End-To-End Deep Learning-Based Adaptation Control for Frequency-Domain Adaptive System Identification

T. Haubner, A. Brendel, and W. Kellermann

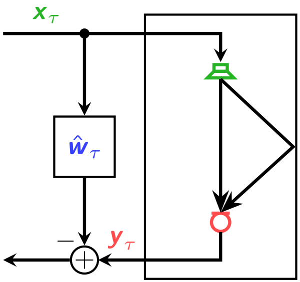
Chair of Multimedia Communications and Signal Processing

Friedrich-Alexander-Universität
Technische Fakultät

icassp 2022
Singapore

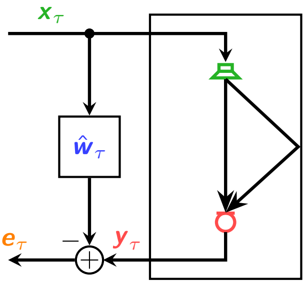
Motivation: Acoustic Echo Cancellation

- **Problem**: Identify acoustic transfer function (ATF) between loudspeaker and microphone signal



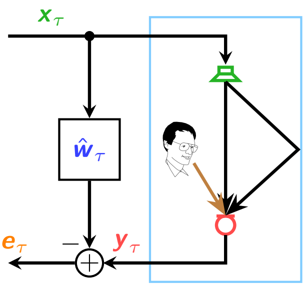
Motivation: Acoustic Echo Cancellation

- **Problem**: Identify acoustic transfer function (ATF) between loudspeaker and microphone signal
- **Approach:**
  - Minimization of error signal power
  - Iterative update of ATF estimate: $\hat{\boldsymbol{w}}_\tau = \hat{\boldsymbol{w}}_{\tau-1} + \delta\hat{\boldsymbol{w}}_\tau$



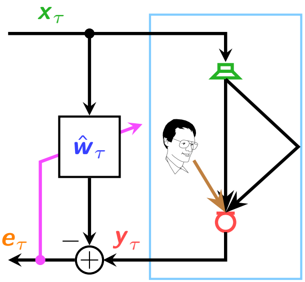
Motivation: Acoustic Echo Cancellation

- **Problem**: Identify acoustic transfer function (ATF) between loudspeaker and microphone signal
- **Approach:**
  - Minimization of error signal power
  - Iterative update of ATF estimate: $\hat{\boldsymbol{w}}_\tau = \hat{\boldsymbol{w}}_{\tau-1} + \delta\hat{\boldsymbol{w}}_\tau$
- **Challenges:**
  - Interfering signals, e.g., local speech or noise
  - Time-varying acoustic environments



Motivation: Acoustic Echo Cancellation

- **Problem**: Identify acoustic transfer function (ATF) between loudspeaker and microphone signal
- **Approach:**
  - Minimization of error signal power
  - Iterative update of ATF estimate: $\hat{\mathbf{w}}_\tau = \hat{\mathbf{w}}_{\tau-1} + \delta\hat{\mathbf{w}}_\tau$
- **Challenges:**
  - Interfering signals, e.g., local speech or noise
  - Time-varying acoustic environments



Robust adaptation control for improved convergence rate

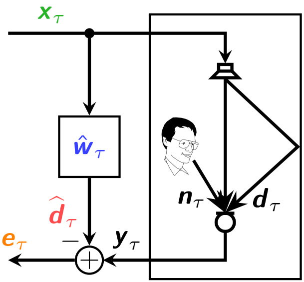
Outline

- Acoustic System Identification
- Proposed DNN-Based Adaptation Control
- Experimental Evaluation
- Conclusion

Iterative ATF Estimation

$$\widehat{\boldsymbol{d}}_\tau = \quad \boldsymbol{Q}(\boldsymbol{x}_\tau \odot \hat{\boldsymbol{w}}_{\tau-1})$$

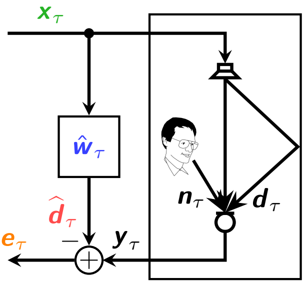
1. Estimate echo $\widehat{\boldsymbol{d}}_\tau$ by linear convolution of
   - loudspeaker signal block $\boldsymbol{x}_\tau$
   - with ATF estimate $\hat{\boldsymbol{w}}_{\tau-1}$.
   - and linear convolution constraint matrix $\boldsymbol{Q}$



