Technische Universität Braunschweig

Institut für Nachrichtentechnik

Learning to Dequantize Speech Signals by Primal-Dual Networks: An Approach for Acoustic Sensor Networks

Christoph Brauer*, Ziyue Zhao†, Dirk Lorenz* and Tim Fingscheidt†

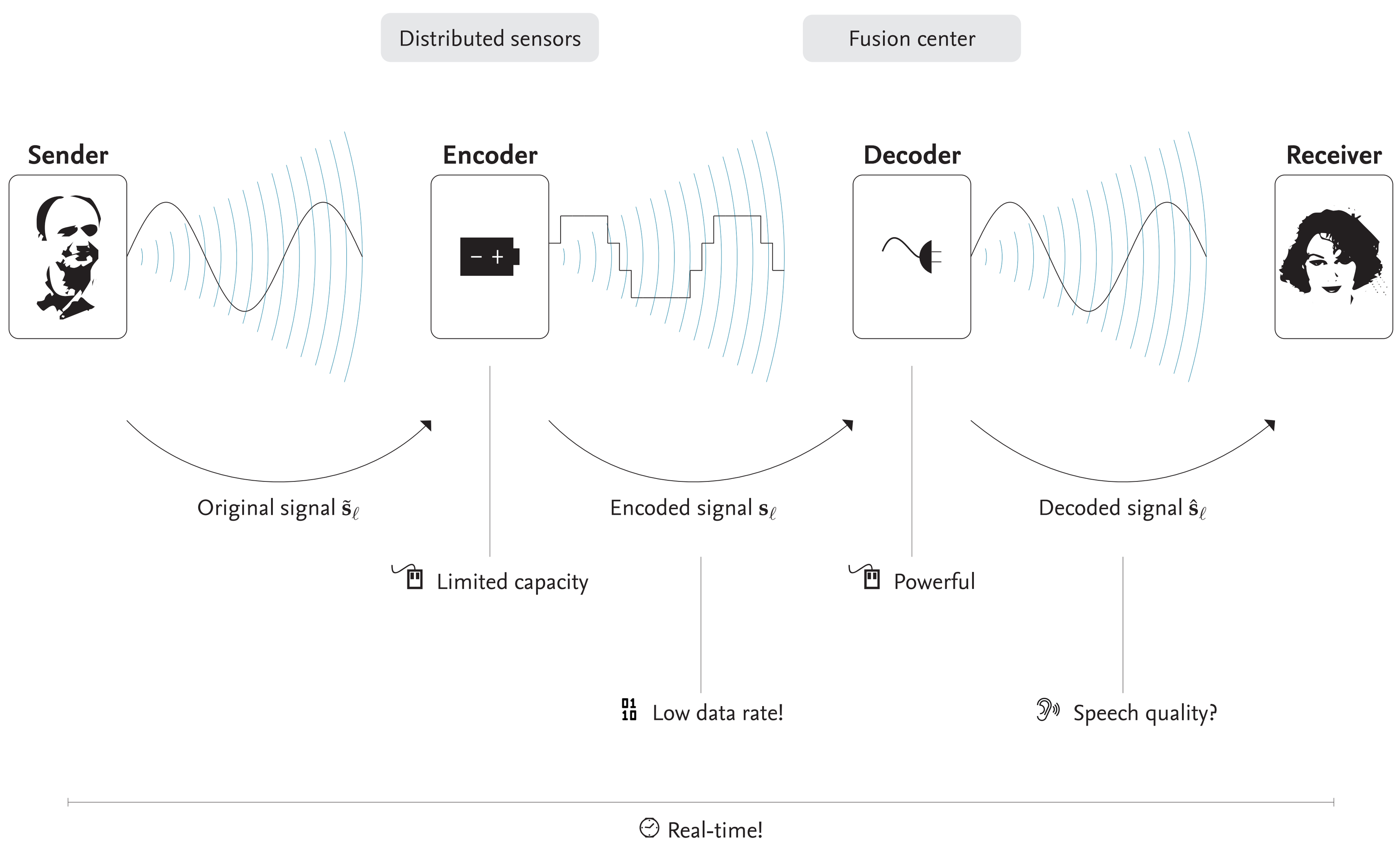
* Institute for Analysis and Algebra | Technische Universität Braunschweig | Braunschweig, Germany

