Real-time Single-channel Speech Enhancement with Recurrent Neural Networks

Yangyang (Raymond) Xia
Mentored by Sebastian Braun
MSR Audio and Acoustics Research Group

Recording has started.
Participation in the meeting indicates your consent to being included in the meeting recording.

Outline

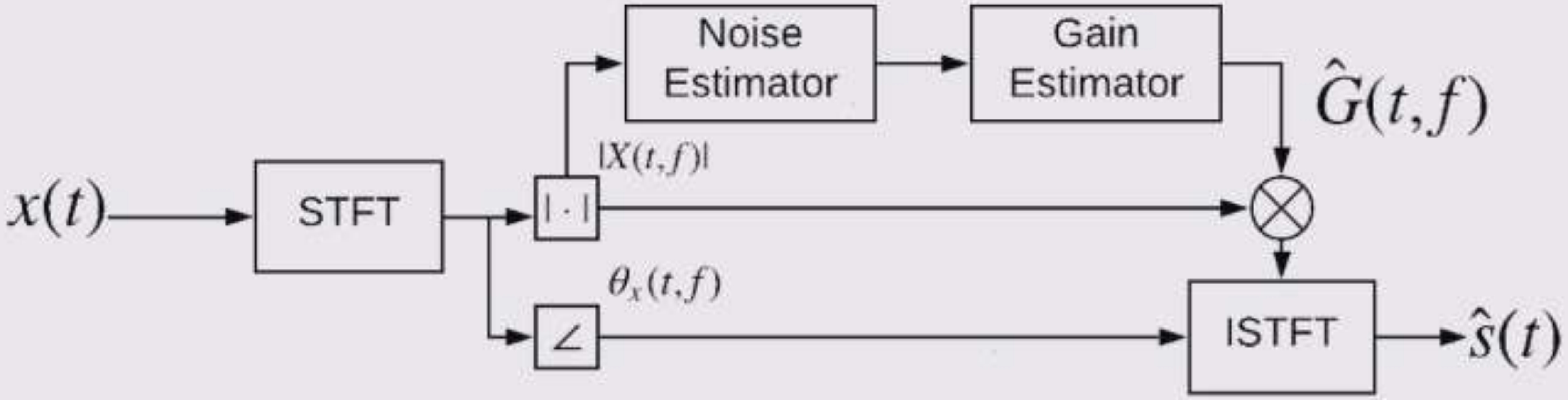
- Introduction to Single-channel Speech Enhancement
  - Classical signal processing vs. Deep learning
  - Considerations for online processing
- Our Method
  - Feature Representations
  - Learning Machines
  - Learning Objectives
  - Training Considerations
- Evaluation
  - Data
  - Metrics
  - Results
- Findings and Conclusions

Single-channel Speech Enhancement (SE)

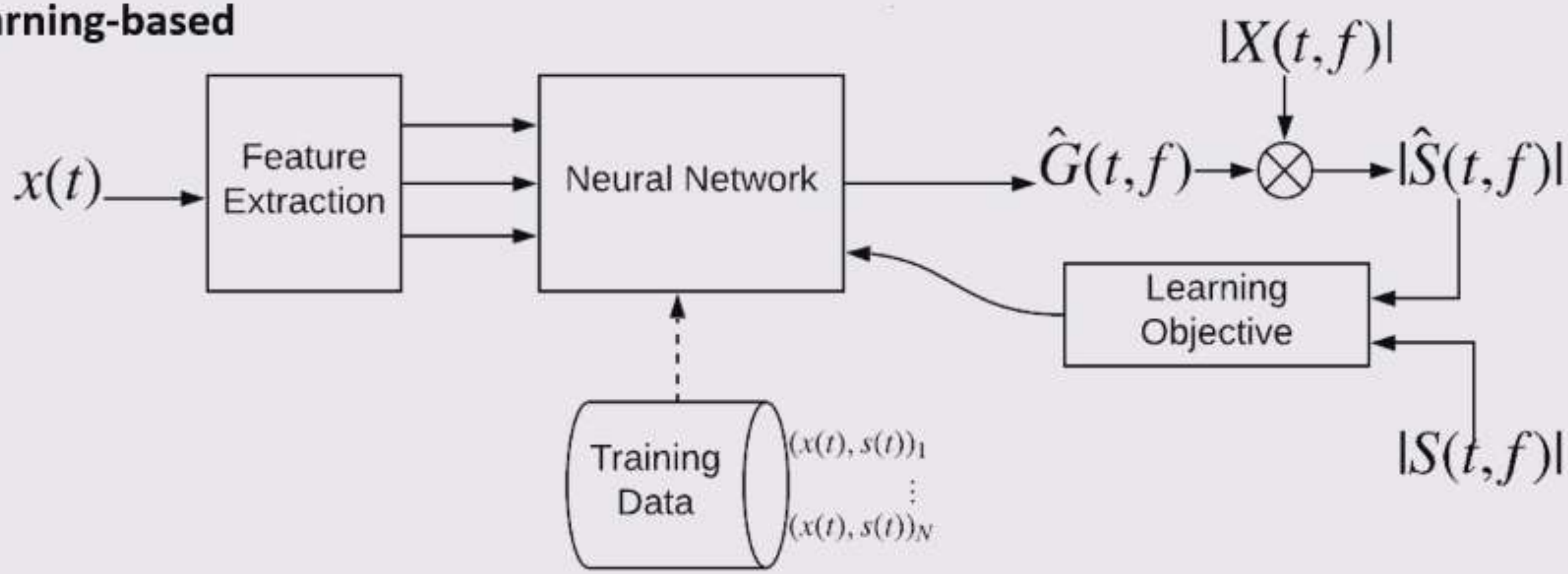
- Assumptions
  - Noisy speech = Speech + Noise
  - Noise attributes change slower than speech
- Goals
  - Suppress noise
  - Retain speech
  - Improve human or/and machine perception
- **Our goal: enhancing speech quality for human listeners**