Iterative ATF Estimation

$$\boldsymbol{e}_\tau = \boldsymbol{y}_\tau - \widehat{\boldsymbol{d}}_\tau = \boldsymbol{y}_\tau - \boldsymbol{Q}(\boldsymbol{x}_\tau \odot \hat{\boldsymbol{w}}_{\tau-1})$$

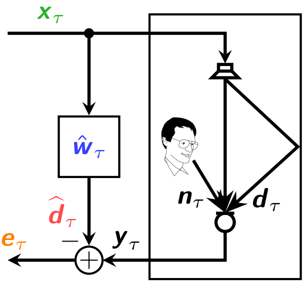
1. Estimate echo $\widehat{\boldsymbol{d}}_\tau$ by linear convolution of
   - loudspeaker signal block $\boldsymbol{x}_\tau$
   - with ATF estimate $\hat{\boldsymbol{w}}_{\tau-1}$.
   - and linear convolution constraint matrix $\boldsymbol{Q}$
2. Compute error signal block $\boldsymbol{e}_\tau$



Iterative ATF Estimation

$$\boldsymbol{e}_\tau = \boldsymbol{y}_\tau - \widehat{\boldsymbol{d}}_\tau = \boldsymbol{y}_\tau - \boldsymbol{Q}(\boldsymbol{x}_\tau \odot \hat{\boldsymbol{w}}_{\tau-1})$$
$$\hat{\boldsymbol{w}}_\tau = \hat{\boldsymbol{w}}_{\tau-1} + \boldsymbol{G} \quad (\boldsymbol{x}_\tau^* \odot \boldsymbol{e}_\tau)$$

1. Estimate echo $\widehat{\boldsymbol{d}}_\tau$ by linear convolution of
   - loudspeaker signal block $\boldsymbol{x}_\tau$
   - with ATF estimate $\hat{\boldsymbol{w}}_{\tau-1}$.
   - and linear convolution constraint matrix $\boldsymbol{Q}$
2. Compute error signal block $\boldsymbol{e}_\tau$
3. Compute stochastic gradient $\boldsymbol{G}\,(\boldsymbol{x}_\tau^* \odot \boldsymbol{e}_\tau)$ with FIR filter projection matrix $\boldsymbol{G}$
4. Perform gradient descent



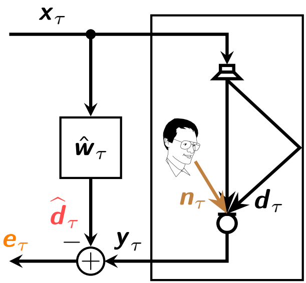
Adaptation Control

$$\boldsymbol{e}_\tau = \boldsymbol{y}_\tau - \widehat{\boldsymbol{d}}_\tau = \boldsymbol{n}_\tau + \left(\boldsymbol{d}_\tau - \widehat{\boldsymbol{d}}_\tau\right)$$

$$\hat{\boldsymbol{w}}_\tau = \hat{\boldsymbol{w}}_{\tau-1} + \boldsymbol{G} \quad (\boldsymbol{x}_\tau^* \odot \boldsymbol{e}_\tau)$$

Large error powers $||\boldsymbol{e}_\tau||^2$ could result from

- system mismatch, i.e., imprecise echo estimates $\widehat{\boldsymbol{d}}_\tau$
- interfering signals $\boldsymbol{n}_\tau$



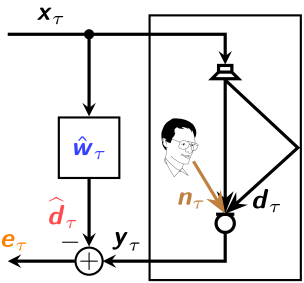
Adaptation Control

$$\boldsymbol{e}_\tau = \boldsymbol{y}_\tau - \widehat{\boldsymbol{d}}_\tau = \boldsymbol{n}_\tau + \left(\boldsymbol{d}_\tau - \widehat{\boldsymbol{d}}_\tau\right)$$

$$\hat{\boldsymbol{w}}_\tau = \hat{\boldsymbol{w}}_{\tau-1} + \boldsymbol{G} \quad (\boldsymbol{x}_\tau^* \odot \boldsymbol{e}_\tau)$$