† Institute for Communications Technology | Technische Universität Braunschweig | Braunschweig, Germany

{ch.brauer, ziyue.zhao, d.lorenz, t.fingscheidt}@tu-bs.de

Motivation

- In **acoustic sensor networks**, transmitters must operate with very low computational complexity due to battery lifetime constraints, while some central decoders can consume much higher resources.
- We use short frames $\mathbf{s}_\ell \in \mathbb{R}^N$ of uniformly quantized speech signals to reconstruct associated ground truth frames $\tilde{\mathbf{s}}_\ell \in \mathbb{R}^N$ via **learned sparse reconstruction.** This is done by **unrolling** and learning the parameters of an iterative algorithm applied to the underlying convex optimization problem.
- The **perceptual weighting filter** from code-excited linear predictive (CELP) speech coding is integrated into the loss function of the neural network, achieving **perceptually improved** reconstructed speech.

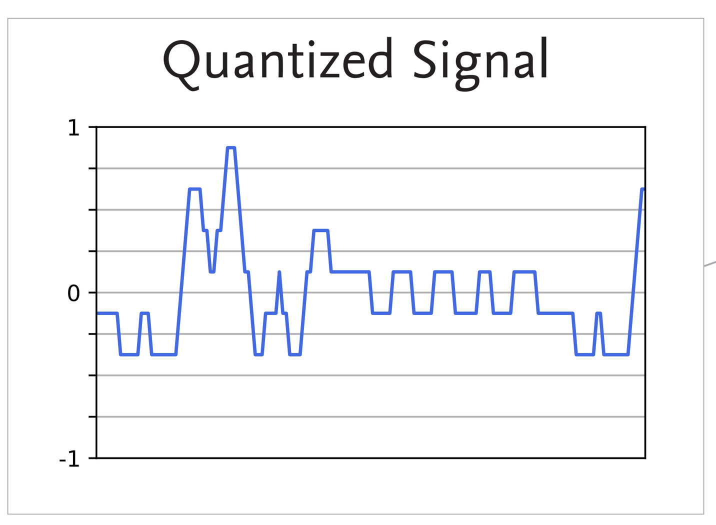


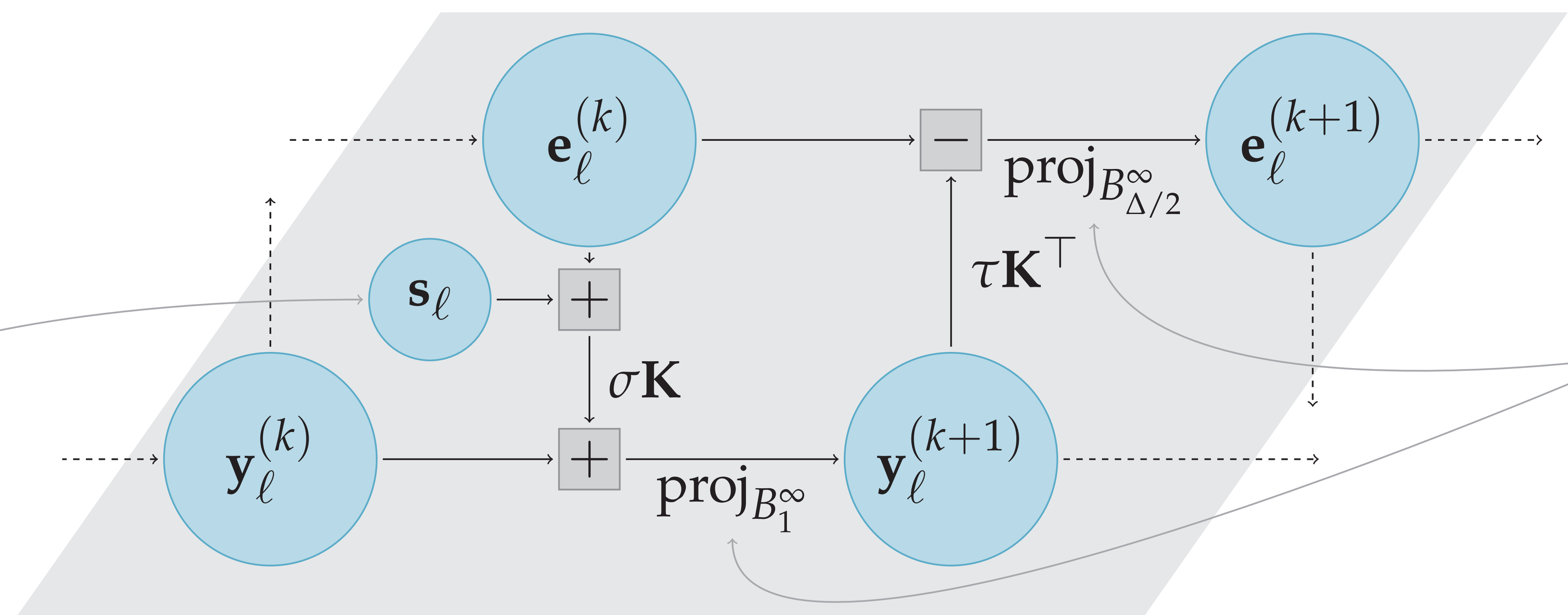
Network Architecture

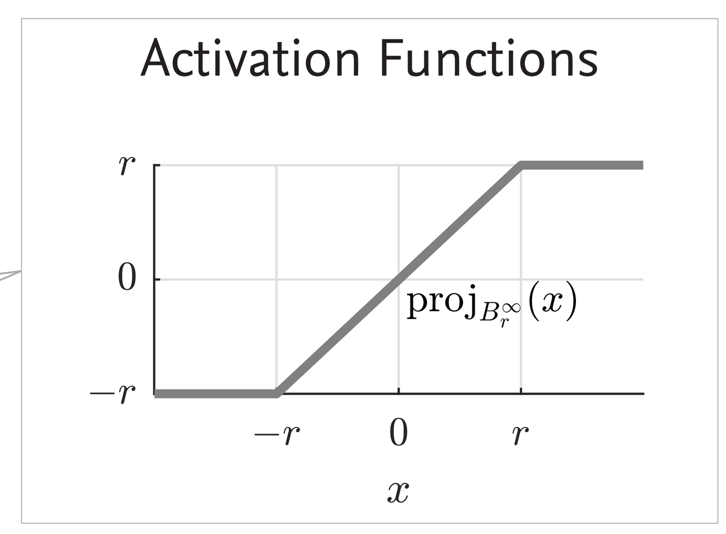
- The residual $\mathbf{e}_\ell = \tilde{\mathbf{s}}_\ell - \mathbf{s}_\ell$ is estimated in terms of the convex optimization problem $\hat{\mathbf{e}}_\ell \in S(\mathbf{K}, \mathbf{s}_\ell) = \arg\min_{\mathbf{e} \in \mathbb{R}^N} \|\mathbf{K}\mathbf{e} + \mathbf{K}\mathbf{s}_\ell\|_1 \quad \text{s.t. } \|\mathbf{e}\|_\infty \leq \frac{\Delta}{2}$.