Generic SE Pipelines

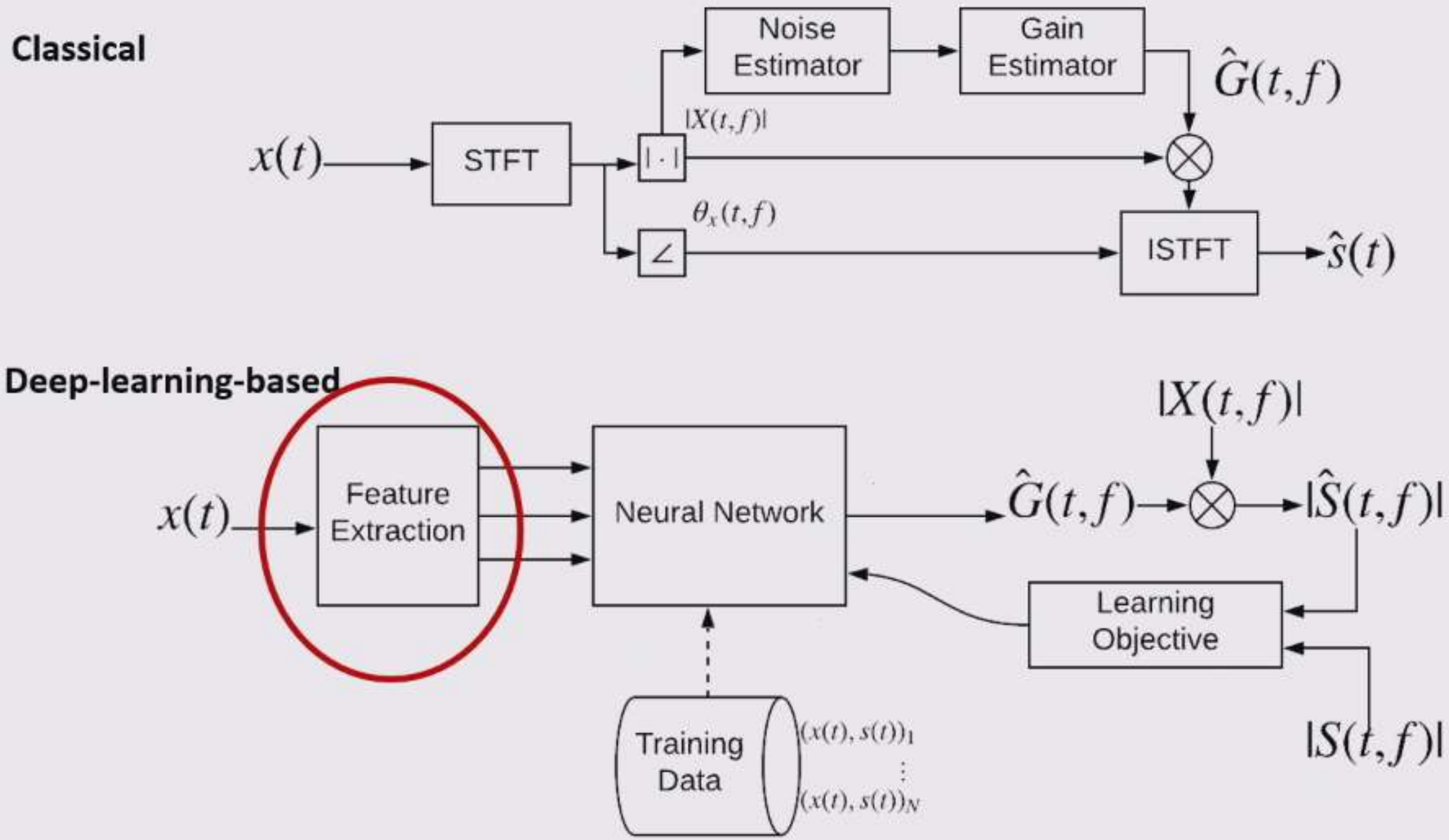
Classical



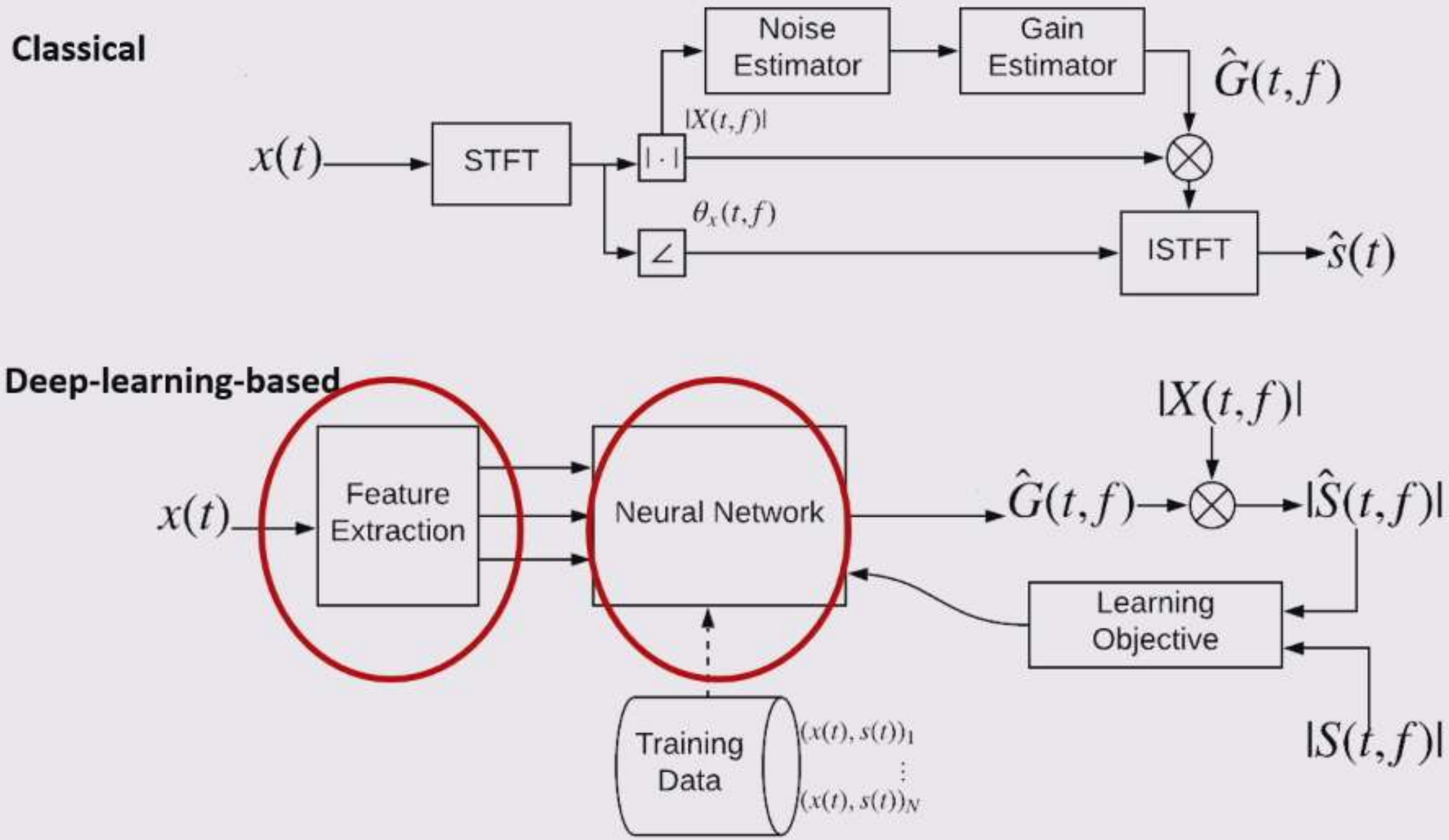
Deep-learning-based



Generic SE Pipelines

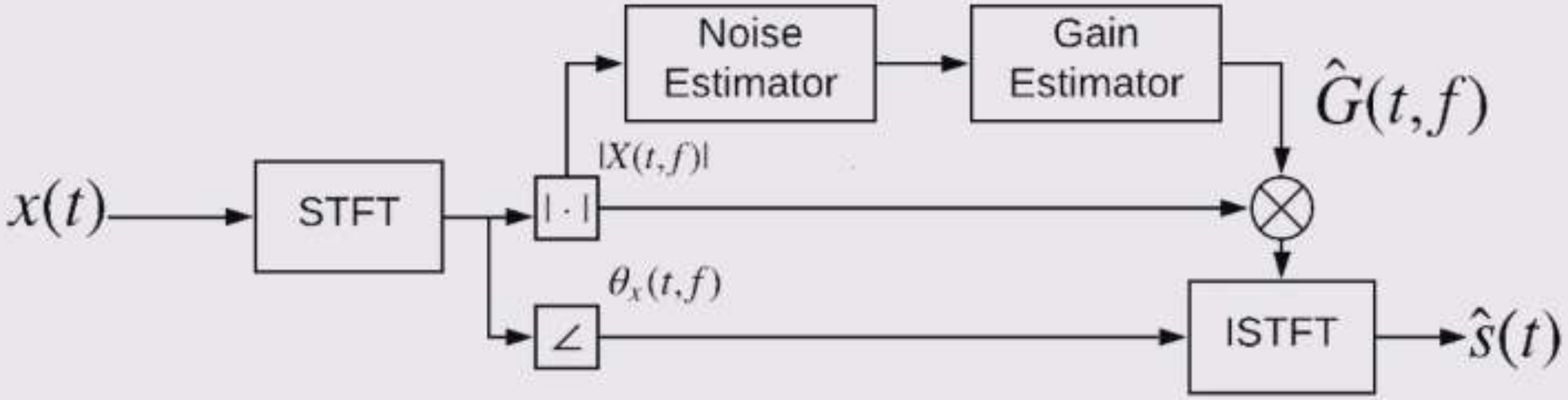


Generic SE Pipelines

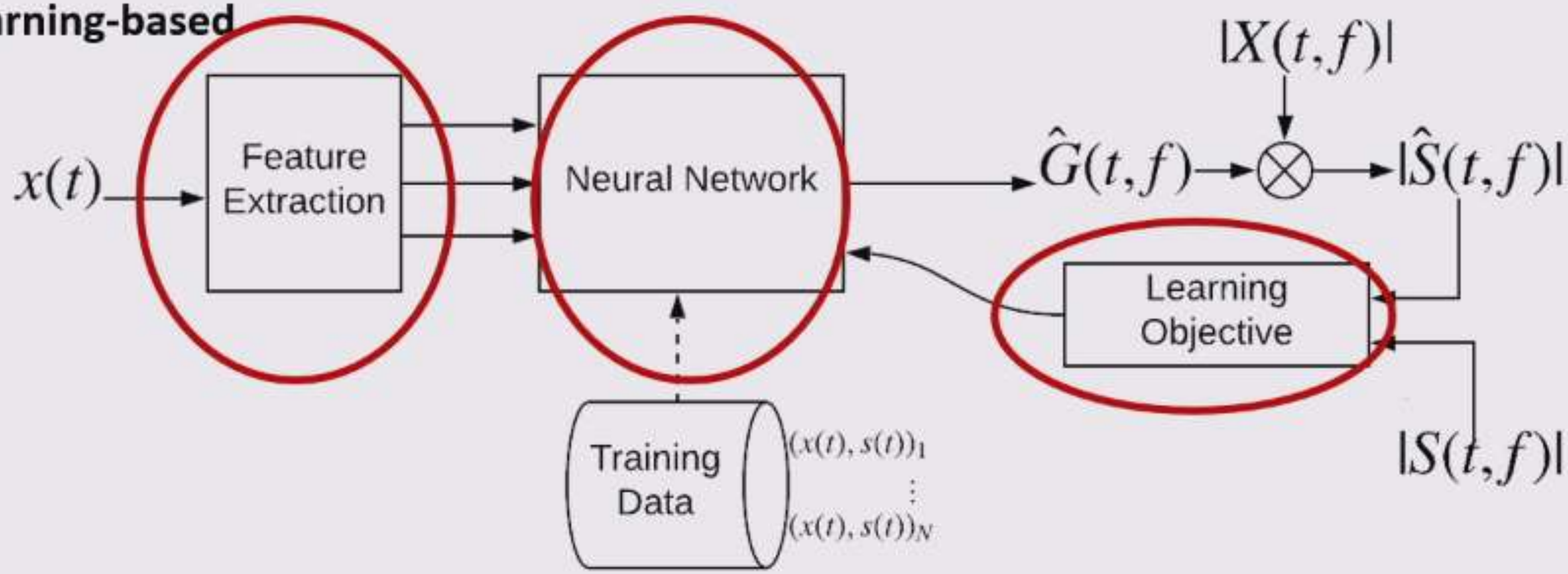


Generic SE Pipelines

Classical

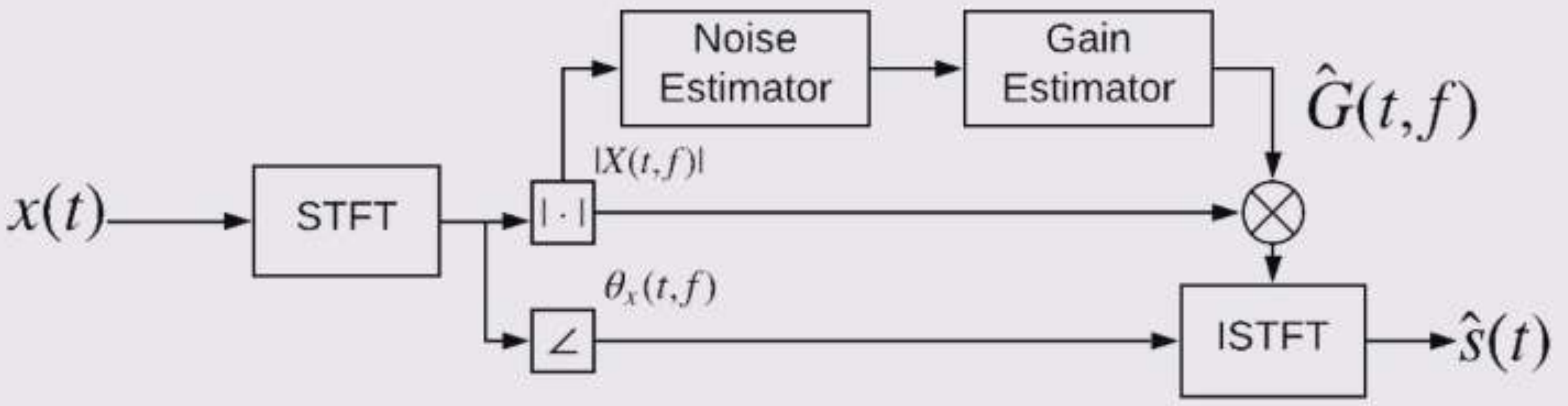


Deep-learning-based

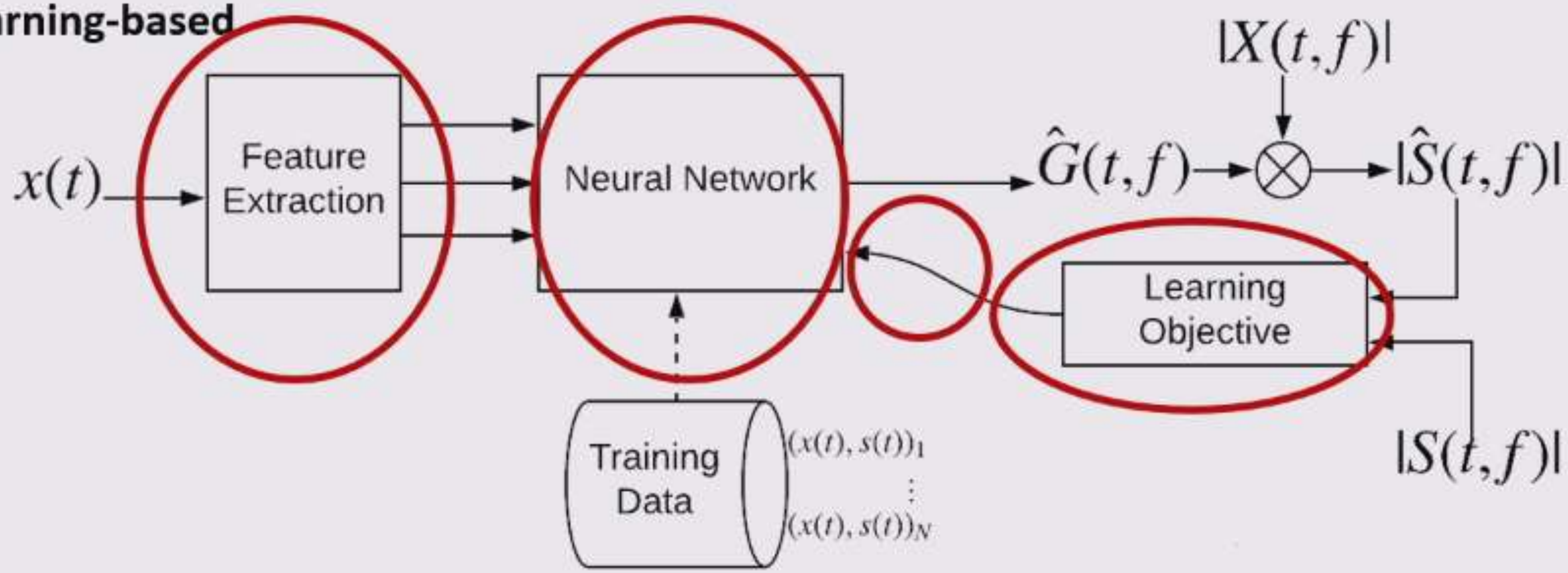


Generic SE Pipelines

Classical



Deep-learning-based



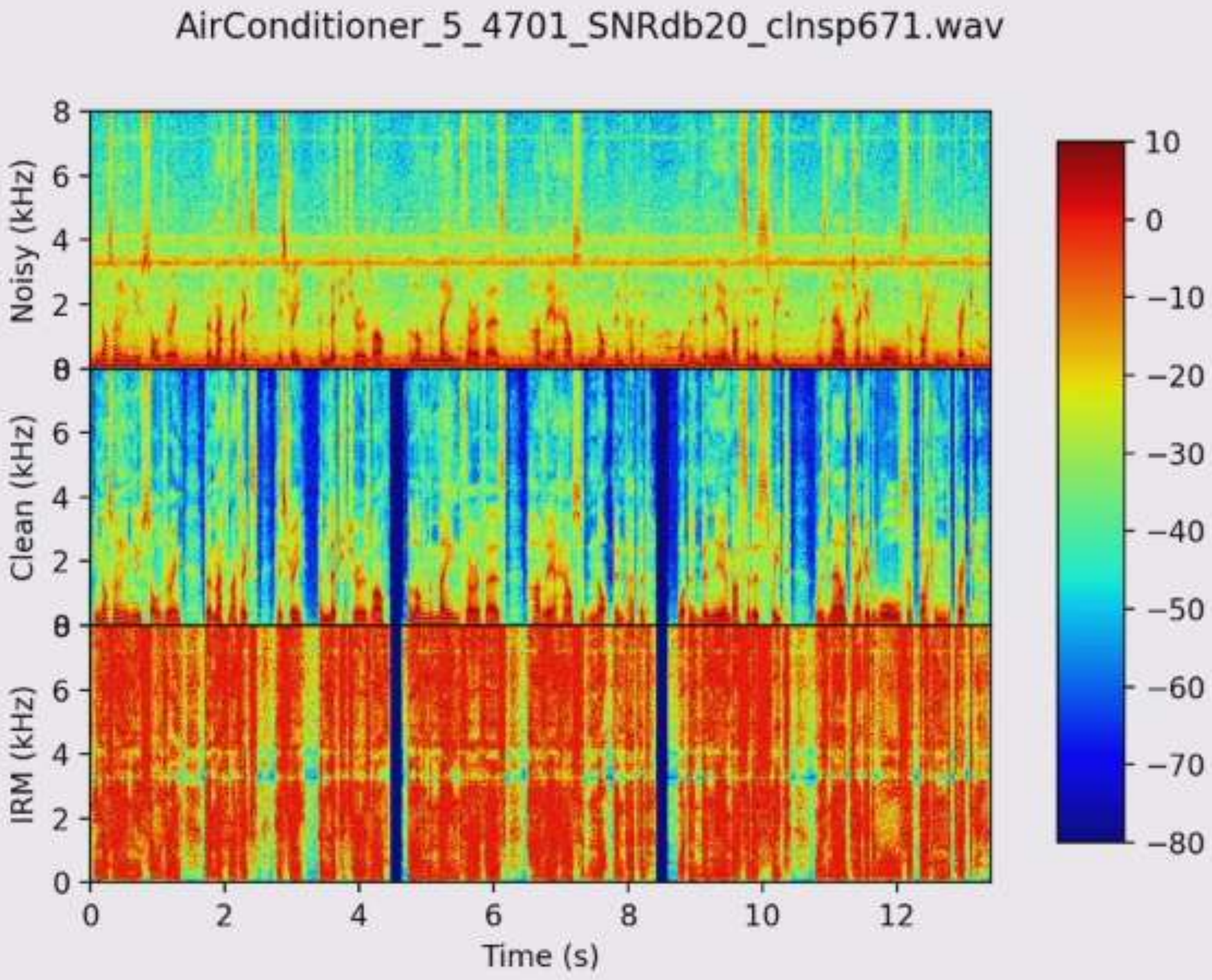
SP vs. DL for Online Enhancement

Name	Method related to Online Processing	Data-driven?	Online?
Spectral subtraction [Boll1979]	Estimate noise magnitude spectra by a **moving average** filter	No	Yes
Decision-directed [Ephraim1984]	Estimate SNRs by **smoothing instantaneous measurements** of SNRs	No	Yes
Deep clustering [Hershey2016]	Cluster each time-frequency bin based on feature embeddings generated from a **100-frame** spectrograms	Yes	No
Audio-visual speech separation [Ephrat2018]	**2-D convolution cross time and frequency** of a spectrogram	Yes	No
RNNoise [Valin2018]	Recurrent units **output one frame from one input frame**	Yes	Yes
SEGAN [Pascual2017]	Generate enhanced speech from a waveform segment	Yes	Maybe, if trained on a single frame.

Outline

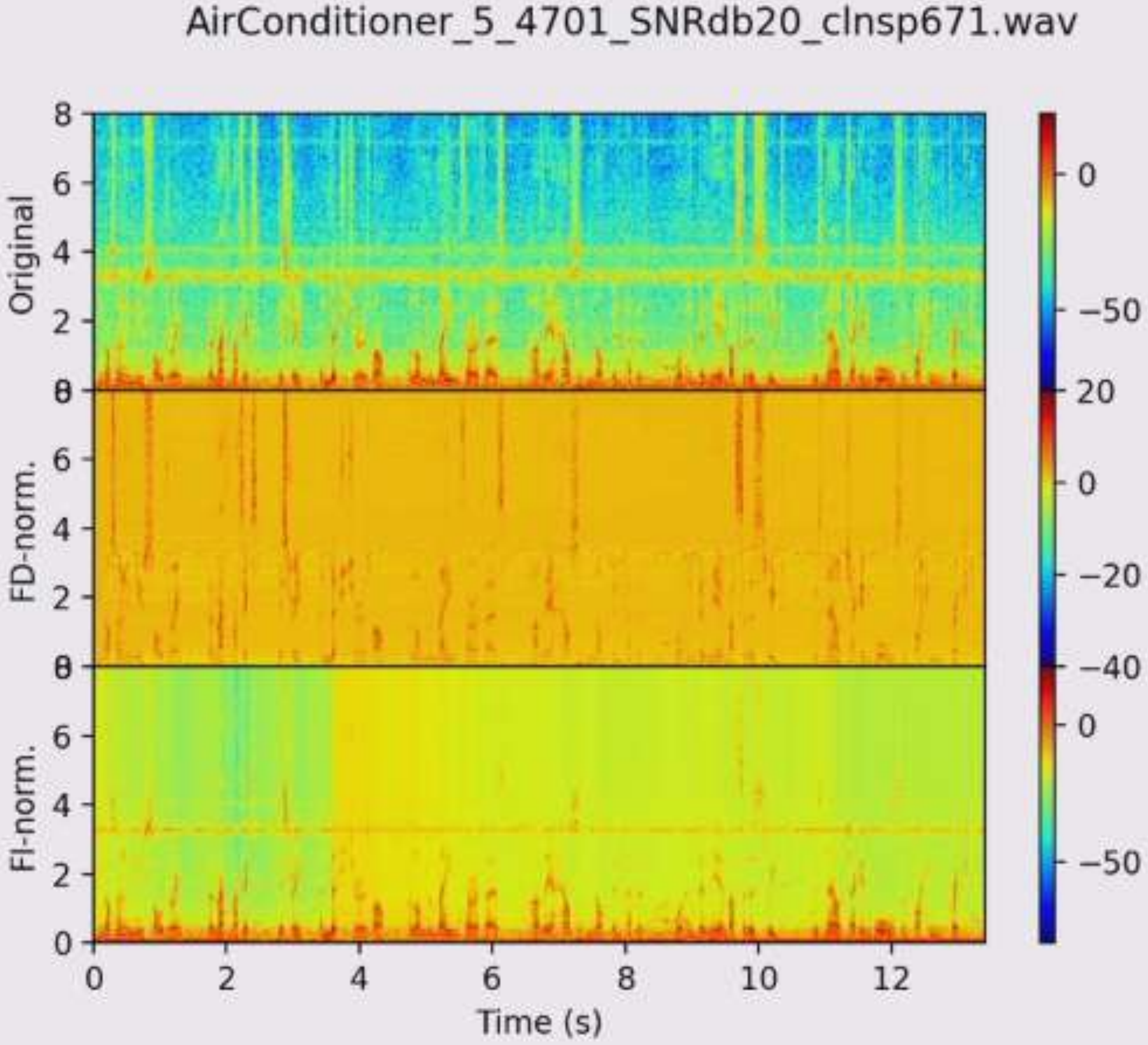
- Introduction to Single-channel Speech Enhancement
  - Classical signal processing vs. Deep learning
  - Considerations for online processing
- **Our Method**
  - Feature Representations
  - Learning Machines
  - Learning Objectives
  - Training Considerations
- Evaluation
  - Data
  - Metrics
  - Results
- Findings and Conclusions

Input (Feature) & Output (Target)



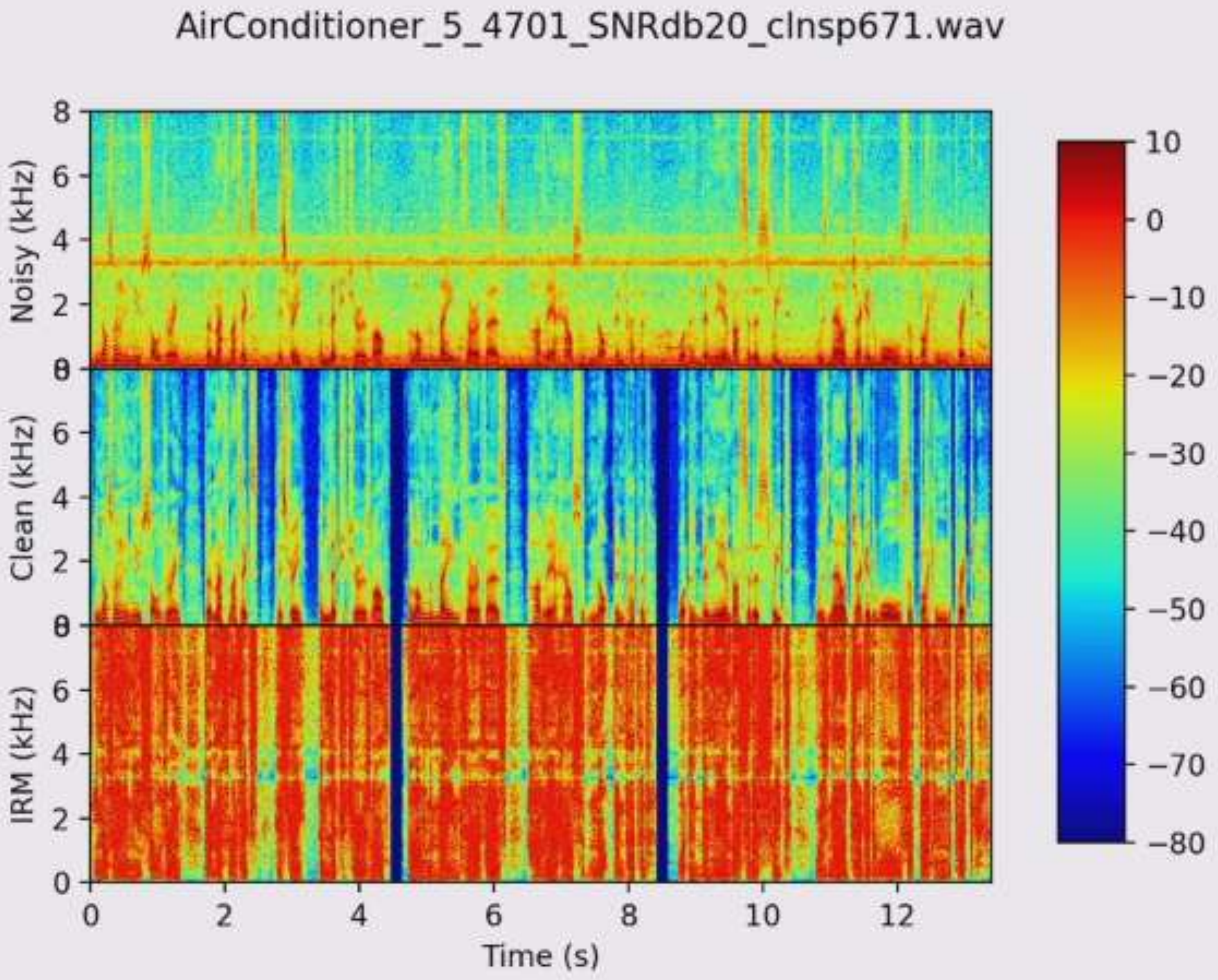
- (In) Short-time Fourier transform magnitude (STFTM)
- (In) Short-time log power spectra (LPS) with -80 dB floor
- (Out) Real magnitude gain function in [0, 1]
- Technical details:
  - 16 KHz sampling rate
  - 32-ms analysis frame
  - Hamming window
  - 75% overlap between frames