Large error powers $||\boldsymbol{e}_\tau||^2$ could result from

- system mismatch, i.e., imprecise echo estimates $\widehat{\boldsymbol{d}}_\tau$ $\Rightarrow$ Update filter coefficients
- interfering signals $\boldsymbol{n}_\tau$ $\Rightarrow$ Stall filter adaptation



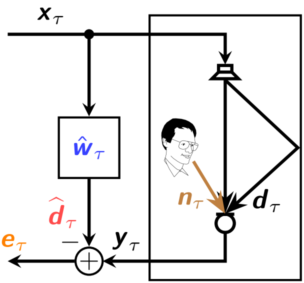
Adaptation Control

$$\mathbf{e}_\tau = \mathbf{y}_\tau - \widehat{\mathbf{d}}_\tau = \mathbf{n}_\tau + \left(\mathbf{d}_\tau - \widehat{\mathbf{d}}_\tau\right)$$

$$\hat{\mathbf{w}}_\tau = \hat{\mathbf{w}}_{\tau-1} + \mathbf{G}\boldsymbol{\Lambda}_\tau(\mathbf{x}_\tau^* \odot \mathbf{e}_\tau)$$

Large error powers $||\mathbf{e}_\tau||^2$ could result from

- system mismatch, i.e., imprecise echo estimates $\widehat{\mathbf{d}}_\tau$ ⇒ Update filter coefficients
- interfering signals $\mathbf{n}_\tau$ ⇒ Stall filter adaptation



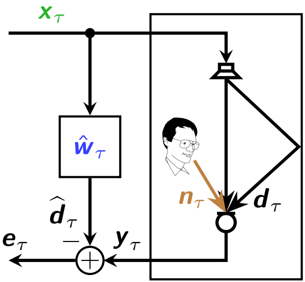
Solution: Control adaptation by time- and frequency-dependent step-sizes $[\boldsymbol{\Lambda}_\tau]_{ff}$

Traditional Model-Based Adaptation Control

$$\boldsymbol{\Lambda}_\tau^{\mathrm{MB}} = f_{\mathrm{MB}}\left(\boldsymbol{\Psi}_\tau^{\mathrm{XX}}, \boldsymbol{\Psi}_\tau^{\mathrm{NN}}, \boldsymbol{\Psi}_\tau^{\Delta\mathrm{W}\Delta\mathrm{W}}, \dots\right)$$

Compute step-size matrix $\boldsymbol{\Lambda}_\tau^{\mathrm{MB}}$ as a function of

- loudspeaker PSD $\boldsymbol{\Psi}_\tau^{\mathrm{XX}}$
- interference PSD $\boldsymbol{\Psi}_\tau^{\mathrm{NN}}$
- filter estimation uncertainty $\boldsymbol{\Psi}_\tau^{\Delta\mathrm{W}\Delta\mathrm{W}}$
- ...



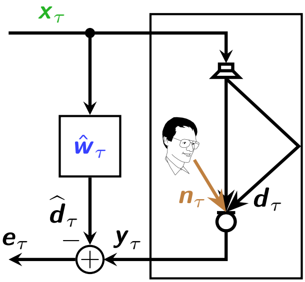
Prominent examples:
FDAF, DFT-domain Kalman filter, ...

Traditional Model-Based Adaptation Control

$$\mathbf{\Lambda}_\tau^{\mathrm{MB}} = f_{\mathrm{MB}}\left(\mathbf{\Psi}_\tau^{\mathrm{XX}}, \mathbf{\Psi}_\tau^{\mathrm{NN}}, \mathbf{\Psi}_\tau^{\Delta\mathrm{W}\Delta\mathrm{W}}, \dots\right)$$

Challenges

- Estimation of signal statistics of unknown quantities, e.g., $\mathbf{\Psi}_\tau^{\mathrm{NN}}$
- Mismatch of assumed model properties



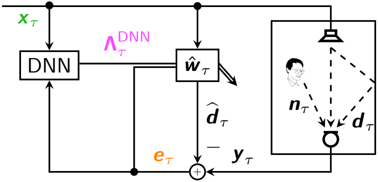
Outline

- Acoustic System Identification
- Proposed DNN-Based Adaptation Control
- Experimental Evaluation
- Conclusion