- As opposed to an earlier approach where the use of $\mathbf{K} = \mathrm{DCT}$ has been shown to lead to perceptually enhanced speech, the matrix $\mathbf{K}$ is not a priori fixed but shall be learned from training data $\{(\mathbf{s}_1, \tilde{\mathbf{s}}_1), \ldots, (\mathbf{s}_m, \tilde{\mathbf{s}}_m)\} \subseteq \mathbb{R}^N \times \mathbb{R}^N$ in this work.
- This gives rise to the bilevel optimization problem $\min_{\mathbf{K} \in \mathbb{R}^{M \times N}} \frac{1}{m} \sum_{\ell=1}^{m} J_\ell(\hat{\mathbf{s}}_\ell, \tilde{\mathbf{s}}_\ell) \quad \text{s.t. } \forall \ell : \hat{\mathbf{s}}_\ell \in \mathbf{s}_\ell + S(\mathbf{K}, \mathbf{s}_\ell)$ where the original optimization problem appears as a lower-level problem in the constraints. The loss functions J_ℓ are data-dependent and play a central role in our approach.
- As solving the bilevel problem directly is potentially hard due to various reasons, we approximate it by replacing $S(\mathbf{K}, \mathbf{s}_\ell)$ with $\mathbf{e}_\ell^{(K)}$ which is defined as the K-th iterate of the Chambolle-Pock algorithm applied to the lower-level problem: Initialize $\mathbf{y}_\ell^{(0)} = \mathbf{0}$ and $\mathbf{e}_\ell^{(0)} = \mathbf{0}$ and compute