Input (Feature) & Output (Target)



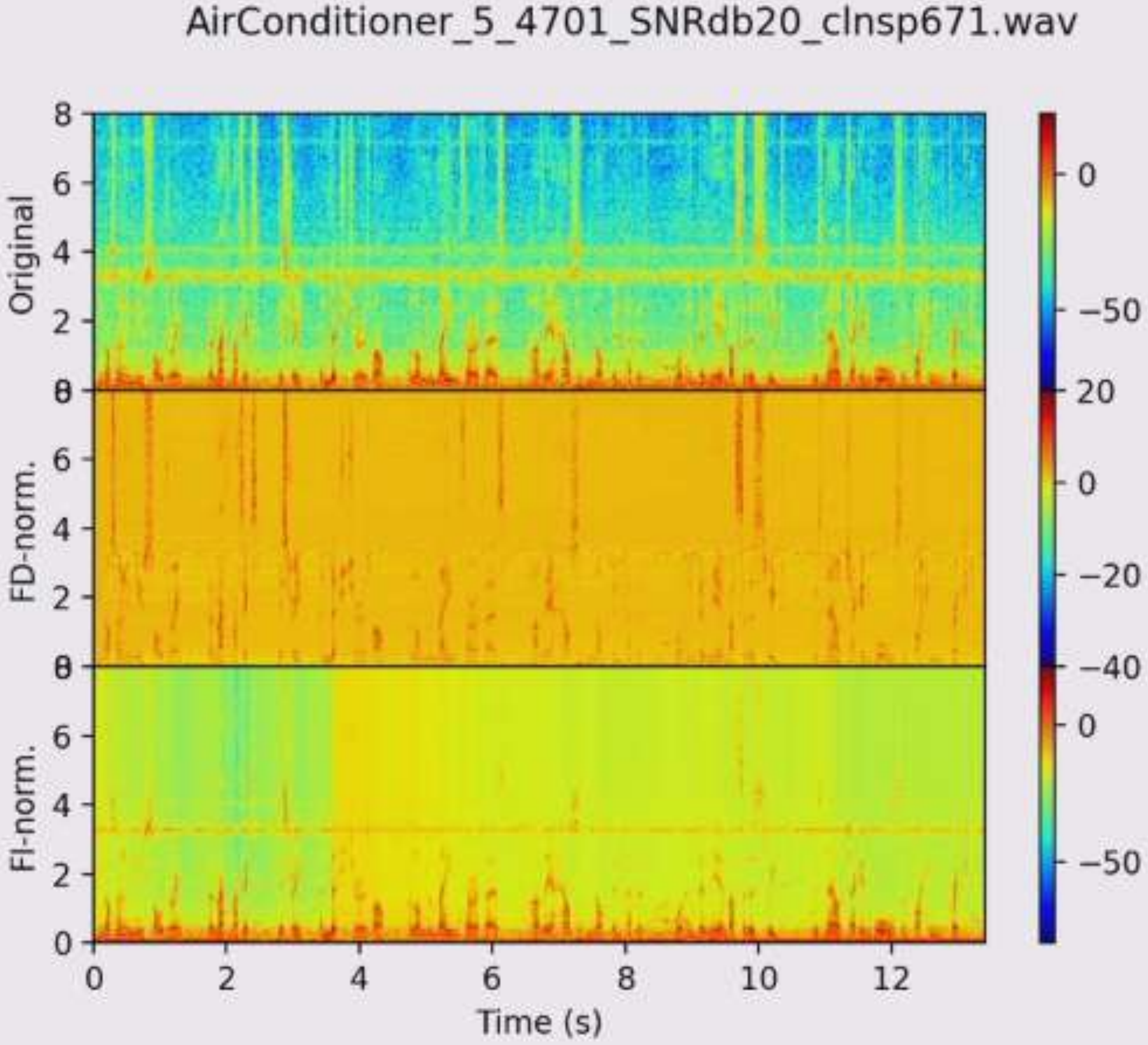
- Global mean and variance normalization
  - Statistics per frequency bin accumulated over 80 hours of randomly sampled speech from the training set
- Online mean and variance normalization
  - 3-second exponential smoothing
  - Frequency-dependent (FD) or frequency-independent (FI)

Input (Feature) & Output (Target)



- (In) Short-time Fourier transform magnitude (STFTM)
- (In) Short-time log power spectra (LPS) with -80 dB floor
- (Out) Real magnitude gain function in [0, 1]
- Technical details:
  - 16 KHz sampling rate
  - 32-ms analysis frame
  - Hamming window
  - 75% overlap between frames

Input (Feature) & Output (Target)



- Global mean and variance normalization
  - Statistics per frequency bin accumulated over 80 hours of randomly sampled speech from the training set
- Online mean and variance normalization
  - 3-second exponential smoothing
  - Frequency-dependent (FD) or frequency-independent (FI)

Learning Machines

- Recurrent neural network (RNN) the most "natural" choice
  - Ability to encode long-term temporal patterns
  - Information exchange across frequencies
- Example: RNNoise [Valin2018]
  - GRUs [Bahdanau2014] encode temporal patterns
  - Full-connected (dense) layers transform composite features to a gain

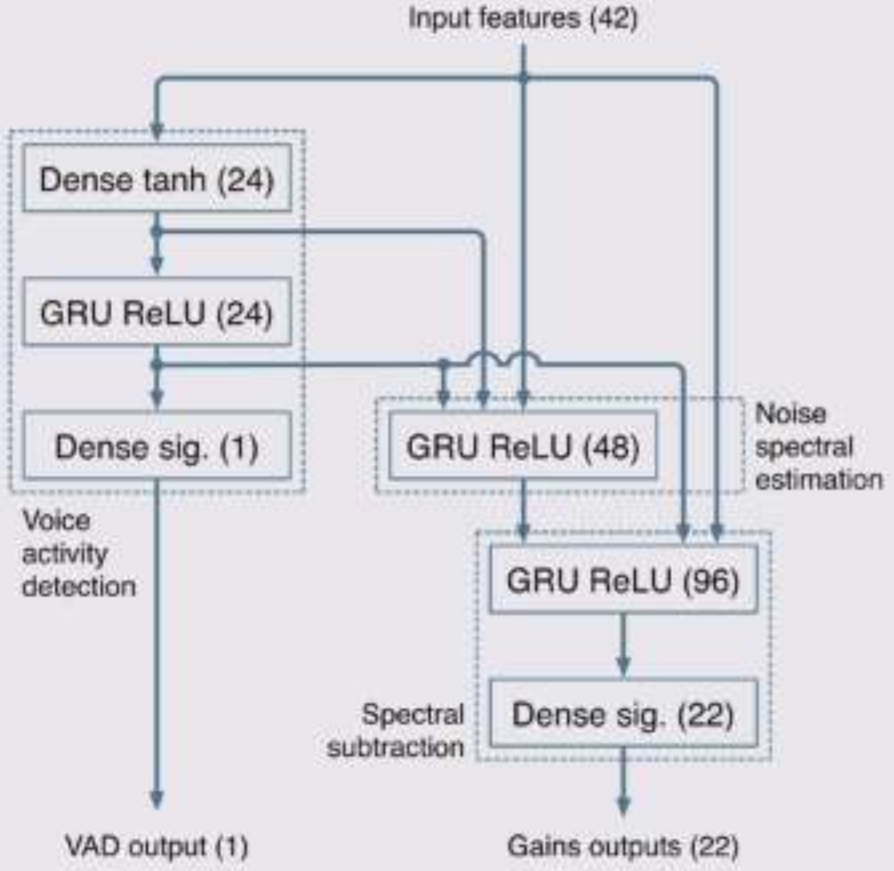


Image credit: [Valin2018]

Recurrent Units with Residual Connections

- Residual connections facilitate learning deep networks [He2016]
  - Depth = sequence length in our context
- Existing work using RNN + residuals
  - Sequence classification [Wang2016]
  - Automatic speech recognition [Kim2017]
  - Feature compensation for ASR [Chen2017]

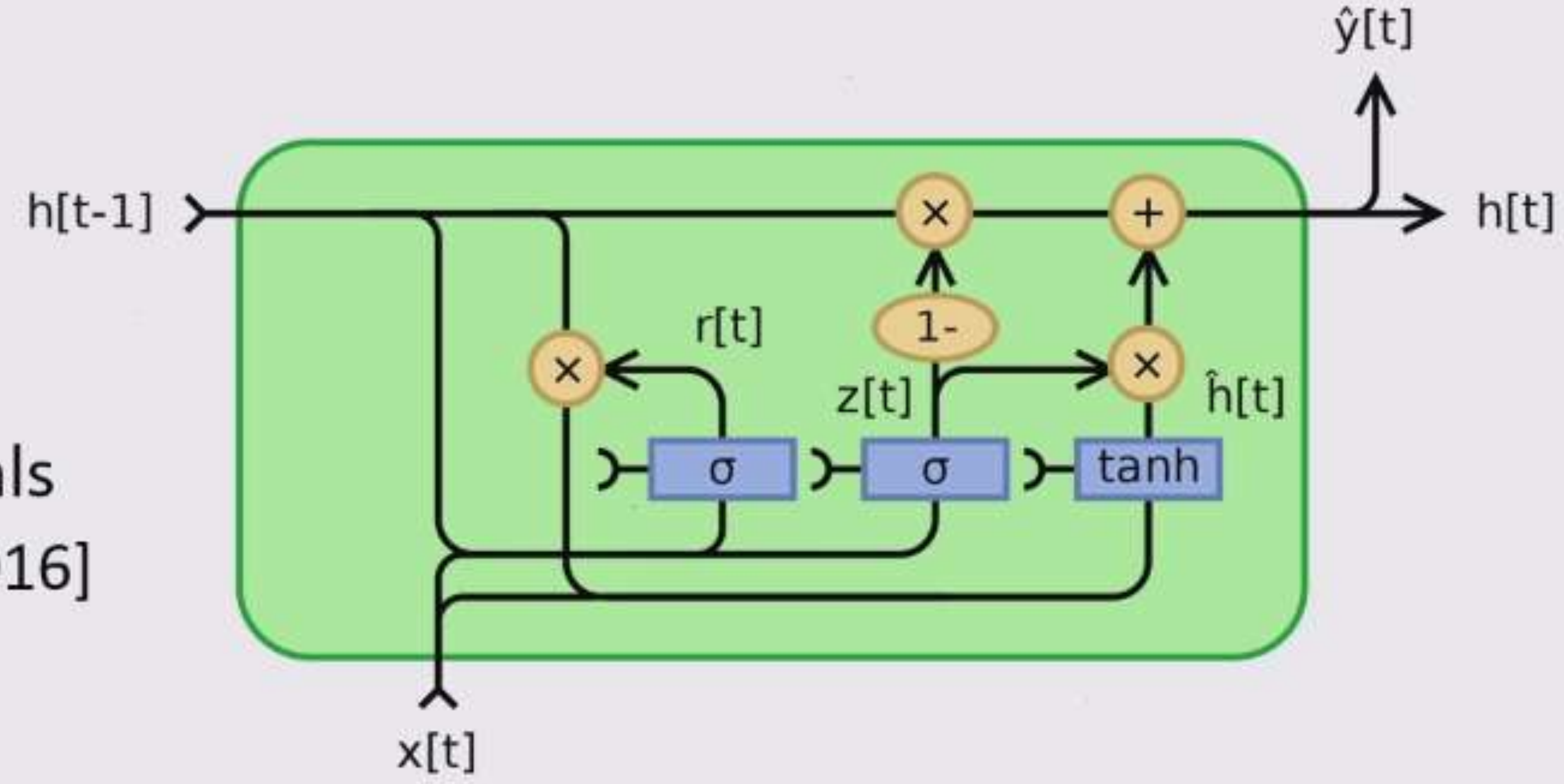


Image credit: https://en.wikipedia.org/wiki/Gated_recurrent_unit#/media/File:Gated_Recurrent_Unit,_base_type.svg. The original website is down.

Learning Machines

- Recurrent neural network (RNN) the most "natural" choice
  - Ability to encode long-term temporal patterns
  - Information exchange across frequencies
- Example: RNNoise [Valin2018]
  - GRUs [Bahdanau2014] encode temporal patterns
  - Full-connected (dense) layers transform composite features to a gain

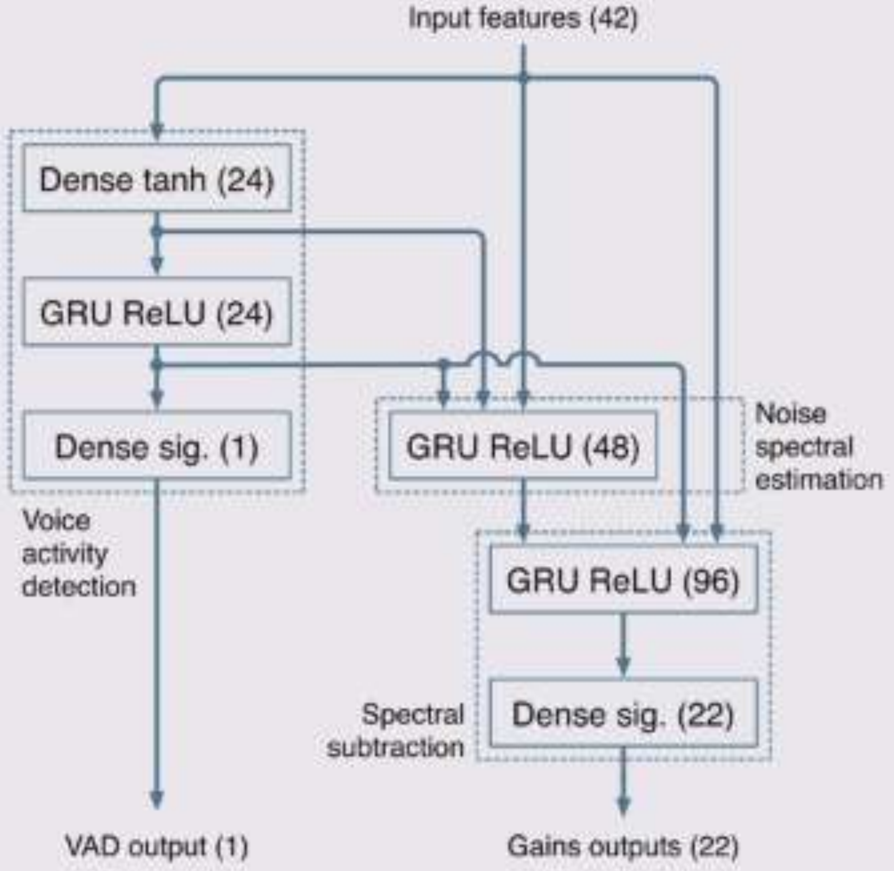


Image credit: [Valin2018]

Recurrent Units with Residual Connections

- Residual connections facilitate learning deep networks [He2016]
  - Depth = sequence length in our context
- Existing work using RNN + residuals
  - Sequence classification [Wang2016]
  - Automatic speech recognition [Kim2017]
  - Feature compensation for ASR [Chen2017]