Proposed General Concept

$$\boldsymbol{\Lambda}_\tau^{\text{DNN}} = f_{\text{DNN}}\left(\boldsymbol{x}_1, \boldsymbol{e}_1, \ldots, \boldsymbol{x}_\tau, \boldsymbol{e}_\tau; \boldsymbol{\theta}\right)$$

- Learn mapping f_{DNN} of observed signal sequences to step-size matrices
- DNN parameters $\boldsymbol{\theta}$ are learned from training data



Proposed Step-Size Structure

$$\left[\mathbf{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = m_{f,\tau}^{\mu}$$

DNN provides raw step-size $m_{f,\tau}^{\mu} \in [0, \mu_{\mathrm{MAX}}]$

Proposed Step-Size Structure

$$\left[\mathbf{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = m_{f,\tau}^{\mu}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals

DNN provides raw step-size $m_{f,\tau}^{\mu} \in [0, \mu_{\mathrm{MAX}}]$

Proposed Step-Size Structure

$$\left[\boldsymbol{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}}}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals

Normalize raw step-size $m_{f,\tau}^{\mu}$ by **loudspeaker PSD** estimate $\hat{\Psi}_{f,\tau}^{\mathrm{XX}}$

Proposed Step-Size Structure

$$\left[\boldsymbol{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}}}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals
- rapid changes due to non-stationarity of interfering signals

Proposed Step-Size Structure

$$\left[\boldsymbol{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}} + \hat{\Psi}_{f,\tau}^{\mathrm{II}}}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals
- rapid changes due to non-stationarity of interfering signals

Traditional Approach: Normalization by interference PSD estimate $\hat{\Psi}_{f,\tau}^{\mathrm{II}}$

Proposed Step-Size Structure

$$\left[\mathbf{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}} + \frac{M}{R}\left|m_{f,\tau}^{e}\left[\boldsymbol{e}_{\tau}\right]_{f}\right|^{2}}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals
- rapid changes due to non-stationarity of interfering signals

Normalization by **masked error power** $\left|m_{f,\tau}^{e}\left[\boldsymbol{e}_{\tau}\right]_{f}\right|^{2}$

Proposed Step-Size Structure

$$\left[\boldsymbol{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}} + \frac{M}{R}\left|m_{f,\tau}^{e}\left[\boldsymbol{e}_{\tau}\right]_{f}\right|^{2}}$$

Challenges: DNN needs to model

- large numerical range due to non-whiteness of loudspeaker signals
- rapid changes due to non-stationarity of interfering signals

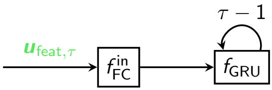
Exploitation of **domain knowledge** from traditional adaptation control

DNN Architecture

$$\xrightarrow{\boldsymbol{u}_{\mathrm{feat},\tau}} \boxed{f_{\mathrm{FC}}^{\mathrm{in}}}$$

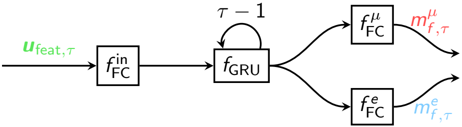
- Feature vector $\boldsymbol{u}_{\mathrm{feat},\tau}$
  - Log. loudspeaker power spectrum
  - Log. error power spectrum
- Extract temporal information by GRU layers
- DNN provides masks $m^{\mu}_{f,\tau} \in [0, \mu_{\mathrm{MAX}}]$ and $m^{e}_{f,\tau} \in [0, 1]$
- Time- and frequency-dependent step-sizes $[\boldsymbol{\Lambda}^{\mathrm{DNN}}_{\tau}]_{ff}$

DNN Architecture



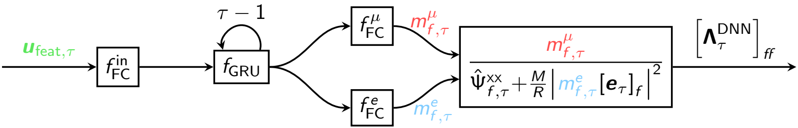
- Feature vector $\boldsymbol{u}_{\text{feat},\tau}$
  - Log. loudspeaker power spectrum
  - Log. error power spectrum
- Extract temporal information by GRU layers
- DNN provides masks $m^{\mu}_{f,\tau} \in [0, \mu_{\text{MAX}}]$ and $m^{e}_{f,\tau} \in [0, 1]$
- Time- and frequency-dependent step-sizes $[\boldsymbol{\Lambda}^{\text{DNN}}_{\tau}]_{ff}$