$$\mathbf{y}_\ell^{(k+1)} = \mathrm{proj}_{B_1^\infty}\left(\mathbf{y}_\ell^{(k)} + \sigma \mathbf{K}(\mathbf{e}_\ell^{(k)} + \mathbf{s}_\ell)\right)$$
$$\mathbf{e}_\ell^{(k+1)} = \mathrm{proj}_{B_{\Delta/2}^\infty}\left(\mathbf{e}_\ell^{(k)} - \tau \mathbf{K}^\top \mathbf{y}_\ell^{(k+1)}\right)$$

  for $k = 1, \ldots, K$.
- Unrolling the first K iterates of this procedure can be considered a specific recurrent neural network with skip connections and output $\mathbf{e}_\ell^{(K)}$ which makes it possible to minimize the objective $\frac{1}{m} \sum_{\ell=1}^{m} J_\ell(\mathbf{s}_\ell + \mathbf{e}_\ell^{(K)}, \tilde{\mathbf{s}}_\ell)$ via gradient based optimization methods.







Technische Universität Braunschweig

Institut für Nachrichtentechnik

Learning to Dequantize Speech Signals by Primal-Dual Networks: An Approach for Acoustic Sensor Networks

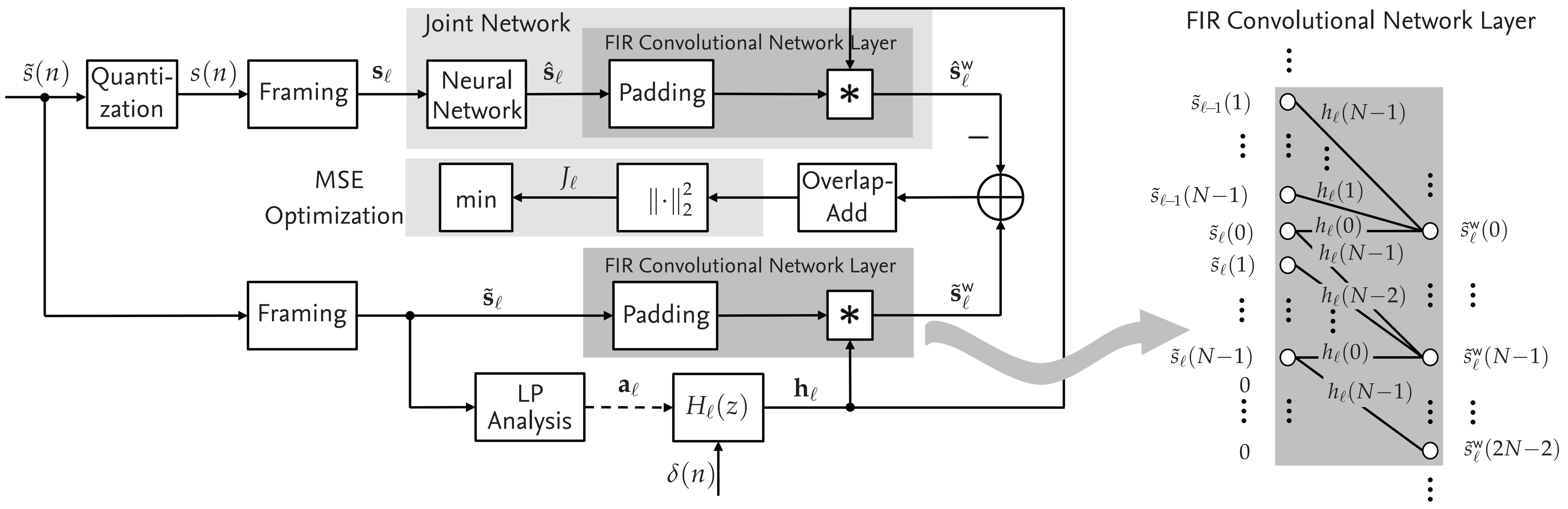
Christoph Brauer*, Ziyue Zhao†, Dirk Lorenz* and Tim Fingscheidt†

