Performance Comparison of Single-Microphone Speech Enhancement Using Speech-Distortion Weighted Inter-Frame Wiener Filters

Klaus Brümann, Dörte Fischer, Simon Doclo
Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

Introduction

- Filters for single-microphone speech enhancement can be applied to suppress background noise while preventing speech distortion [1, 2, 3, 4]
- Single- & multi-frame trade-off filters (also known as speech-distortion weighted inter-frame Wiener filters) can increase **noise reduction** while limiting **speech distortion**
- Single-frame filters **disregard correlation** across time frames and/or frequency bins
- Multi-frame filters can **exploit temporal correlation** across present and past frames

IN THIS POSTER

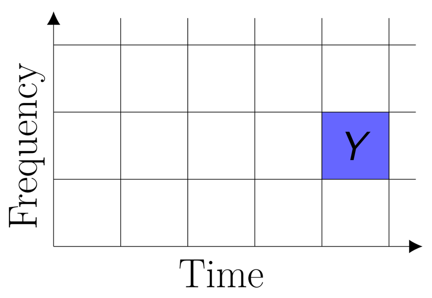
Real- and complex-valued **trade-off filters** are analysed using **low-delay** STFT filterbanks for **speech enhancement** with a single-microphone.

Signal Models

Single-Frame Signal Model

- Assumption: no correlation over time frames and/or frequency bins
- Signal model:

$$\underbrace{Y}_{\text{noisy}} = \underbrace{S}_{\text{speech}} + \underbrace{N}_{\text{noise}}$$



- Estimate speech S by applying gain G to to each noisy speech STFT coefficient Y independently:

$$\hat{S} = GY$$

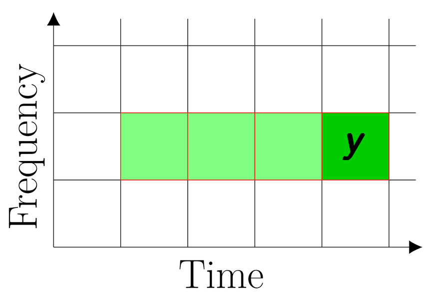
Multi-Frame Signal Model

- Assumption: speech correlated over consecutive time frames
- Speech correlation vector:

$$\boldsymbol{\gamma}_s = \frac{\mathbb{E}[\boldsymbol{s}S^*]}{\mathbb{E}[|S|^2]} = \frac{\boldsymbol{R}_s\boldsymbol{e}}{\phi_S}$$

- Signal model of noisy speech $\boldsymbol{y}$ in terms of speech $\boldsymbol{s}$ and noise $\boldsymbol{n}$ or correlated speech $\boldsymbol{\gamma}_s S$ and undesired $\boldsymbol{u}$:

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} = \boldsymbol{\gamma}_s S + \boldsymbol{u}$$



- Estimate speech S by applying filter $\boldsymbol{h}$ to noisy speech vector:

$$\hat{S} = \boldsymbol{h}^H \boldsymbol{y}$$

Trade-off Filters

$$\text{Less speech distortion} \xleftarrow{\quad} \overset{\textit{Trade-off}}{\underset{\mu}{\longleftrightarrow}} \xrightarrow{\quad} \text{More noise reduction}$$

- **Single-Frame Trade-off Filter:**

$$G = \underset{G}{\operatorname{argmin}} \left\{ \underbrace{\mathbb{E}\left[|GS - S|^2\right]}_{\text{Filtered speech}} + \mu \underbrace{\mathbb{E}\left[|GN|^2\right]}_{\text{Filtered noise}} \right\}$$

Real-Valued Trade-off filter (R-TF)

- Implemented in STFT filterbank with **high frequency resolution** & **high time resolution**
- Parameters are estimated using [5], [6]

$$\boxed{G = \frac{\phi_S}{\mu\phi_N + \phi_S}}$$

- For $\mu = 0$ applies **no noise reduction**
- For $\mu = 1$ is a **Wiener gain**