DNN Architecture



- Feature vector $\boldsymbol{u}_{\text{feat},\tau}$
  - Log. loudspeaker power spectrum
  - Log. error power spectrum
- Extract temporal information by GRU layers
- DNN provides masks $m_{f,\tau}^{\mu} \in [0, \mu_{\text{MAX}}]$ and $m_{f,\tau}^{e} \in [0, 1]$
- Time- and frequency-dependent step-sizes $[\boldsymbol{\Lambda}_{\tau}^{\text{DNN}}]_{ff}$

DNN Architecture

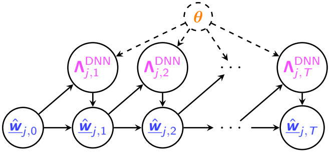


- Feature vector $\boldsymbol{u}_{\text{feat},\tau}$
  - Log. loudspeaker power spectrum
  - Log. error power spectrum
- Extract temporal information by GRU layers
- DNN provides masks $m_{f,\tau}^{\mu} \in [0, \mu_{\text{MAX}}]$ and $m_{f,\tau}^{e} \in [0, 1]$
- Time- and frequency-dependent step-sizes $[\boldsymbol{\Lambda}_\tau^{\text{DNN}}]_{ff}$

DNN Training

Challenges:

- Choice of optimum target step-sizes
- Cost function design



Proposed Solution

End-to-end training of DNN parameters $\boldsymbol{\theta}$ w.r.t. achieved system identification performance

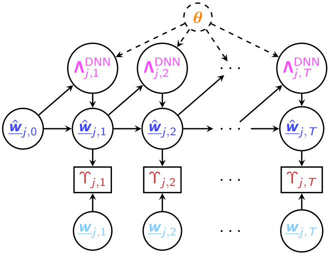
Cost function:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{TJ} \sum_{j=1}^{J} \sum_{\tau=1}^{T} 10 \log_{10} \left(\Upsilon_{j,\tau} \right)$$

with normalized system distance

$$\Upsilon_{j,\tau} = \frac{||\underline{\boldsymbol{w}}_{j,\tau} - \hat{\underline{\boldsymbol{w}}}_{j,\tau}||_2^2}{||\underline{\boldsymbol{w}}_{j,\tau}||_2^2}$$

j: sequence index, τ: block index

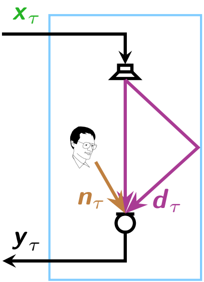


Outline

- Acoustic System Identification
- Proposed DNN-Based Adaptation Control
- Experimental Evaluation
- Conclusion

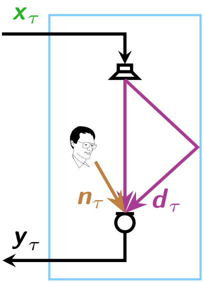
Experimental Evaluation: AEC Application

- Loudspeaker signal: 143 different speakers
- Acoustic environment: 201 measured AIRs
  - Sampling frequency: $f_s = 16$ kHz
  - Reverberation time $T_{60} \in [120\text{ms}, 780\text{ms}]$
  - Acoustic scene change in the interval $[7.2\text{s}, \ 8.8\text{s}]$
- Interfering signal
  - 145 different speakers (echo-to-near-end power ratio $\in [-10\text{dB}, 10\text{dB}]$)
  - White Gaussian noise (echo-to-noise power ratio $\in [25\text{dB}, 35\text{dB}]$)



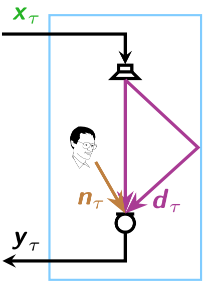
Experimental Evaluation: AEC Application