* Institute for Analysis and Algebra | Technische Universität Braunschweig | Braunschweig, Germany

† Institute for Communications Technology | Technische Universität Braunschweig | Braunschweig, Germany

{ch.brauer, ziyue.zhao, d.lorenz, t.fingscheidt}@tu-bs.de

A Loss Function Applying the Perceptual Weighting Filter



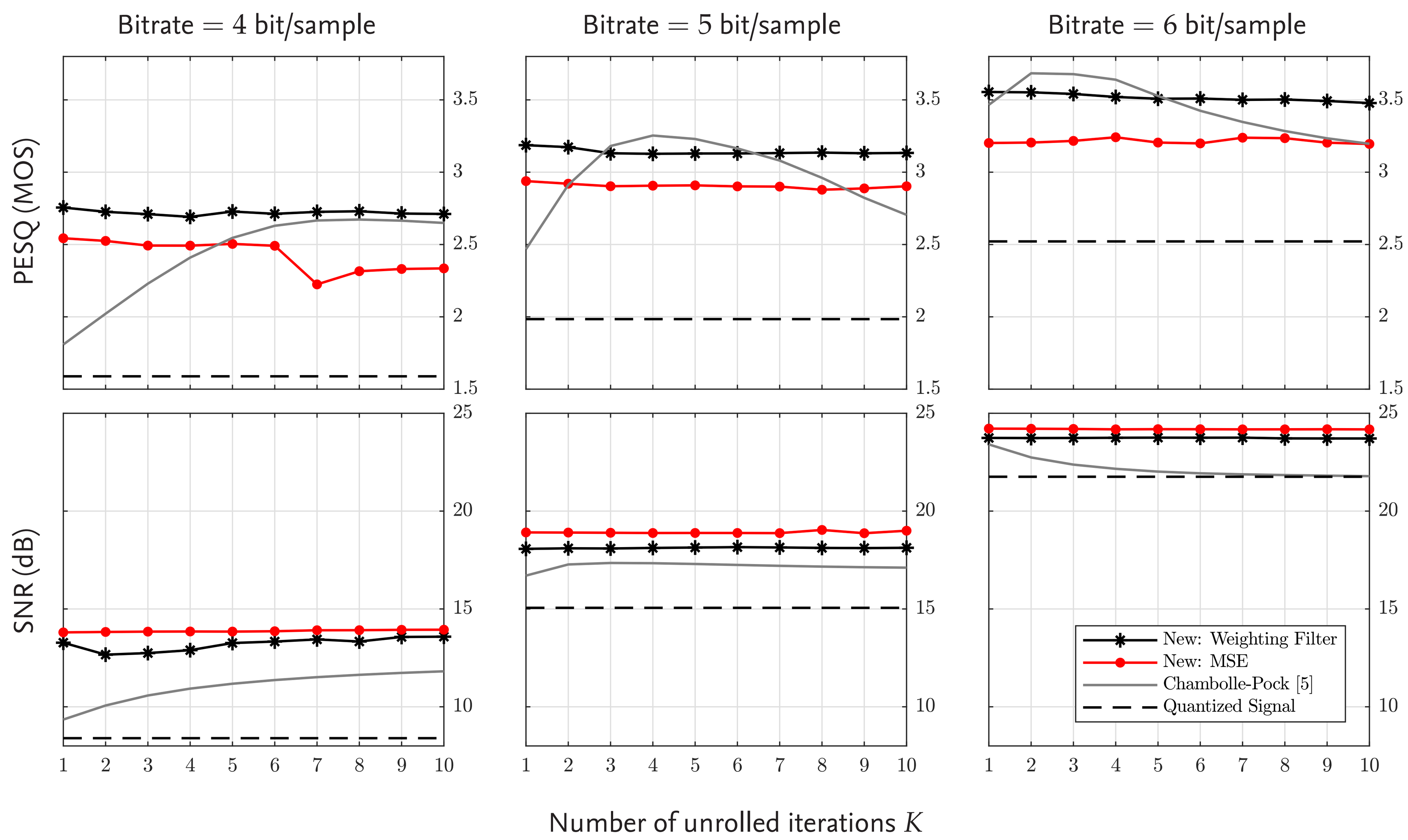
- $\tilde{s}(n)$: Original speech
- $\tilde{\mathbf{s}}_\ell$: Original speech frame
- $\tilde{\mathbf{s}}^{\mathrm{w}}_\ell$: Original weighted speech frame
- $s(n)$: Quantized speech
- $\mathbf{s}_\ell$: Quantized speech frame
- $\hat{\mathbf{s}}_\ell$: Reconstructed speech frame
- $\hat{\mathbf{s}}^{\mathrm{w}}_\ell$: Reconstructed weighted speech frame
- $\mathbf{a}_\ell$: LPC coefficients
- $\mathbf{h}_\ell$: Finite impulse response (FIR) of the weighting filter
- N: Frame length
- $\delta(n)$: $\delta(0) = 1$ and $\delta(n) = 0$ if $n \neq 0$
- OLA: Overlap-add operation