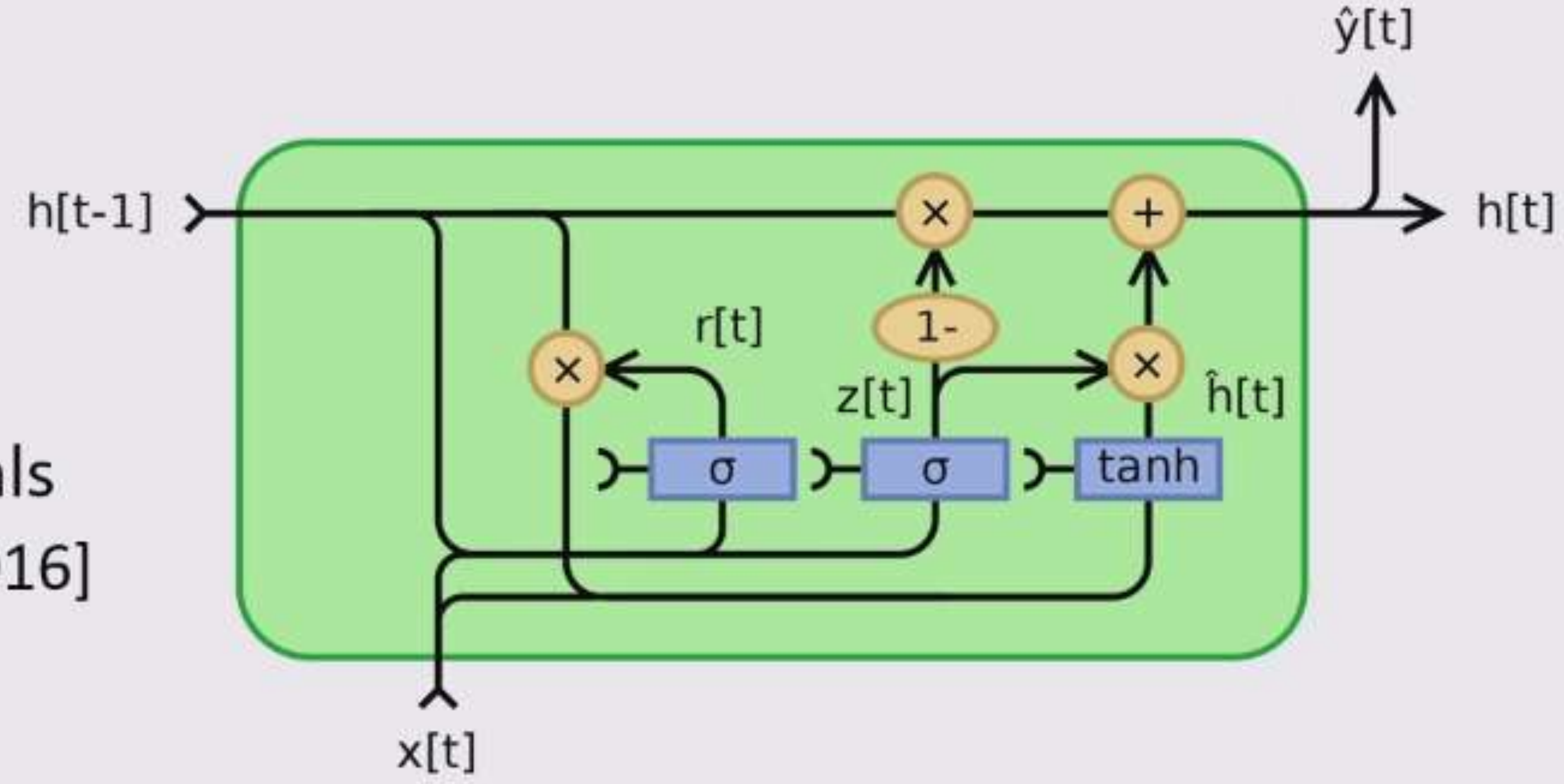


Image credit: https://en.wikipedia.org/wiki/Gated_recurrent_unit#/media/File:Gated_Recurrent_Unit,_base_type.svg. The original website is down.

Learning Machines

- Recurrent neural network (RNN) the most "natural" choice
  - Ability to encode long-term temporal patterns
  - Information exchange across frequencies
- Example: RNNoise [Valin2018]
  - GRUs [Bahdanau2014] encode temporal patterns
  - Full-connected (dense) layers transform composite features to a gain

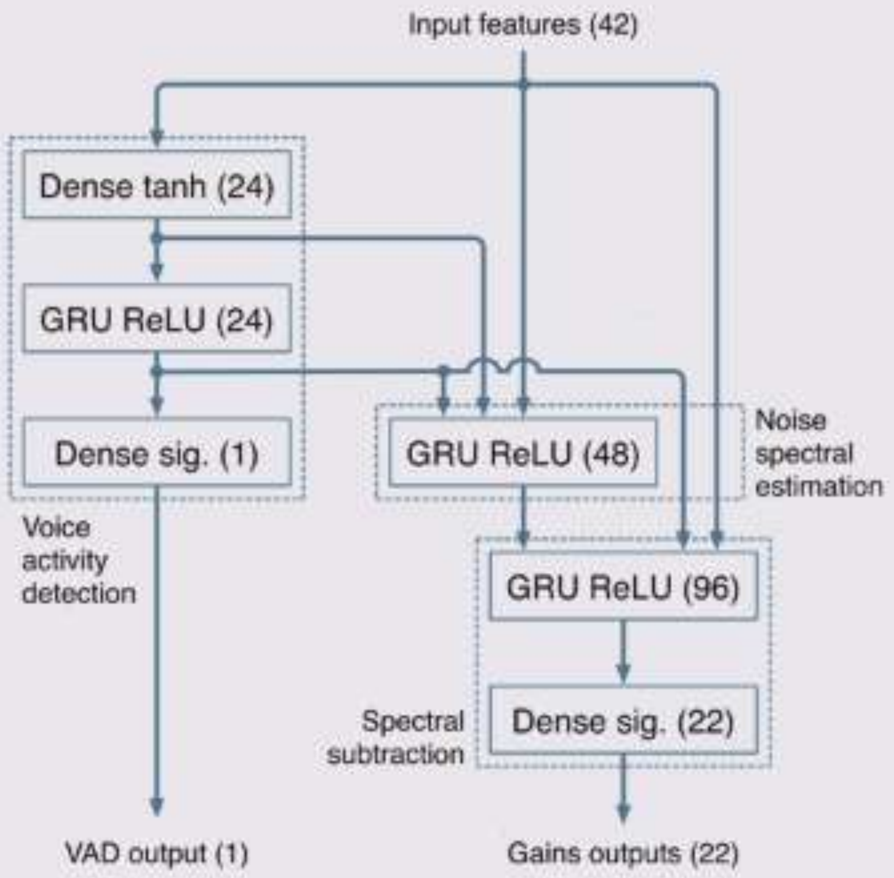


Image credit: [Valin2018]

Recurrent Units with Residual Connections

- Residual connections facilitate learning deep networks [He2016]
  - Depth = sequence length in our context
- Existing work using RNN + residuals
  - Sequence classification [Wang2016]
  - Automatic speech recognition [Kim2017]
  - Feature compensation for ASR [Chen2017]

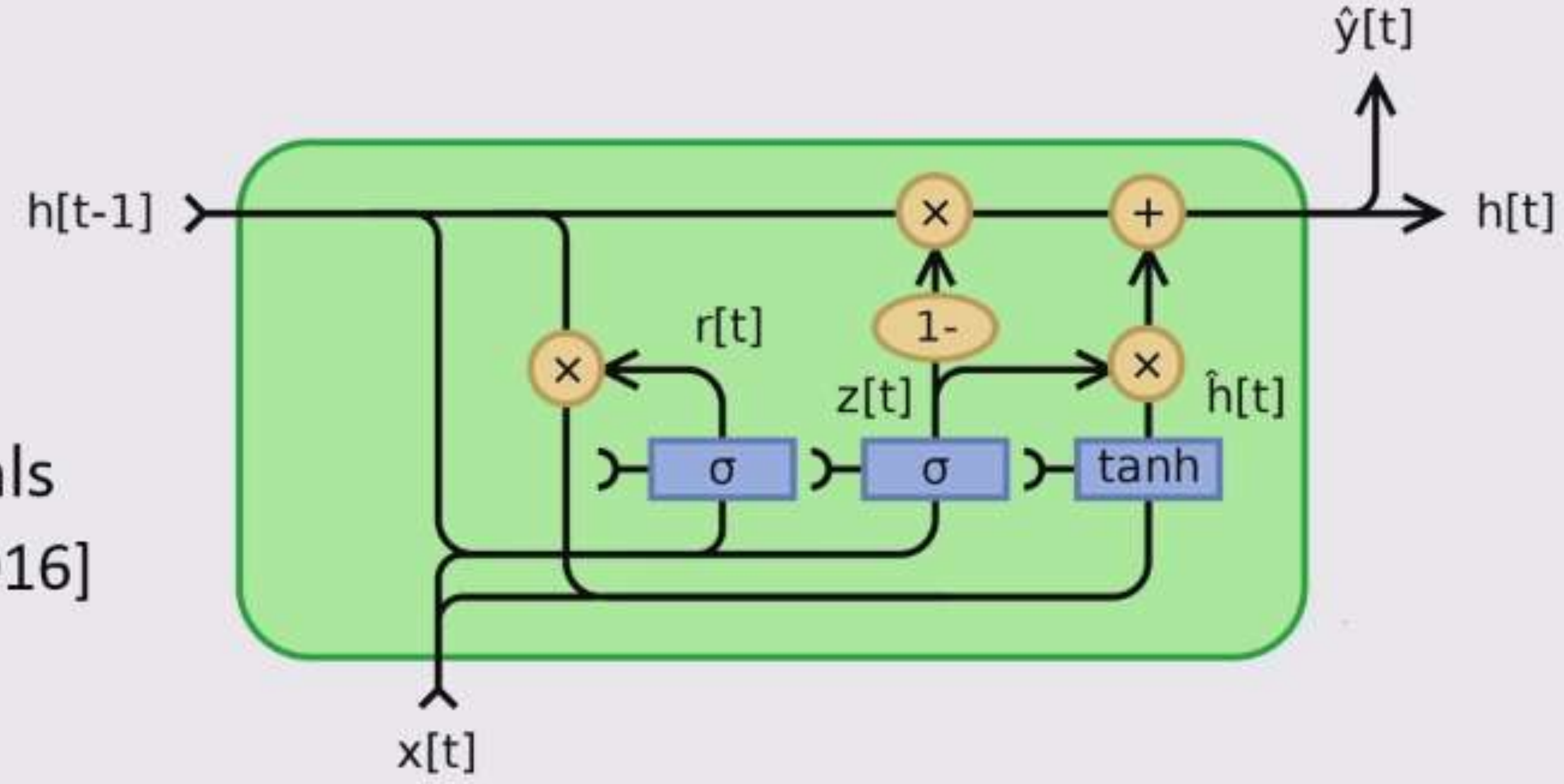


Image credit: https://en.wikipedia.org/wiki/Gated_recurrent_unit#/media/File:Gated_Recurrent_Unit,_base_type.svg. The original website is down.

Recurrent Units with Residual Connections

- Residual connections facilitate learning deep networks [He2016]
  - Depth = sequence length in our context
- Existing work using RNN + residuals
  - Sequence classification [Wang2016]
  - Automatic speech recognition [Kim2017]
  - Feature compensation for ASR [Chen2017]

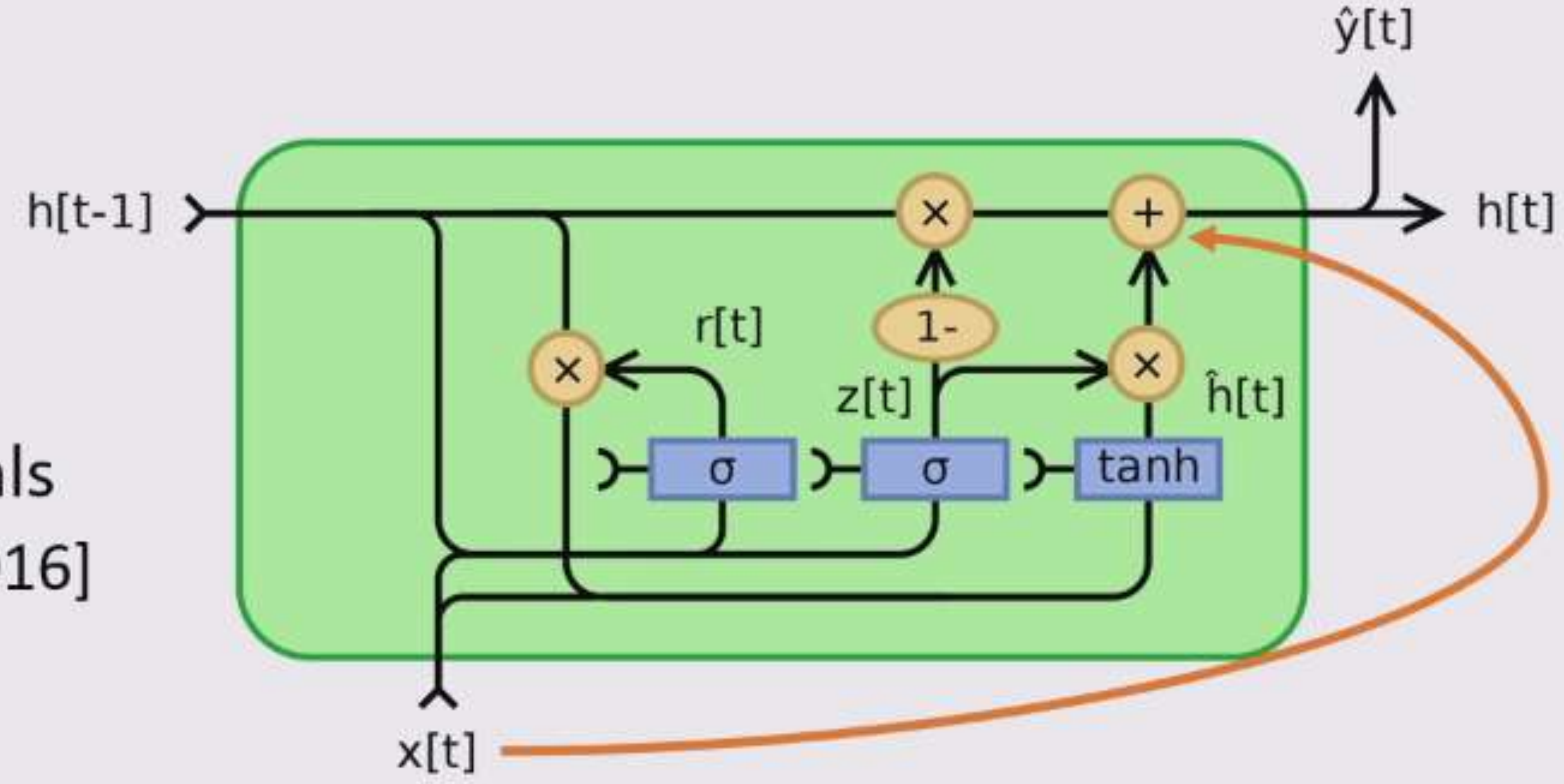
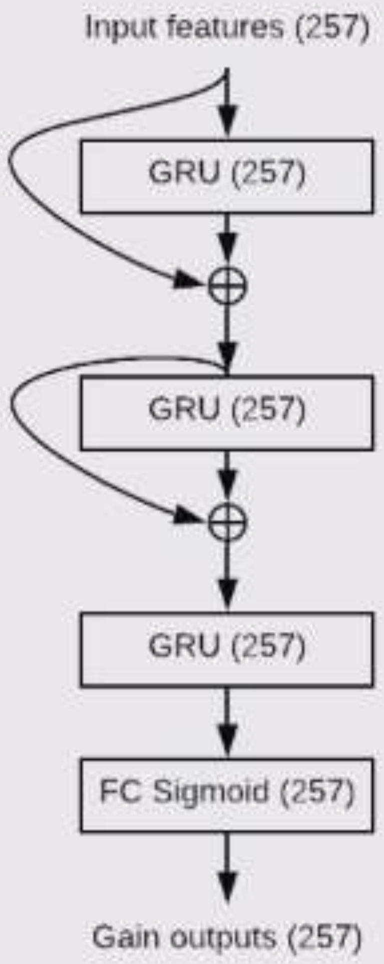


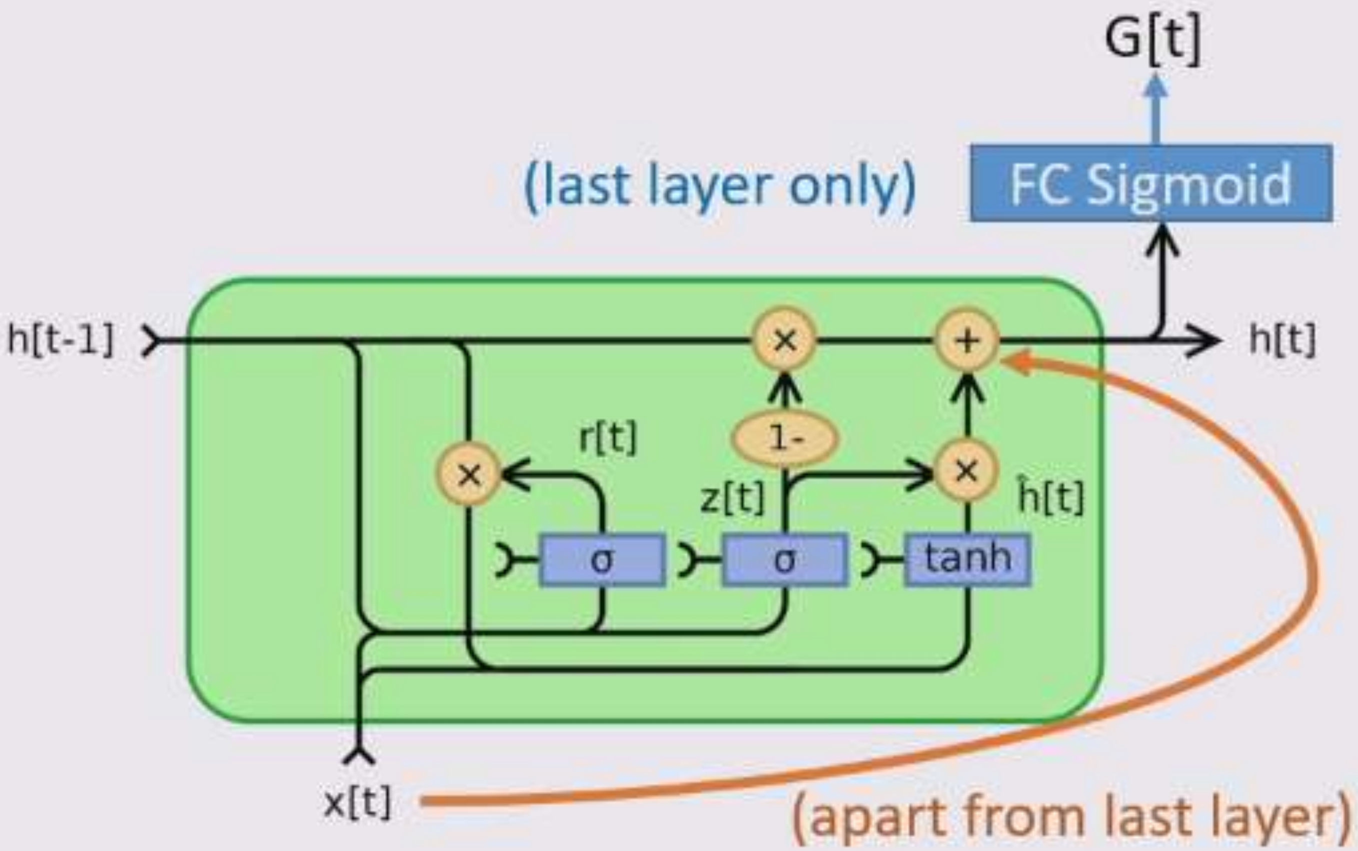
Image credit: https://en.wikipedia.org/wiki/Gated_recurrent_unit#/media/File:Gated_Recurrent_Unit,_base_type.svg. The original website is down.