- **Multi-Frame Trade-off Filter:**

$$\boldsymbol{h}^{\text{MFTF}} = \underset{\boldsymbol{h}}{\operatorname{argmin}} \left\{ \underbrace{\mathbb{E}\left[|\boldsymbol{h}^H\boldsymbol{\gamma}_s S - S|^2\right]}_{\text{Filtered correlated speech}} + \mu \underbrace{\mathbb{E}\left[|\boldsymbol{h}^H\boldsymbol{u}|^2\right]}_{\text{Filtered undesired}} \right\}$$

Complex-Valued Multi-Frame Trade-off Filter (C-MFTF)

- Implemented in STFT filterbank with **low frequency resolution** & **high time resolution**
- Parameters are estimated as in [1]

$$\boxed{\boldsymbol{h}^{\text{C-MFTF}} = \frac{\boldsymbol{R}_y^{-1}\boldsymbol{\gamma}_s}{\boldsymbol{\gamma}_s^H\boldsymbol{R}_y^{-1}\boldsymbol{\gamma}_s} \frac{\phi_S}{\mu\phi_U^{\text{out}} + \phi_S}}$$

- For $\mu = 0$ is a **C-MFMPDR filter**
- For $\mu = 1$ is a **C-MFWF**

Real-Valued Multi-Frame Trade-off Filter (R-MFTF)

- Implemented in STFT filterbank with **high frequency resolution** & **high time resolution**
- Parameters are estimated as in [2]
- Solution is obtained by applying DFT to cost function: $\boldsymbol{g} = \boldsymbol{D}\boldsymbol{h}$ ($\boldsymbol{D}$: DFT matrix)

$$\boxed{\boldsymbol{g}^{\text{R-MFTF}} = \frac{\boldsymbol{\Phi}_y^{-1}\boldsymbol{\Phi}_s\boldsymbol{1}}{\mu + (1-\mu)\boldsymbol{1}^T\boldsymbol{\Phi}_s\boldsymbol{\Phi}_y^{-1}\boldsymbol{\Phi}_s\boldsymbol{1}A}, \quad A = \frac{1}{\boldsymbol{1}^T\boldsymbol{\Phi}_s\boldsymbol{1}}}$$

- Real-valued filter vector $\boldsymbol{g}^{\text{R-MFTF}}$ can be overlapped into high-resolution scalar gain: $G^{\text{R-MFTF}}$
- For $\mu = 0$ is a **R-MFMPDR filter**
- For $\mu = 1$ is a **R-MFWF**