- Loudspeaker signal: 143 different speakers
- Acoustic environment: 201 measured AIRs
  - Sampling frequency: $f_s = 16$ kHz
  - Reverberation time $T_{60} \in [120\text{ms}, 780\text{ms}]$
  - Acoustic scene change in the interval $[7.2\text{s},\ 8.8\text{s}]$
- Interfering signal
  - 145 different speakers (echo-to-near-end power ratio $\in [-10\text{dB}, 10\text{dB}]$)
  - White Gaussian noise (echo-to-noise power ratio $\in [25\text{dB}, 35\text{dB}]$)



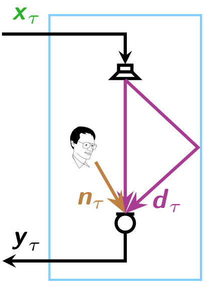
Experimental Evaluation: AEC Application

- Loudspeaker signal: 143 different speakers
- Acoustic environment: 201 measured AIRs
  - Sampling frequency: $f_s = 16$ kHz
  - Reverberation time $T_{60} \in [120\text{ms}, 780\text{ms}]$
  - Acoustic scene change in the interval $[7.2\text{s},\ 8.8\text{s}]$
- Interfering signal
  - 145 different speakers (echo-to-near-end power ratio $\in [-10\text{dB}, 10\text{dB}]$)
  - White Gaussian noise (echo-to-noise power ratio $\in [25\text{dB}, 35\text{dB}]$)



Experimental Evaluation: AEC Application

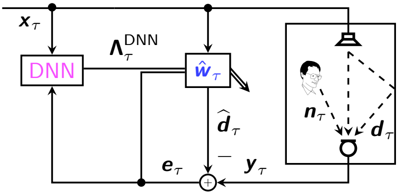
- Loudspeaker signal: 143 different speakers
- Acoustic environment: 201 measured AIRs
  - Sampling frequency: $f_s = 16$ kHz
  - Reverberation time $T_{60} \in [120\text{ms}, 780\text{ms}]$
  - Acoustic scene change in the interval $[7.2\text{s},\ 8.8\text{s}]$
- Interfering signal
  - 145 different speakers (echo-to-near-end power ratio $\in [-10\text{dB}, 10\text{dB}]$)
  - White Gaussian noise (echo-to-noise power ratio $\in [25\text{dB}, 35\text{dB}]$)



Separation into disjoint training and testing data sets

Algorithmic Settings

- FIR filter length: 2048 samples
- Frame shift: 1024 samples
- #DNN parameters: 2.4 million
- Training optimizer: *Adam*



Performance Measures

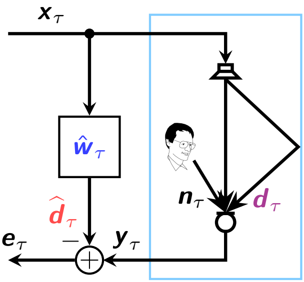
- **Normalized system distance** (the lower the better)

$$\Upsilon_{\mathrm{ZP},\tau} = 10\log_{10}\frac{||\tilde{\boldsymbol{w}}_\tau - \boldsymbol{V}\hat{\boldsymbol{w}}_\tau||_2^2}{||\tilde{\boldsymbol{w}}_\tau||_2^2}$$

with $\boldsymbol{V}$ being a zero-padding matrix

- **ERLE** (the higher the better)

$$\mathcal{E}_\tau = 10\log_{10}\frac{\mathbb{E}\left[||\boldsymbol{d}_\tau||_2^2\right]}{\mathbb{E}\left[||\boldsymbol{d}_\tau - \widehat{\boldsymbol{d}}_\tau||_2^2\right]}$$



Performance Measures

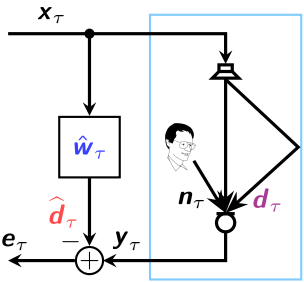
- **Normalized system distance** (the lower the better)

$$\Upsilon_{\text{ZP},\tau} = 10 \log_{10} \frac{||\tilde{\boldsymbol{w}}_\tau - \boldsymbol{V} \hat{\boldsymbol{w}}_\tau||_2^2}{||\tilde{\boldsymbol{w}}_\tau||_2^2}$$