GRU + Residuals for Speech Enhancement

- Global view



- Zoomed-in view



Learning Objectives

Learning Objectives

- Mean squared error (MSE)

Learning Objectives

- Mean squared error (MSE)

$$L_{MSE}(\Theta; X, Y) = \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} ||X_{t,f} - G(Y_{t,f}; \Theta) Y_{t,f}||^2$$

Learning Objectives

- Mean squared error (MSE)

$$L_{MSE}(\Theta; X, Y) = \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} ||X_{t,f} - G(Y_{t,f}; \Theta) Y_{t,f}||^2$$

- Classical statistical methods based on minimum MSE [Ephraim1984]
  - Assumes complex STFTs of speech/noise have Gaussian distribution
  - Assumes complex STFTs of speech and noise are uncorrelated
  - Solves for the optimal solution in MMSE sense

Learning Objectives

- Mean squared error (MSE)

$$L_{MSE}(\Theta; X, Y) = \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} ||X_{t,f} - G(Y_{t,f}; \Theta) Y_{t,f}||^2$$

- Classical statistical methods based on minimum MSE [Ephraim1984]
  - Assumes complex STFTs of speech/noise have Gaussian distribution
  - Assumes complex STFTs of speech and noise are uncorrelated
  - Solves for the optimal solution in MMSE sense
- Deep learning methods based on MSE
  - No assumptions about distributions
  - Solves for a “good” solution by stochastic gradient descent
    - Good = small MSE for both seen and unseen examples
  - Stable convergence (if able to learn at all)

Separating Speech and Noise Objectives

Separating Speech and Noise Objectives

- Re-writing the MSE

Separating Speech and Noise Objectives

- Re-writing the MSE

Cross terms ignored

$$E[(S-\hat{S})^2] = E[(S-G(S+N))^2] \approx E[(S-GS)^2] + E[(GN)^2]$$

Separating Speech and Noise Objectives

- Re-writing the MSE

Cross terms ignored

$$E[(S-\hat{S})^2] = E[(S-G(S+N))^2] \approx E[(S-GS)^2] + E[(GN)^2]$$

- Assign weighting to separated speech distortion and noise suppression terms

Separating Speech and Noise Objectives

- Re-writing the MSE

Cross terms ignored

$$E[(S-\hat{S})^2] = E[(S-G(S+N))^2] \approx E[(S-GS)^2] + E[(GN)^2]$$

- Assign weighting to separated speech distortion and noise suppression terms

$$L_{SN}(\Theta; S, N) = \alpha||S-GS||^2 + (1-\alpha)||GN||^2$$

Separating Speech and Noise Objectives

- Re-writing the MSE

Cross terms ignored

$$E[(S-\hat{S})^2] = E[(S-G(S+N))^2] \approx E[(S-GS)^2] + E[(GN)^2]$$

- Assign weighting to separated speech distortion and noise suppression terms

$$L_{SN}(\Theta; S, N) = \alpha||S-GS||^2 + (1-\alpha)||GN||^2$$

- Compute speech distortion only when speech is active (SA)
  - Energy-based SA detector: [300, 5000] Hz, max. 30dB below max. power

Separating Speech and Noise Objectives

- Re-writing the MSE

Cross terms ignored

$$E[(S-\hat{S})^2] = E[(S-G(S+N))^2] \approx E[(S-GS)^2] + E[(GN)^2]$$

- Assign weighting to separated speech distortion and noise suppression terms

$$L_{SN}(\Theta; S, N) = \alpha||S-GS||^2 + (1-\alpha)||GN||^2$$

- Compute speech distortion only when speech is active (SA)
  - Energy-based SA detector: [300, 5000] Hz, max. 30dB below max. power

$$\boxed{L_{SN}(\Theta; S_{SA}, N) = \alpha||S_{SA}-GS_{SA}||^2 + (1-\alpha)||GN||^2}$$

SNR-weighted Objectives

SNR-weighted Objectives

- The weighting is static, but our goal varies across different scenarios
  - We want little speech distortion when only speech is present (SNR → $+\infty$)
  - We want aggressive suppression when only noise is present (SNR → 0)
  - Existing work in classical SP approach [Low2011]

SNR-weighted Objectives

- The weighting is static, but our goal varies across different scenarios
  - We want little speech distortion when only speech is present (SNR → $+\infty$)
  - We want aggressive suppression when only noise is present (SNR → 0)
  - Existing work in classical SP approach [Low2011]
- Adapt the loss of each example pair {speech, noise} by the global SNR:

SNR-weighted Objectives

- The weighting is static, but our goal varies across different scenarios
  - We want little speech distortion when only speech is present (SNR → $+\infty$)
  - We want aggressive suppression when only noise is present (SNR → 0)
  - Existing work in classical SP approach [Low2011]
- Adapt the loss of each example pair {speech, noise} by the global SNR:

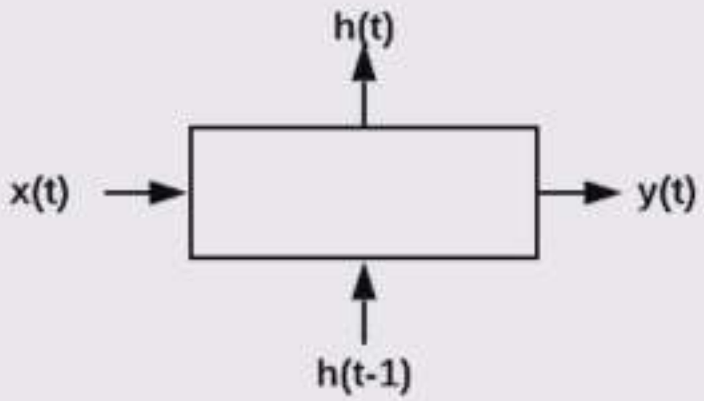
$$\boxed{L_{SNR}^{(i)}(\Theta; S^{(i)}, N^{(i)}) = \alpha \frac{\sigma^2_{S^{(i)}}}{\sigma^2_{N^{(i)}}} ||S_{SA}^{(i)} - GS_{SA}^{(i)}||^2 + (1-\alpha) \frac{\sigma^2_{S^{(i)}}}{\sigma_{N^{(i)}}} ||GN^{(i)}||^2}$$

Training Consideration

- Classical decision-directed approach [Ephraim1984]:
  - Transparent "hidden states" – *a priori* SNR, *a posteriori* SNR
  - Hidden states from the previous estimates affect the current by recursive smoothing
    - "Short-term memory that decays exponentially" in DL lingo
- RNN-based learning approach:
  - Black-box hidden states
  - LSTM/GRU are capable of learning long temporal patterns [Gers1999]
  - Patterns are learned through backpropagation through time [Werbos1990]

Training Consideration

- Backpropagation through time:



$$h(t) = y(t)$$

$$y(t) = f(x(t)+h(t-1))$$

$$\begin{aligned} y'(t) &= f'(x(t)+h(t-1))h'(t-1) \\ &= f'(x(t)+h(t-1))[f'(x(t-1)+h(t-2))h'(t-2)] \\ &= \dots \\ &= g(x(t), x(t-1), \dots, x(0)) \end{aligned}$$

- We want to compare a small batch of long sequences to a large batch of short sequences, given the same amount of information per batch.

Outline

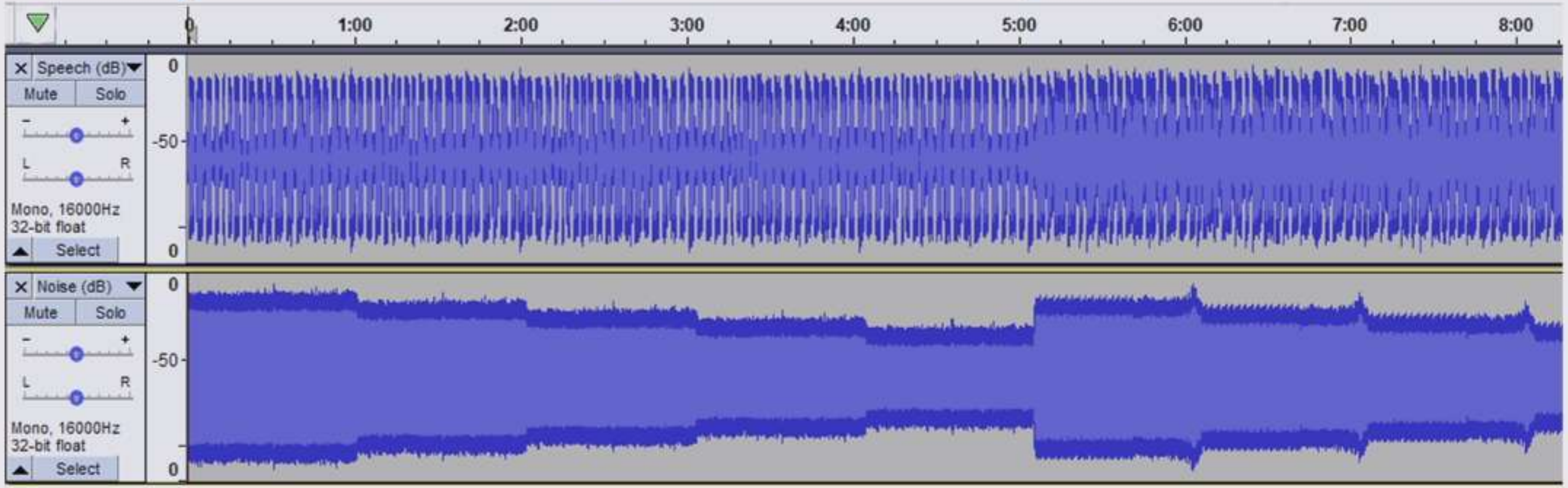
- Introduction to Single-channel Speech Enhancement
  - Classical signal processing vs. Deep learning
  - Considerations for online processing
- Our Method
  - Feature Representations
  - Learning Machines
  - Learning Objectives
  - Training Considerations
- **Evaluation**
  - Data
  - Metrics
  - Results
- Findings and Conclusions

Evaluation: Data

- 84 hours of training data
  - Speech: Edinburgh 56 Speakers Corpus
  - Noise: 14 noise types from DEMAND Database and Freesound
    - Air Conditioner, airport announcements, appliances, car noise, copy machine, door shutting, eating, multi-talker babble, neighbor speaking, squeaky chair, traffic, road, typing, vacuum cleaner.
- 18 hours of test data in 5500 clips
  - Speech: Graz University 20 Speakers Corpus
  - Noise: 9 challenging classes from DEMAND and Freesound
    - Air conditioner, airport announcements, babble, copy machine, munching, neighbor, shutting door, typing, vacuum cleaner.
    - All clips are unseen in training
- SNR: {40, 30, 20, 10, 0} dB
- All clips sampled at 16 kHz

Evaluation: Data & Data Augmentation

Same utterance from the same speaker repeated 5 times

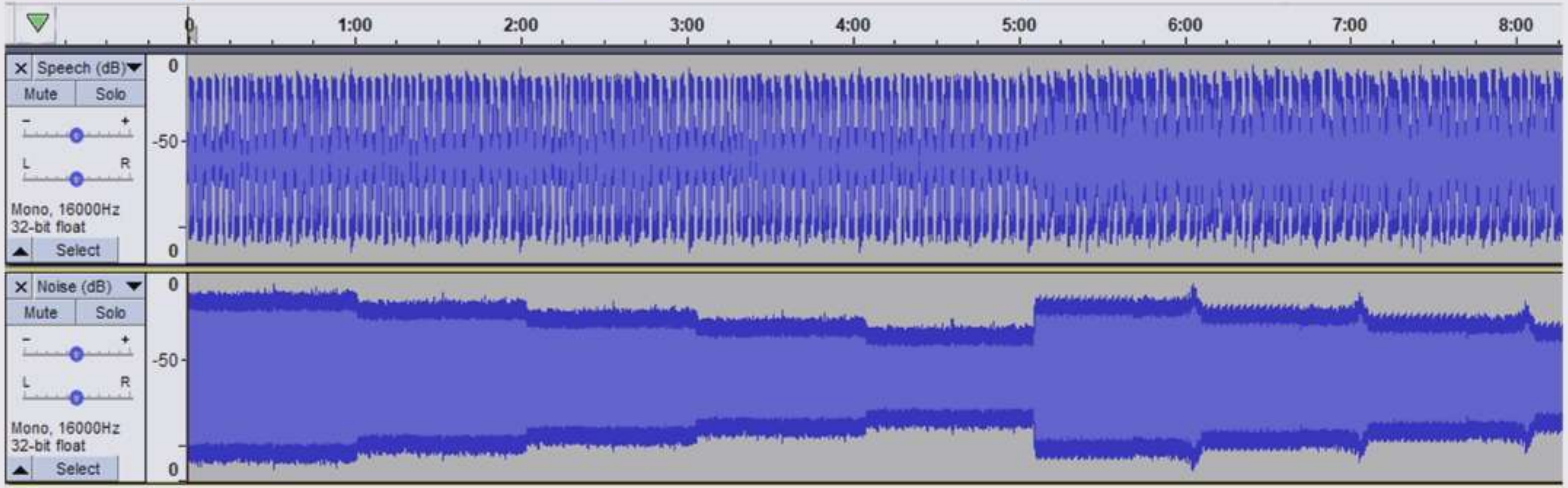


Same noise repeated 5 times with five discrete SNRs (assuming point-wise addition)

Evaluation: Data & Data Augmentation

During training, we randomly sample x-second speech and noise, respectively, and remix.

Same utterance from the same speaker repeated 5 times



Same noise repeated 5 times with five discrete SNRs (assuming point-wise addition)

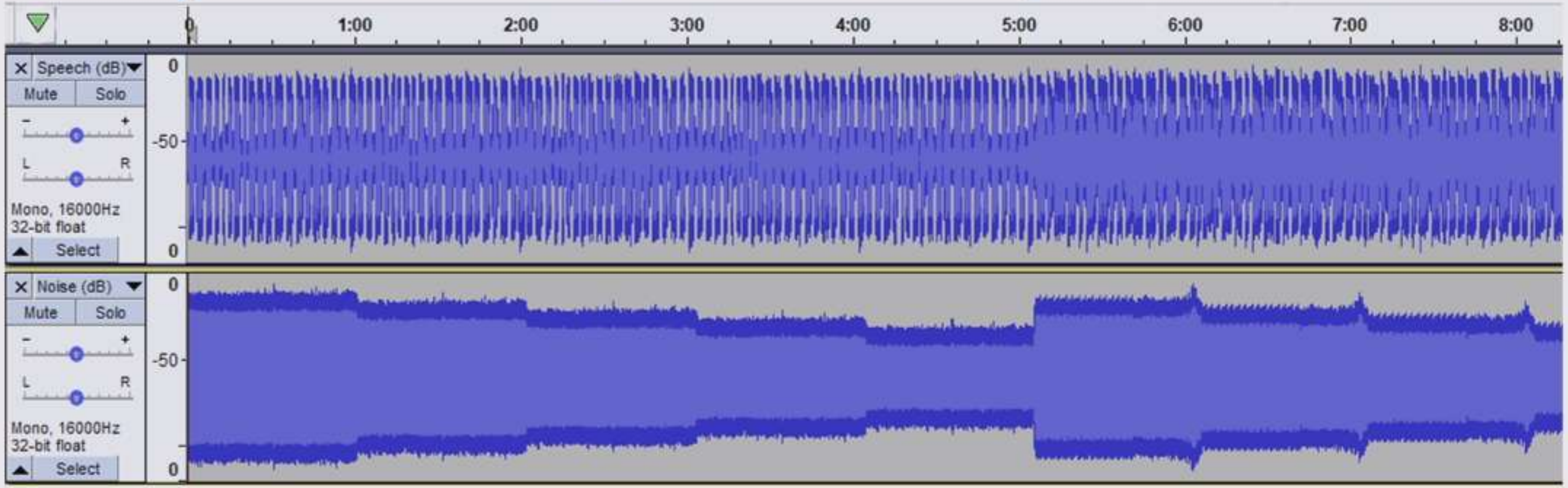
Baseline Systems

- Noisy
- MSR's statistical-based
- Proposed
  - Log spectra with global and FD online normalization; Twelve 5-second segments/batch; various objectives
- RNN
  - Same as proposed, except no residual connections; MSE loss
- RNNoise [Valin2018]
  - Online enhancement of 22-dimensional energy envelope with 42-dimensional features
  - No augmented data
- Simplified RNNoise
  - Full-band (257) enhancement; same network architecture as RNNoise
  - No VAD during training
- Oracle information + Wiener filter rule

Evaluation: Data & Data Augmentation

Same utterance from the same speaker repeated 5 times

During training, we randomly sample x-second speech and noise, respectively, and remix.