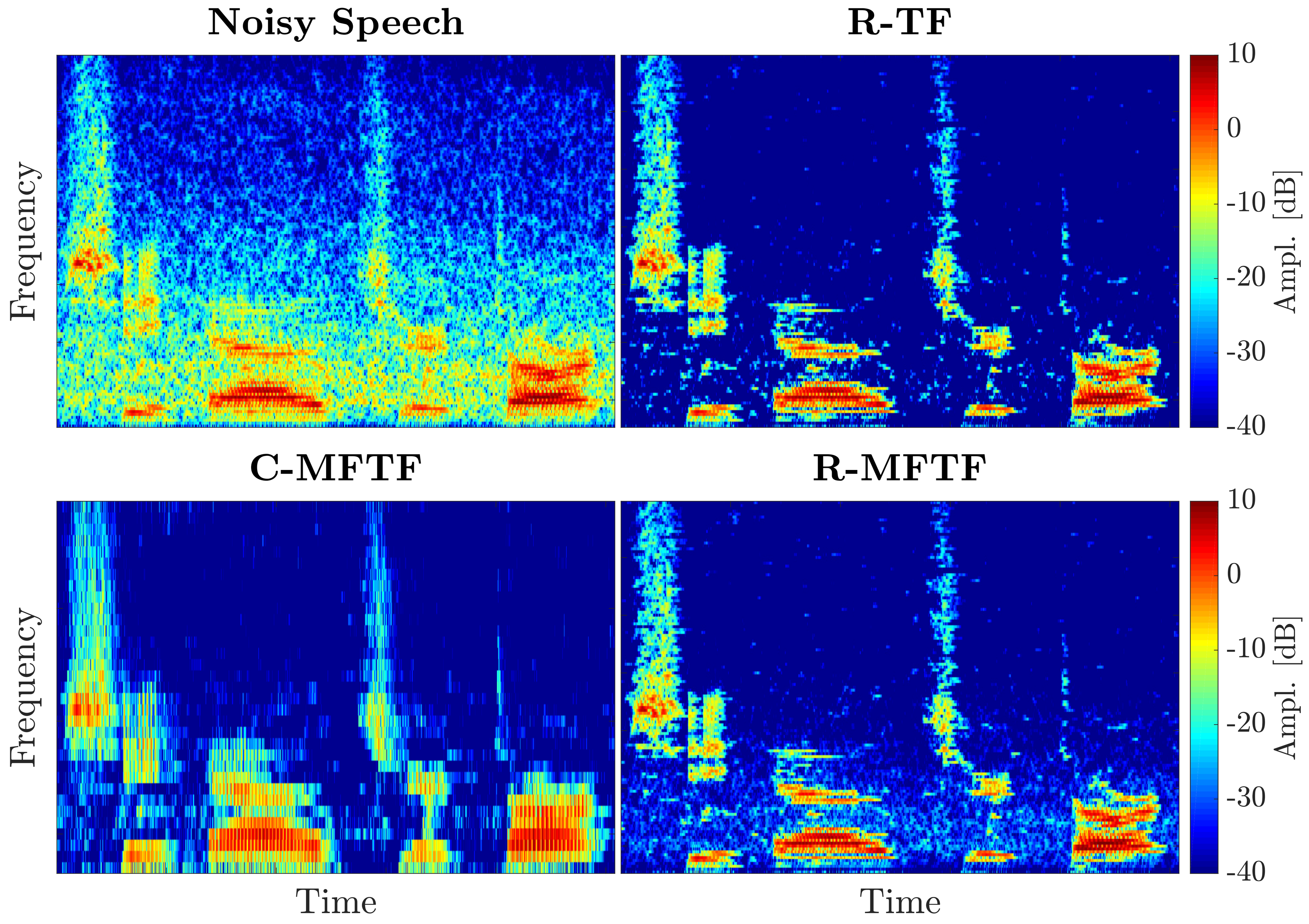
Framework

Filterbank	Analysis window length	Synthesis window length
Low resolution STFT	64 (4ms)	64 (4ms)
High resolution STFT	256 (16ms)	64 (4ms)

- Sampling frequency: 16 kHz
- Both filterbanks use a frame-shift of 16 (1 ms) and have a synthesis delay of 3 ms