with $\boldsymbol{V}$ being a zero-padding matrix

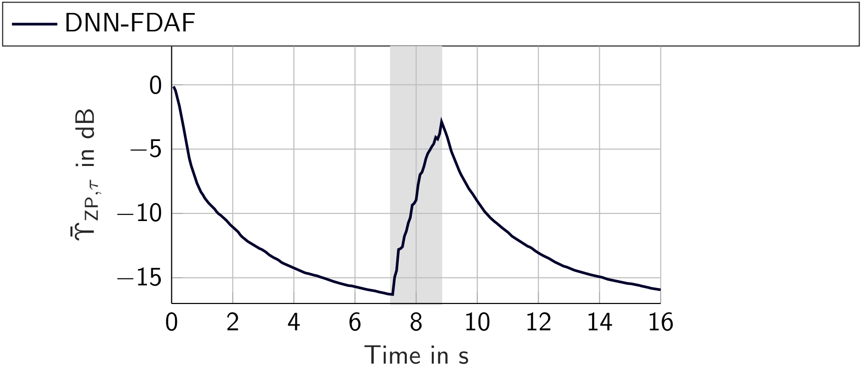
- **ERLE** (the higher the better)

$$\mathcal{E}_\tau = 10 \log_{10} \frac{\mathbb{E}\left[||\boldsymbol{d}_\tau||_2^2\right]}{\mathbb{E}\left[||\boldsymbol{d}_\tau - \widehat{\boldsymbol{d}}_\tau||_2^2\right]}$$



Arithmetic average of **100** experiments with randomly-selected loudspeaker and interfering signals, AIRs and transition times.

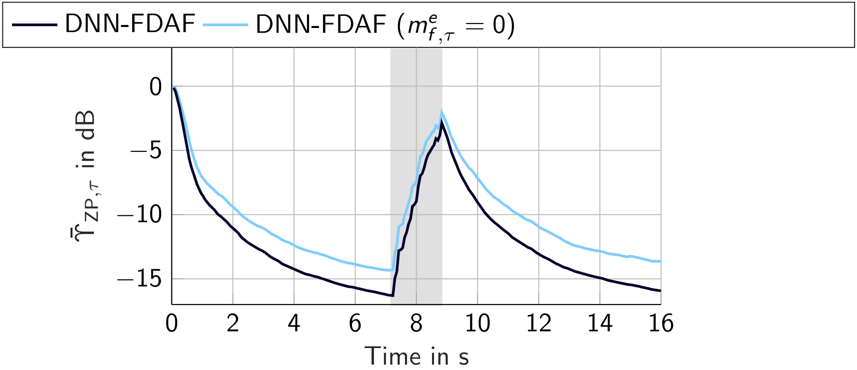
Analysis of Proposed DNN-FDAF



$$\left[\boldsymbol{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m_{f,\tau}^{\mu}}{\hat{\Psi}_{f,\tau}^{\mathrm{XX}} + \frac{M}{R}\left| m_{f,\tau}^{e} \left[\boldsymbol{e}_{\tau}\right]_{f} \right|^{2}}$$

Proposed step-size control

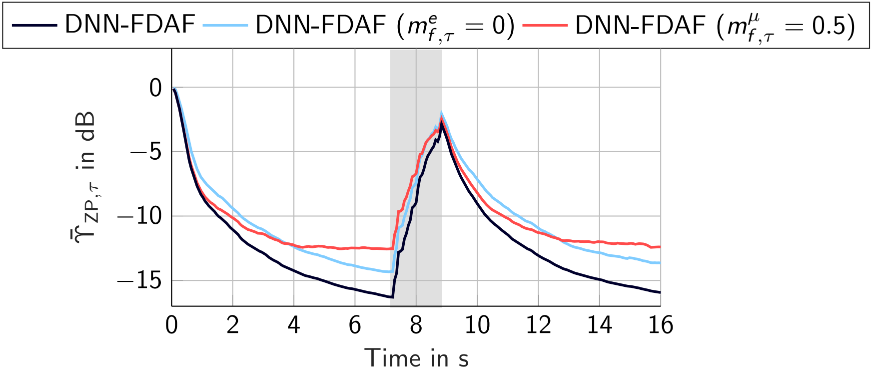
Analysis of Proposed DNN-FDAF



$$\left[\mathbf{\Lambda}_{\tau}^{\mathrm{DNN}}\right]_{ff} = \frac{m^{\mu}_{f,\tau}}{\hat{\Psi}^{\mathrm{XX}}_{f,\tau} + 0}$$