Same noise repeated 5 times with five discrete SNRs (assuming point-wise addition)

Baseline Systems

- Noisy
- MSR's statistical-based
- Proposed
  - Log spectra with global and FD online normalization; Twelve 5-second segments/batch; various objectives
- RNN
  - Same as proposed, except no residual connections; MSE loss
- RNNoise [Valin2018]
  - Online enhancement of 22-dimensional energy envelope with 42-dimensional features
  - No augmented data
- Simplified RNNoise
  - Full-band (257) enhancement; same network architecture as RNNoise
  - No VAD during training
- Oracle information + Wiener filter rule

Evaluation Metrics

- Classical speech quality/intelligibility measures
  - Scale-invariant signal-to-distortion ratio (SI-SDR) [LeRoux2019]
  - Cepstral distance (CD) [Hu2008]
  - Short-time objective intelligibility (STOI) [Taal2010]
  - Perceptual evaluation of speech quality (PESQ) [Rix2001]
- DNN-based mean opinion score (MOS) prediction
  - AudioMOS
    - Trained on MOS by real users
    - 0.89 Pearson correlation coefficient on test data

Baseline Systems

- Noisy
- MSR's statistical-based
- Proposed
  - Log spectra with global and FD online normalization; Twelve 5-second segments/batch; various objectives
- RNN
  - Same as proposed, except no residual connections; MSE loss
- RNNoise [Valin2018]
  - Online enhancement of 22-dimensional energy envelope with 42-dimensional features
  - No augmented data
- Simplified RNNoise
  - Full-band (257) enhancement; same network architecture as RNNoise
  - No VAD during training
- Oracle information + Wiener filter rule

Results: "Best" from each Category

Babble @ 20dB

Name	# Trainable Parameters	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
Noisy	0	9.81	4.56	88.0	2.22	2.40
Statistical-based	0	6.10	4.64	84.7	2.33	2.61
RNNoise	61.2 K	10.4	4.24	84.3	2.33	2.73
RNN	1.26 M	10.4	4.48	88.6	2.39	3.15
Full-band RNNoise	2.64 M	13.0	3.88	89.3	2.56	2.95
Proposed (SNR wt.; a = 0.35)	1.26 M	**14.8**	**3.72**	**90.9**	**2.71**	**3.24**
Oracle Wiener	Oracle	20.5	2.13	98.1	3.82	3.75

Results: “Best” from each Category

Babble @ 20dB

Name	# Trainable Parameters	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
Noisy	0	9.81	4.56	88.0	2.22	2.40
Statistical-based	0	6.10	4.64	84.7	2.33	2.61
RNNoise	61.2 K	10.4	4.24	84.3	2.33	2.73
RNN	1.26 M	10.4	4.48	88.6	2.39	3.15
Full-band RNNoise	2.64 M	13.0	3.88	89.3	2.56	2.95
Proposed (SNR wt.; a = 0.35)	1.26 M	**14.8**	**3.72**	**90.9**	**2.71**	**3.24**
Oracle Wiener	Oracle	20.5	2.13	98.1	3.82	3.75

Enhanced 1 second audio in 39.6 milliseconds on a single CPU
Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, Python 3.6.8

Results: "Best" from each Category

Babble @ 20dB

Name	# Trainable Parameters	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
Noisy	0	9.81	4.56	88.0	2.22	2.40
Statistical-based	0	6.10	4.64	84.7	2.33	2.61
RNNoise	61.2 K	10.4	4.24	84.3	2.33	2.73
RNN	1.26 M	10.4	4.48	88.6	2.39	3.15
Full-band RNNoise	2.64 M	13.0	3.88	89.3	2.56	2.95
Proposed (SNR wt.; a = 0.35)	1.26 M	**14.8**	**3.72**	**90.9**	**2.71**	**3.24**
Oracle Wiener	Oracle	20.5	2.13	98.1	3.82	3.75

Enhanced 1 second audio in 39.6 milliseconds on a single CPU
Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, Python 3.6.8

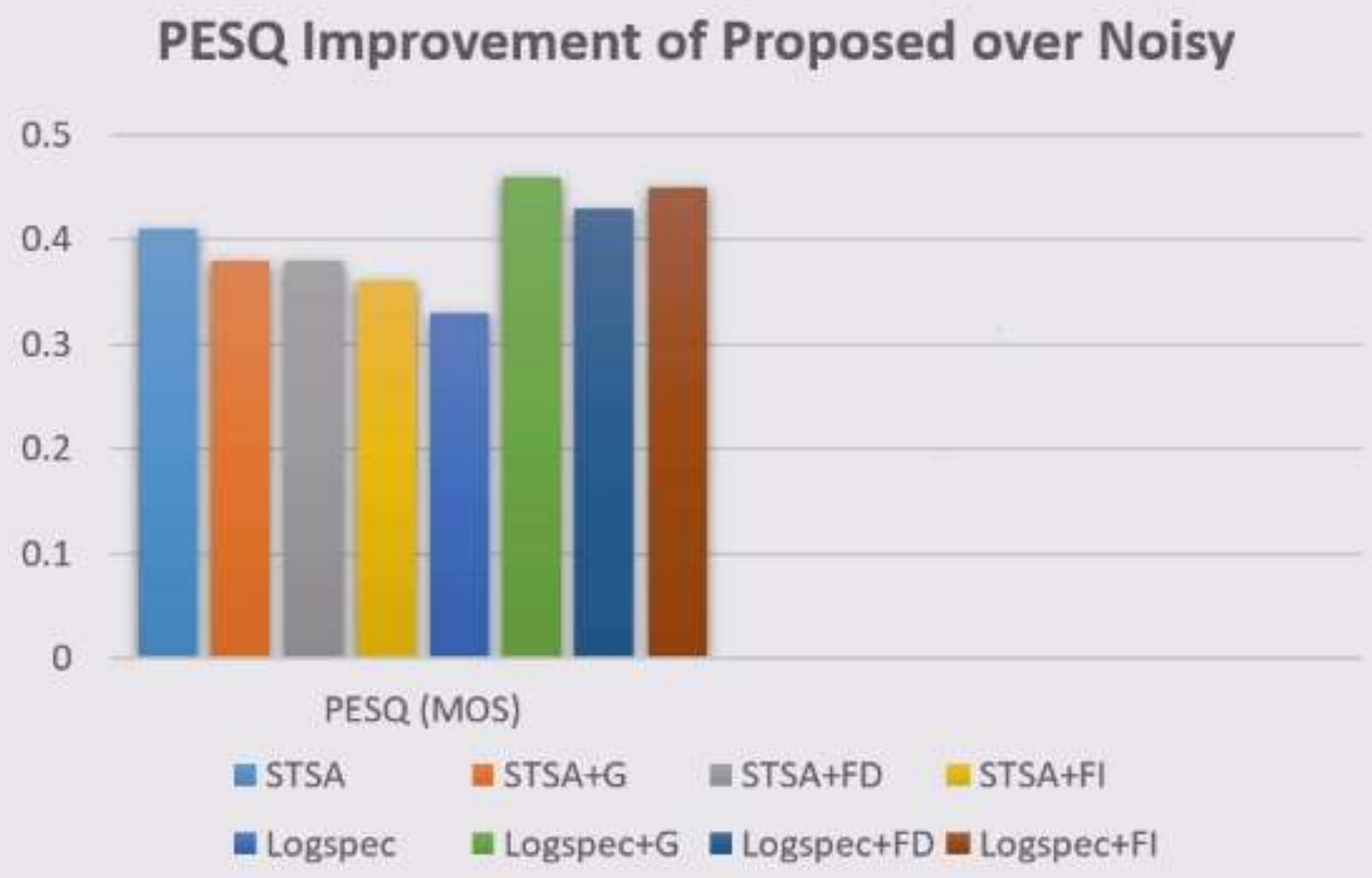
Results: "Best" from each Category

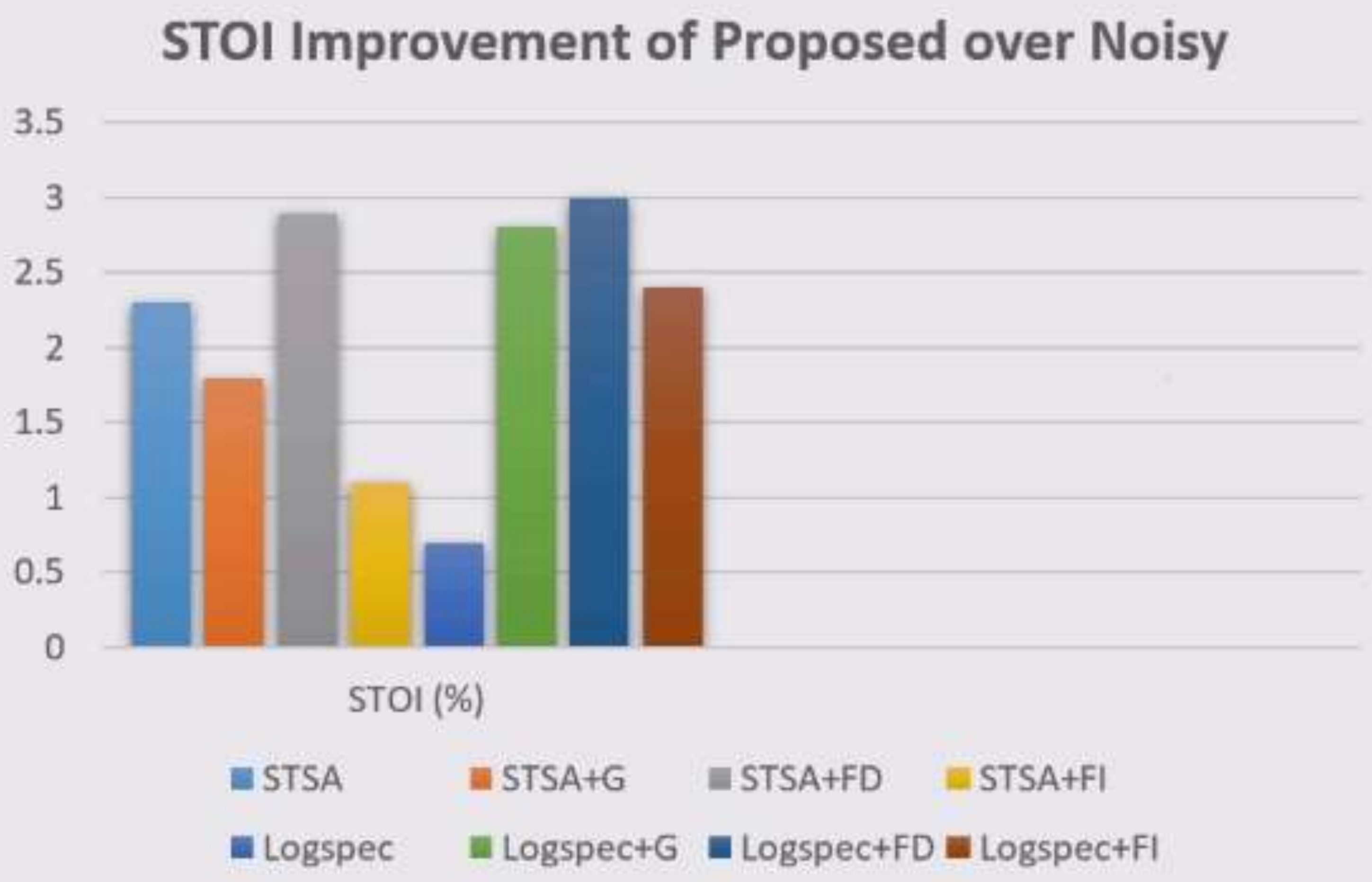
Babble @ 20dB

Name	# Trainable Parameters	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
Noisy	0	9.81	4.56	88.0	2.22	2.40
Statistical-based	0	6.10	4.64	84.7	2.33	2.61
RNNoise	61.2 K	10.4	4.24	84.3	2.33	2.73
RNN	1.26 M	10.4	4.48	88.6	2.39	3.15
Full-band RNNoise	2.64 M	13.0	3.88	89.3	2.56	2.95
Proposed (SNR wt.; a = 0.35)	1.26 M	**14.8**	**3.72**	**90.9**	**2.71**	**3.24**
Oracle Wiener	Oracle	20.5	2.13	98.1	3.82	3.75

Enhanced 1 second audio in 39.6 milliseconds on a single CPU
Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, Python 3.6.8

Results: Effect of Feature Normalization





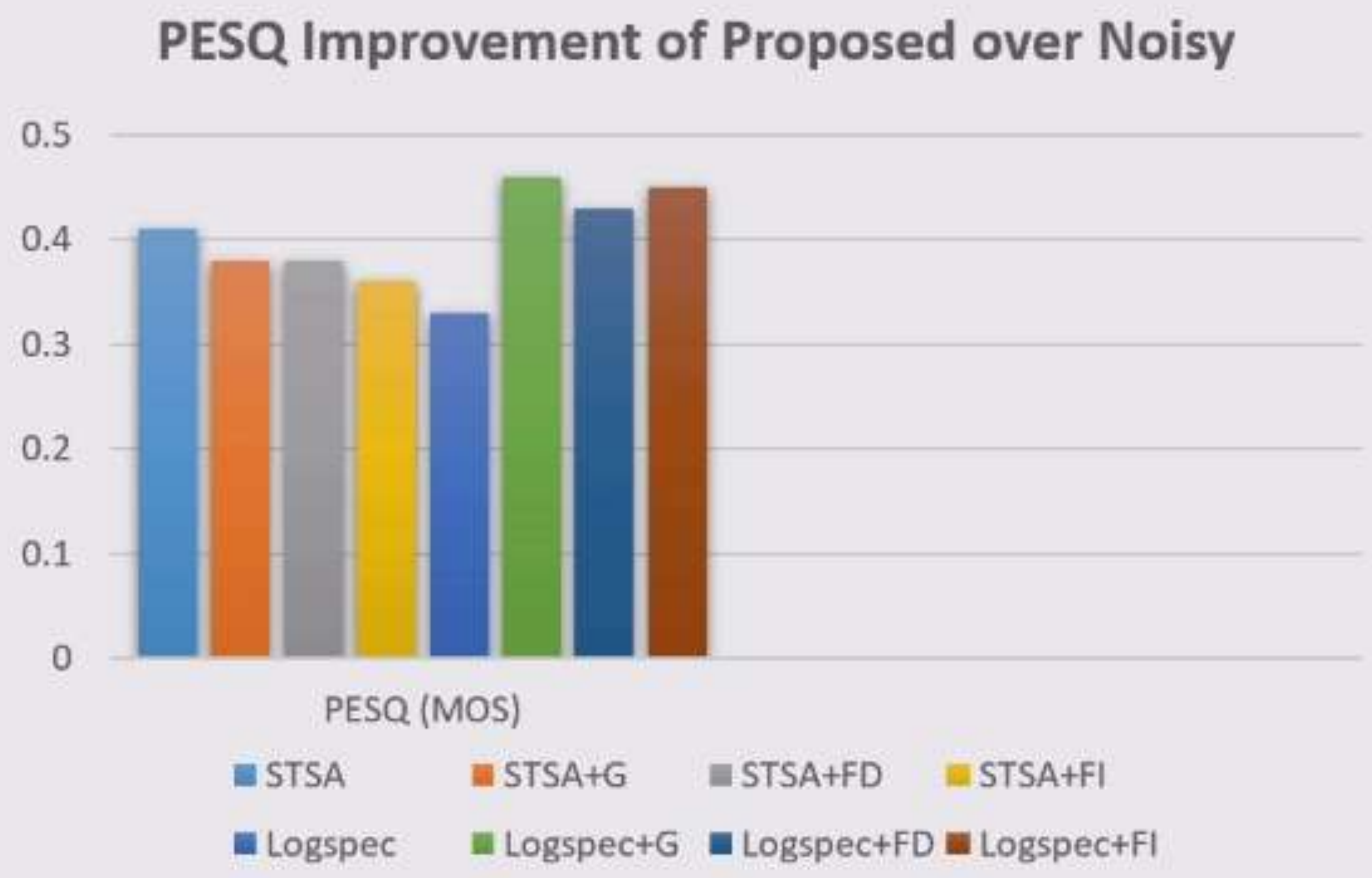
STSA – short-time spectral amplitude
Logspec – short-time log power spectra
G – global normalization
FD – Frequency-dependent online norm.
FI – Frequency-independent online norm.