- Perceptual weighting filter in CELP speech coding [1]
  - The weighting filter $H_\ell(z) = \frac{A_\ell(z/\gamma_1)}{A_\ell(z/\gamma_2)}$: $A_\ell(z/\gamma) = \sum_{i=0}^{16} a_\ell(i)\gamma^i z^{-i}$ and $\gamma_1 = 0.94$, $\gamma_2 = 0.6$.
  - The **inverse** weighting filter has similarities to the structure of the clean speech spectral envelope.
- Loss function in neural network training
  - $\mathbf{h}_\ell$ is obtained by filtering the delta function $\delta(n)$ with $H_\ell(z)$.
  - Final loss function $J_\ell(\hat{\mathbf{s}}_\ell, \tilde{\mathbf{s}}_\ell) = \|\mathrm{OLA}((\hat{\mathbf{s}}_\ell - \tilde{\mathbf{s}}_\ell) * \mathbf{h}_\ell)\|_2^2$.
- **Less audible reconstruction error:** Minimization of the weighted error $\rightarrow$ the weighted error becomes spectrally white $\rightarrow$ final (unweighted) error follows the **inverse** weighting filter and is kept at some level below $\rightarrow$ exploiting the masking property of human ear.

[1] 3GPP, Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions (3GPP TS 26.090, Rel. 14), 3GPP; TSG SA, Mar. 2017.

Experiments

- Our test set includes 108 sentences from the IEEE corpus consisting of male speech and sampled at 16 kHz.
- We train networks with different numbers K of unrolled iterations using 612 sentences from the same corpus. These are compared to plain Chambolle-Pock with fixed $\mathbf{K} = \mathrm{DCT}$ in terms of PESQ and SNR.
- In addition to $\mathbf{K}$, we also learn the step sizes σ and τ.
- To learn the parameters, we experiment with two loss functions. On the one hand, we use the MSE and on the other hand the weighting filter based loss.
- To minimize the respective losses, we perform 3000 epochs of stochastic gradient descent using Adam with standard parameters and learning rate 10^{-4}.



Conclusions

- Networks trained with MSE loss are best in terms of SNR, while networks trained with the weighting filter based loss are best in terms of PESQ.
- Best results are already obtained when using $K = 1$ which is clearly favorable in terms of **realtime applicability** of the trained networks.
- The designed loss applying the weighting filter for neural network training is **perceptually efficient** to improve the reconstructed speech quality.