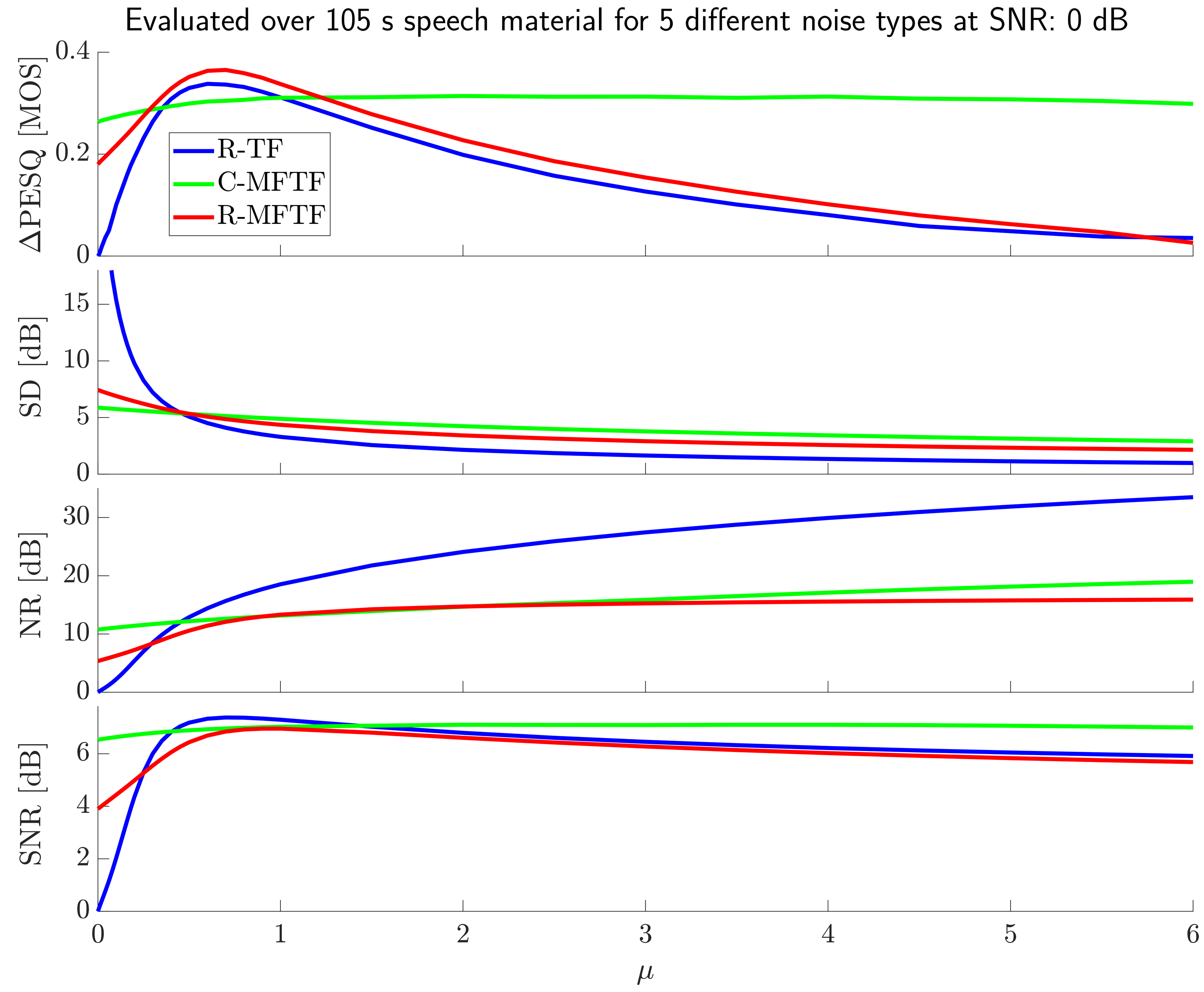
Evaluation

Speech with crossroad noise at SNR: 0 dB. Trade-off parameter $\mu = 0.8$



Results

Objective measures: Perceptual speech quality (PESQ), segmental speech distortion (SD), segmental noise reduction (NR), and segmental signal to noise ratio (SNR). Evaluated over 105 s speech material for 5 different noise types at SNR: 0 dB



- At $\mu = 0$, the C-MFTF (i.e. C-MFMPDR filter) is the most effective filter
- R-MFTF applies best overall PESQ improvement at $\mu = 0.7$
- R-TF can apply highest noise reduction but with high speech distortion

Conclusions

- Complex- and real-valued trade-off filters are effective for speech enhancement
- Increasing trade-off parameter μ increases both noise reduction and speech distortion
- C-MFMPDR filter performs better than R-MFMPDR filter (MFTFs for $\mu = 0$)

References

[1] D. Fischer, K. Brümann, and S. Doclo, "Comparison of parameter estimation methods for Single-Microphone Multi-Frame wiener filtering," in 27th European Signal Processing Conference (EUSIPCO) (Submitted).

[2] K. T. Andersen and M. Moonen, "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction," IEEE Trans. Audio, Speech, Language Process., vol. 26, no. 1, pp. 97–107, Jan. 2018.

[3] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 4, pp. 1256–1269, May 2012.

[4] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," IEEE Trans. Audio, Speech, Language Process., vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443–445, Apr. 1985.

[6] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 4, pp. 1383–1393, May 2012.