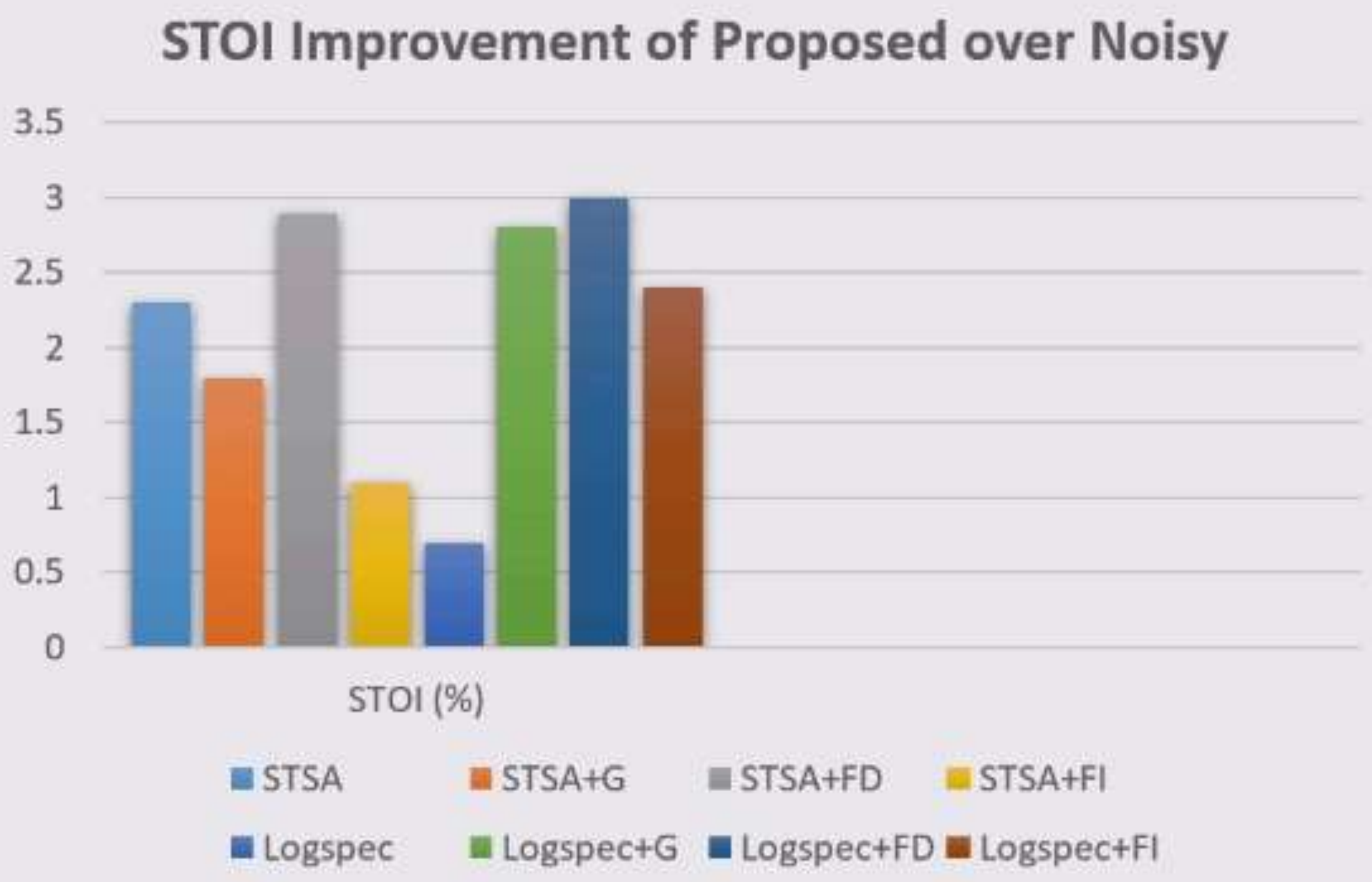
Results: Effect of Sequence Lengths

Duration (s/seg.)	# Seg. Per batch	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
1	60	13.8	3.81	90.6	2.61	2.82
5	12	**14.1**	**3.67**	**91.0**	**2.64**	2.88
15	4	**14.1**	3.74	90.7	**2.64**	**2.96**
30	2	13.8	3.79	90.3	2.60	2.91

Stopped early at 53/100 epochs.

Results: Effect of Feature Normalization





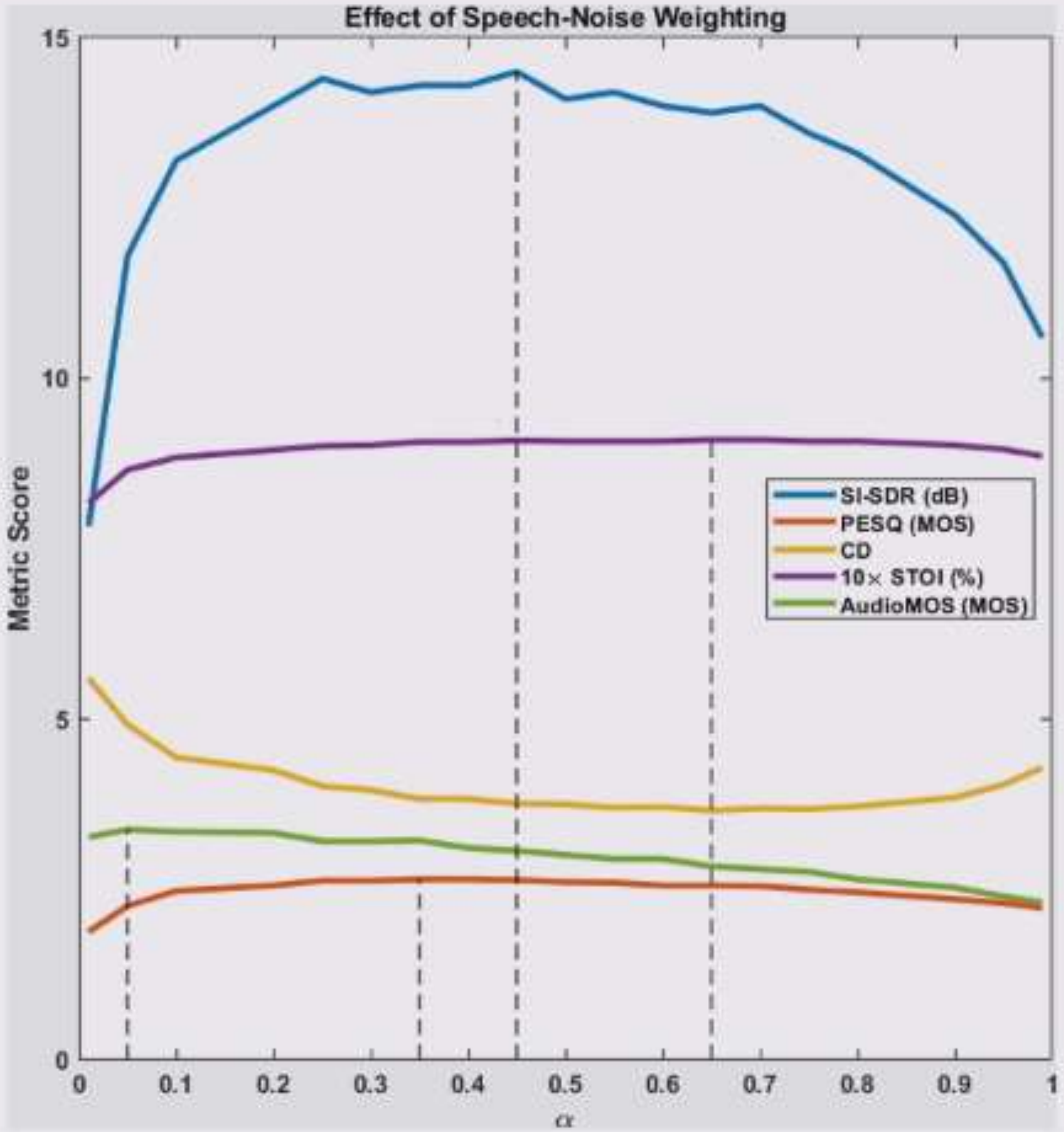
STSA – short-time spectral amplitude
Logspec – short-time log power spectra
G – global normalization
FD – Frequency-dependent online norm.
FI – Frequency-independent online norm.

Results: Effect of Sequence Lengths

Duration (s/seg.)	# Seg. Per batch	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)	AudioMOS (MOS)
1	60	13.8	3.81	90.6	2.61	2.82
5	12	**14.1**	**3.67**	**91.0**	**2.64**	2.88
15	4	**14.1**	3.74	90.7	**2.64**	**2.96**
30	2	13.8	3.79	90.3	2.60	2.91

Stopped early at 53/100 epochs.

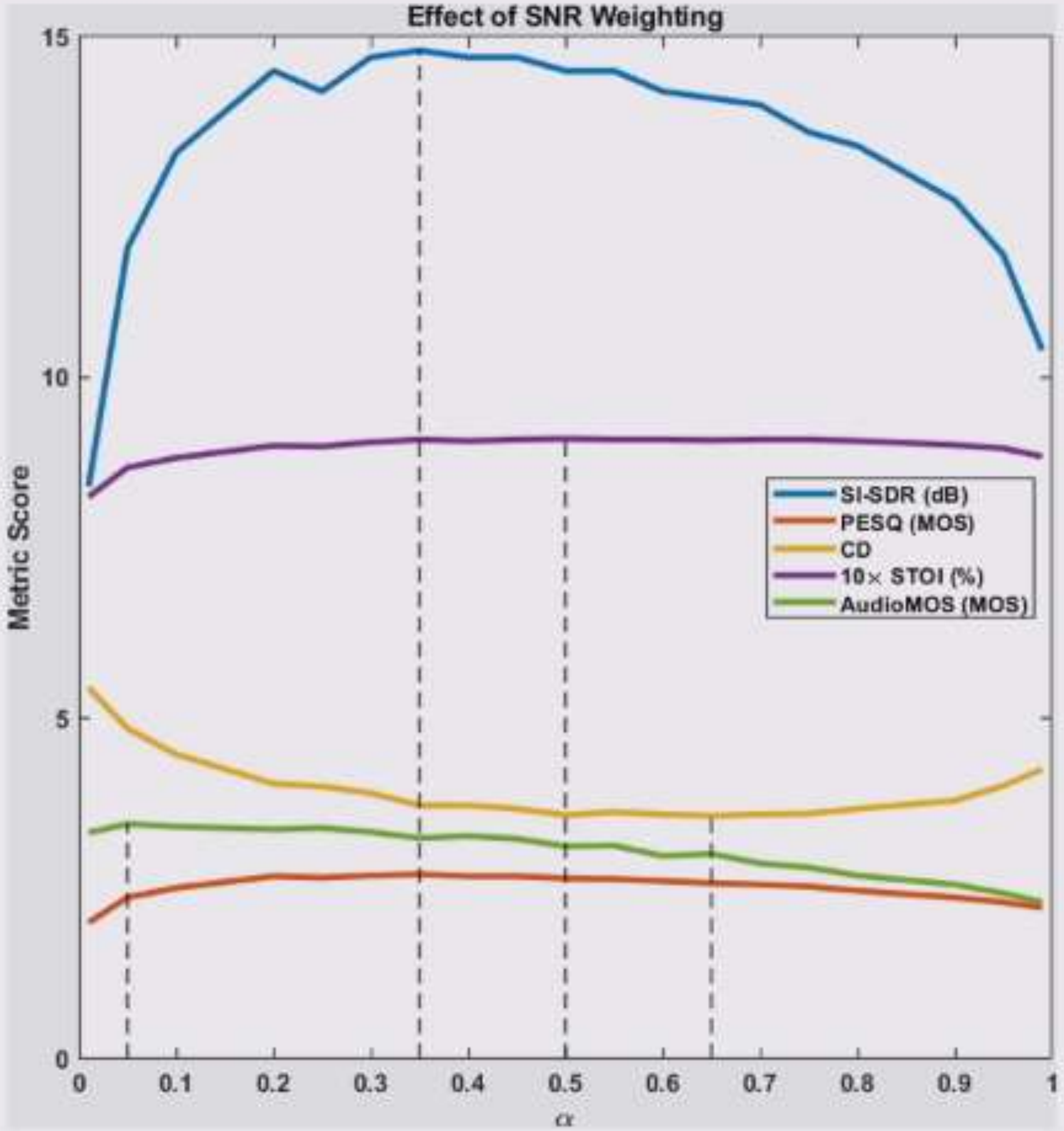
Results: Optimal Speech-Noise Weighting



- DEMO: Air Conditioner Noise

a \ SNR	20dB	10dB	0dB
0.05 (AudioMOS)			
0.35 (PESQ)			
0.45 (SI-SDR)			
0.65 (CD & STOI)			
Noisy			

Results: Optimal SNR-weighted SN Weighting



- DEMO: Airport Announcements

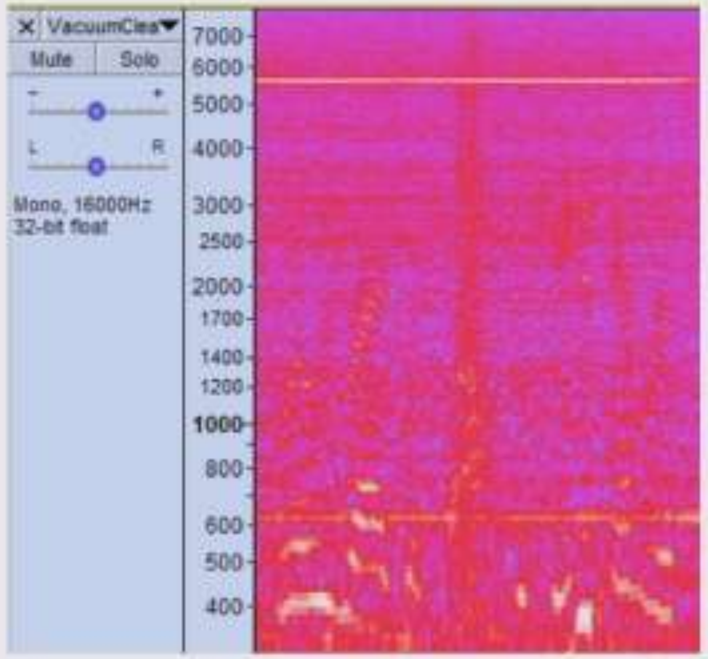
a \ SNR	20dB	10dB	0dB
0.05 (AudioMOS)			
0.35 (SI-SDR & PESQ)			
0.5 (STOI)			
0.65 (CD)			
Noisy			

Outline

- Introduction to Single-channel Speech Enhancement
    - Classical signal processing vs. Deep learning
    - Considerations for online processing
- Our Method
    - Feature Representations
    - Learning Machines
    - Learning Objectives
    - Training Considerations
- Evaluation
    - Data
    - Metrics
    - Results
- **Findings and Conclusions**

