University of Crete, Dept of Computer Science
shifaspv@csd.uoc.gr

hy578: Voice Processsing
Crete, April.5

Outline

Outline

Speech Denoising



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

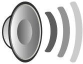
Traditional signal processing approach

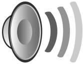
- The noise and speech in the mixture will vary over the time
- The intensity of noise variations will be lower compared to the speech
- Traditional Approach: Estimate the variations of the noise over time and subtract.

  -Spectral Substractions

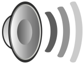
  -Wiener filtering

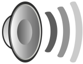
Input:





Output:





The Practical Constraints

Convolutional and LSTM SE model

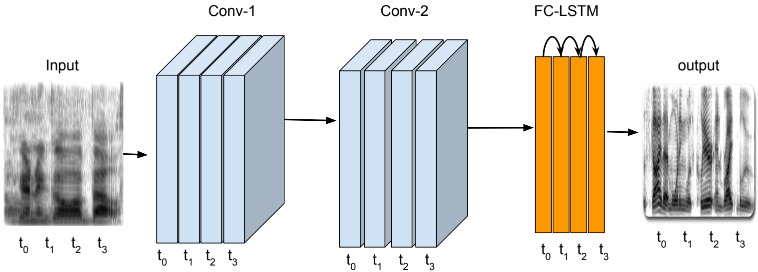


Figure: The convolutional LSTM model architecture[1]

- Temporal recurency was achieved through a Fully Connected LSTM unit followed the casual convolution

[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 instent is 5ms.
- The 160 point FFT is calculated and the magnitude of half of these points are processed: since spectral symmetry.
- The noisy phase is used for reconstruction of the clean prediction at the output.
- The input is the noisy speech spectrogram $(X(t, f))$ and objective is to get clean output

The Target: Maker Training

$$Target : M(t, f) = \frac{Y(t, f)}{Y(t, f) + N(t, f)} \tag{1}$$

Manual Post Processing:

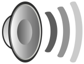
$$Y(t, f) = M(t, f) * X(t, f) \tag{2}$$

Model layer details

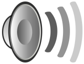
Table: Model Parameter count

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:



Proposed Recurrent Convolution SE model

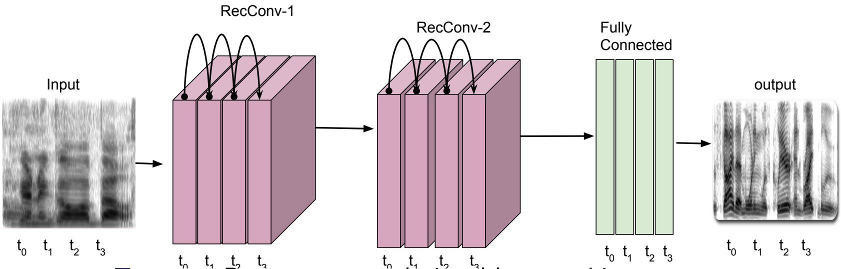


FIGURE: Recurrent convolutional layer architecture

- Temporal recurrency is being modeled while extracting the feature through the convolutional layter
- As the convolutional recurrency is less computational demanding the model is compressed upto 60%, which is benificial for applications like Hearing Aids.

Model layer details

Table: Model Parameter count

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

The Target: Maker Training

$$Target : M(t,f) = \frac{Y(t,f)}{Y(t,f) + N(t,f)} \tag{3}$$

Manual Post Processing:

$$Y(t,f) = M(t,f) * X(t,f) \tag{4}$$

The processed samples

Input:

Output:

Outline

1. The Speech denoising Task
2. Waveform domain models: WaveNet and FFTNet
3. Conclusion

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t | x_{t-1}, \ldots, x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t | x_{t-1}, \ldots, x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t | x_{t-1}, \ldots, x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1}, \ldots, x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t|x_{t-1},\ldots,x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

Sample Domain models

- Neural models are grown up to operates in the sample domain.
- It was not pissible initially since the implementational constraints like gradient vanishing
- Now the Resideual Network baypass the vanishing gradient
- WaveNet and FFTNet are the existing sample domain models as Vocoders (TTS)[a]
- It models the dependency of a sample at t on the r previous samples as:

$$f(y_t | x_{t-1}, \ldots, x_{t-r}) \tag{5}$$

- This conditional dependency is being achieved by different architecture for WaveNet and FFTNet

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

The Wavenet Architecture

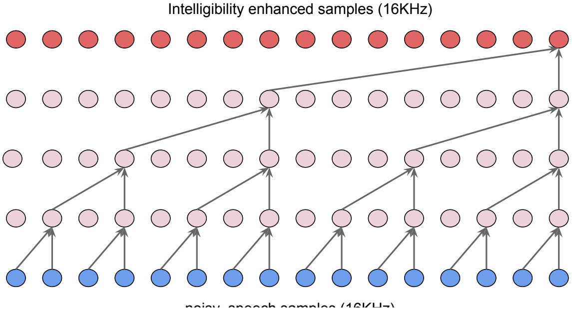


Figure: Causal Wavenet architecture

The FFTNet architecture

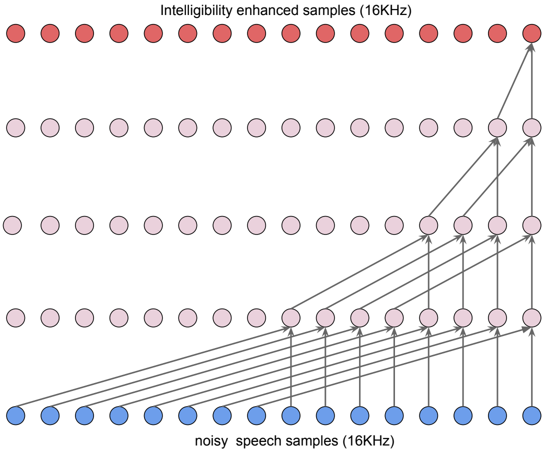


Figure: Causal FFTNet architecture

MODEL DETAILS

- Causal architecture: current sample is generated by considering past dependencies
- Target is the intelliginility enhanced samples (SSDRC modfied) of the clean speech corresponding to a noisy 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: current sample is generated by considering past dependencies
- Target is the intelliginility enhanced samples (SSDRC modfied) of the clean speech corresponding to a noisy 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: current sample is generated by considering past dependencies
- Target is the intelliginility enhanced samples (SSDRC modfied) of the clean speech corresponding to a noisy 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: current sample is generated by considering past dependencies
- Target is the intelliginility enhanced samples (SSDRC modfied) of the clean speech corresponding to a noisy 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[2]
- 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

[2] Valentini-Botinhao, Cassia. "Noisy speech database for training speech enhancement algorithms and TTS models." (2017)

Intelligibility in SSN & SWN

Outline

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Conclusion

Part.I: Speech Denoising Models

- Peroposed a recurrent convolutional architecture for speech denoising
- The model parameters being reduced considerably (50-60%) compared to traditional method while maintaining the perfromance
- It has the potential to be implemented in the DSP processor for hearing aid.

Part.II: Speech intelligibility enhancer

- Proposed two neural network models (WaveNet & FFTNet) for the task of real-life speech (noisy speech) intelligibility enhancement.
- The FFTNet performes better than WaveNet model in task.
- The FFTNet is outperfromed the traditional WBSSDRC

Thank You