Discarding error power normalization

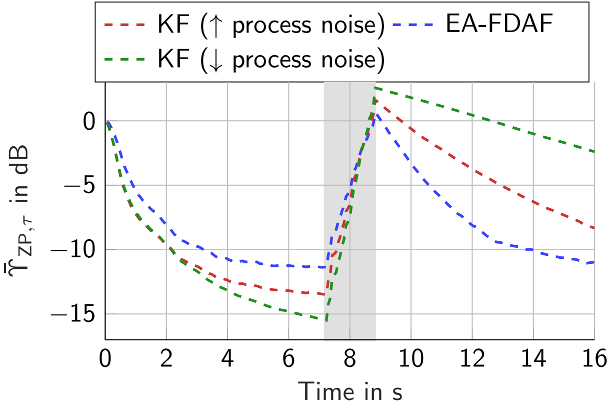
Analysis of Proposed DNN-FDAF



$$\left[\boldsymbol{\Lambda}^{\mathrm{DNN}}_{\tau}\right]_{ff} = \frac{0.5}{\hat{\Psi}^{\mathrm{XX}}_{f,\tau} + \frac{M}{R}\left|m^{e}_{f,\tau}\left[\boldsymbol{e}_{\tau}\right]_{f}\right|^{2}}$$

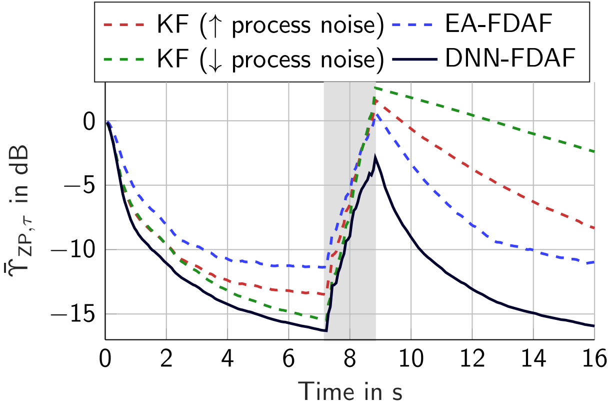
Static and frequency-independent
raw step-size

System Identification Performance



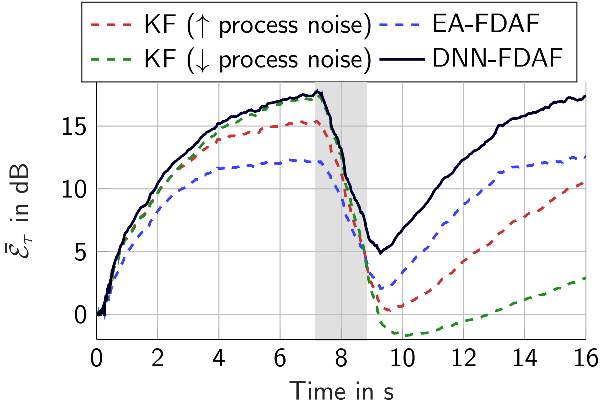
Steady-state and **tracking performance trade-off**

System Identification Performance



DNN-FDAF: **Fast convergence** and **high steady-state** performance

Echo Cancellation Performance



DNN-FDAF: Improved **echo cancellation** performance

Conclusion

Summary

- Novel adaptation control for online system identification by using **DNN-based step-size inference**
- **End-to-end optimization** of DNN parameters w.r.t. average system identification performance

Conclusion

Summary

- Novel adaptation control for online system identification by using **DNN-based step-size inference**
- **End-to-end optimization** of DNN parameters w.r.t. average system identification performance

Outlook

- **Joint control** of system identification and further parts of speech enhancement algorithms, e.g., spectral postfiltering
- Extension to **unsupervised system identification** applications

Conclusion

Summary

- Novel adaptation control for online system identification by using **DNN-based step-size inference**
- **End-to-end optimization** of DNN parameters w.r.t. average system identification performance

Outlook

- **Joint control** of system identification and further parts of speech enhancement algorithms, e.g., spectral postfiltering
- Extension to **unsupervised system identification** applications

Thank you for watching!