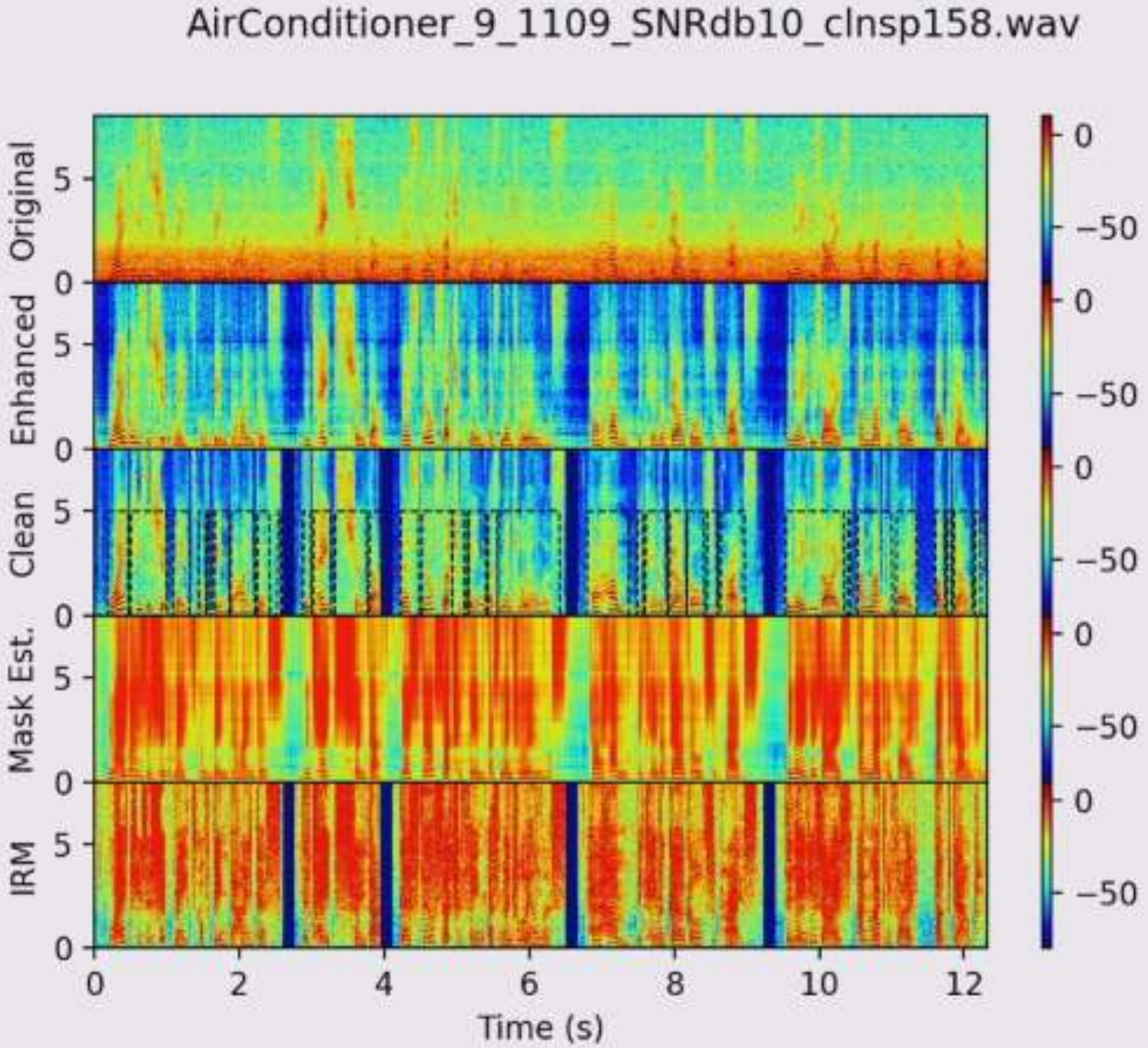
Major Findings

- Residual connections within recurrent cells really, really help
- GRUs are able to encode extremely long temporal patterns in high dimensional space (probably with the aid of residual connections)
  - 5-second waveform = 625 frames of 257-point spectra
- Trust the old faithful for stationary patterns?
  - The model learns to ALWAYS strongly suppress ~6 kHz

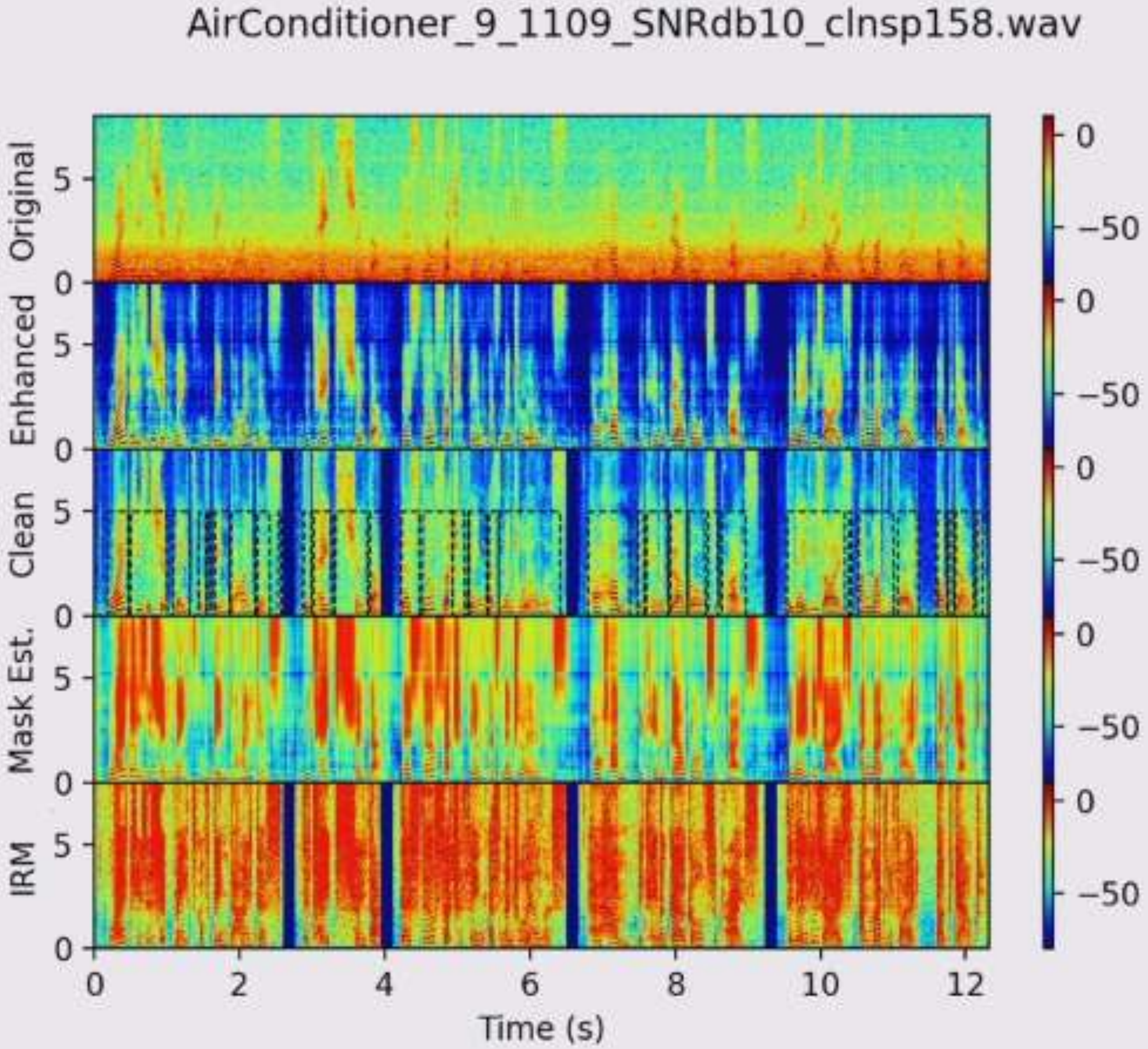


Major Findings

- MSE

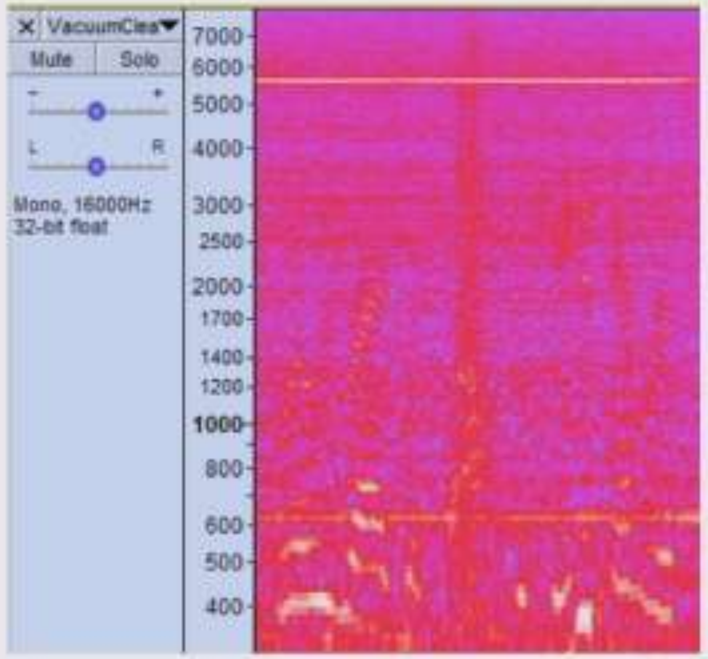


- SNR-weighted (a=0.2)



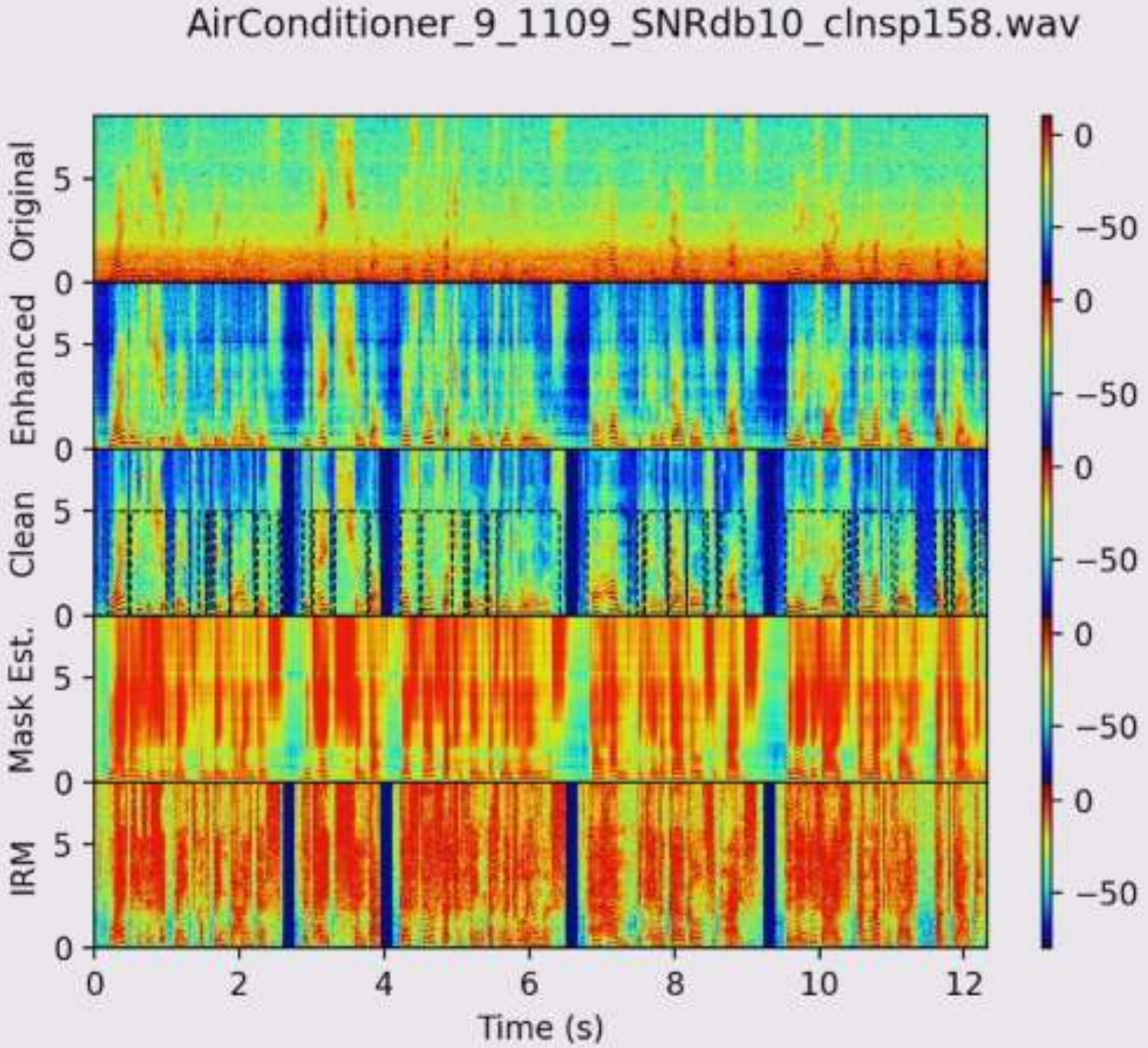
Major Findings

- Residual connections within recurrent cells really, really help
- GRUs are able to encode extremely long temporal patterns in high dimensional space (probably with the aid of residual connections)
  - 5-second waveform = 625 frames of 257-point spectra
- Trust the old faithful for stationary patterns?
  - The model learns to ALWAYS strongly suppress ~6 kHz

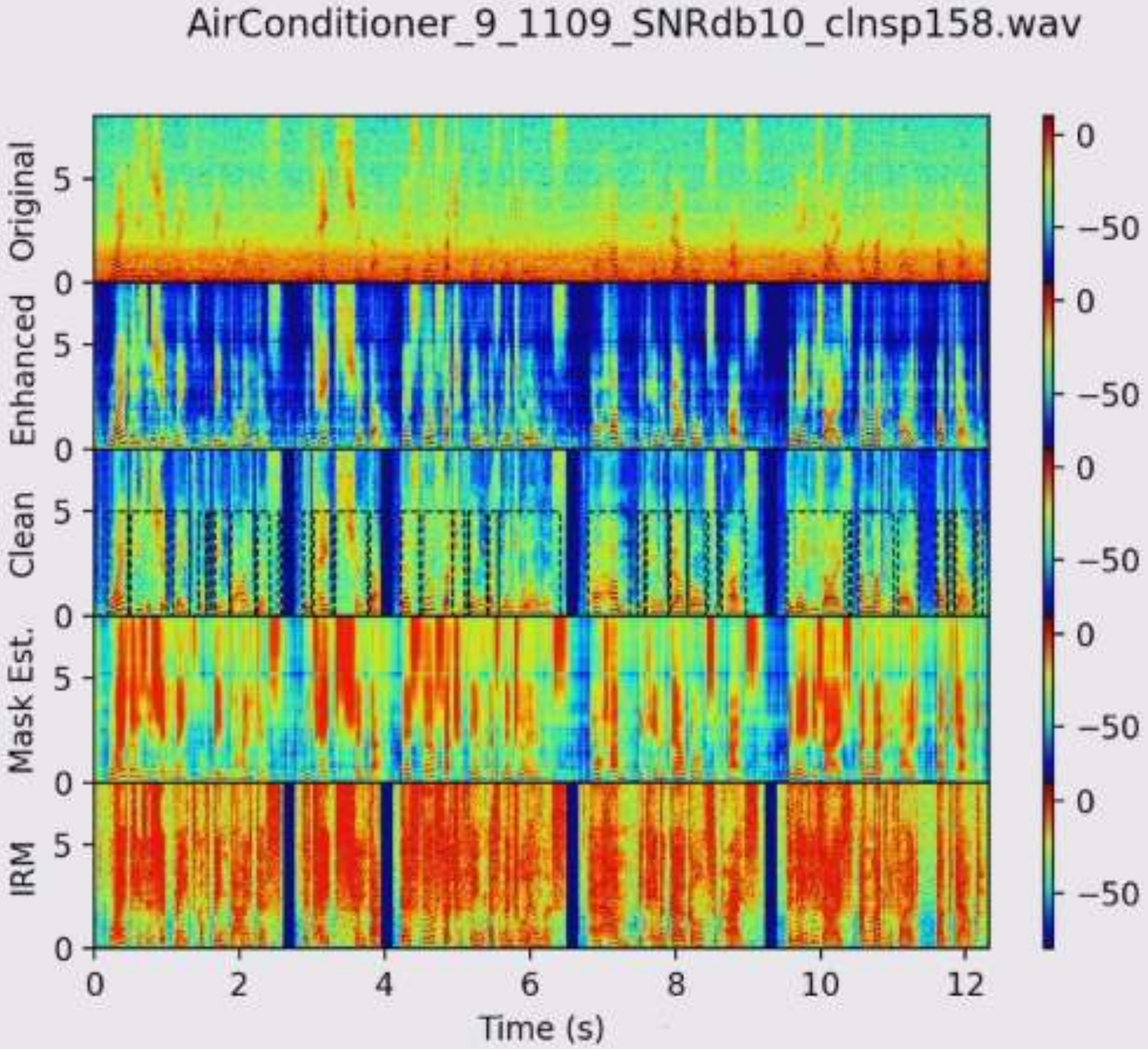


Major Findings

- MSE



- SNR-weighted (a=0.2)



Conclusions

Conclusions

- We proposed a DNN-based online speech enhancement system
  - A compact recurrent network with residual connections
  - Two novel learning objectives motivated by balancing speech distortion and noise suppression
    - The speech-noise weighting happens to coincide (apart from VAD) with a paper published on arxiv a few days ago.

Conclusions

- We proposed a DNN-based online speech enhancement system
  - A compact recurrent network with residual connections
  - Two novel learning objectives motivated by balancing speech distortion and noise suppression
    - The speech-noise weighting happens to coincide (apart from VAD) with a paper published on arxiv a few days ago.
- We studied the impact of multiple factors associated with training a RNN on speech quality
  - Feature normalization, sequence length, objective weightings

Conclusions

- We proposed a DNN-based online speech enhancement system
  - A compact recurrent network with residual connections
  - Two novel learning objectives motivated by balancing speech distortion and noise suppression
    - The speech-noise weighting happens to coincide (apart from VAD) with a paper published on arxiv a few days ago.
- We studied the impact of multiple factors associated with training a RNN on speech quality
  - Feature normalization, sequence length, objective weightings
- We compared multiple competitive SP or DL-based online systems in terms of objective speech quality measures

Future Directions

- Study the speech quality improvement by SNR
- Investigate learning objectives to replace MSE
  - MAE, log-domain and cepstral-domain objectives
- Feature dimensionality reduction
  - Speech energy is sparse and noisy at very high frequencies

Thank you!

- Sebastian, Hannes, and all mentors from the Audio and Acoustics Group
- Ross, Chandan, and Hari from Skype
- All interns from the Audio and Acoustics Group

- Stay in touch!
  - School email: raymondxia@cmu.edu
  - Personal email: raymondxia@pm.me