Microsoft
Microsoft Research

Effect of Noise Suppression Losses on Speech Distortion and ASR Performance

Sebastian Braun, Hannes Gamper

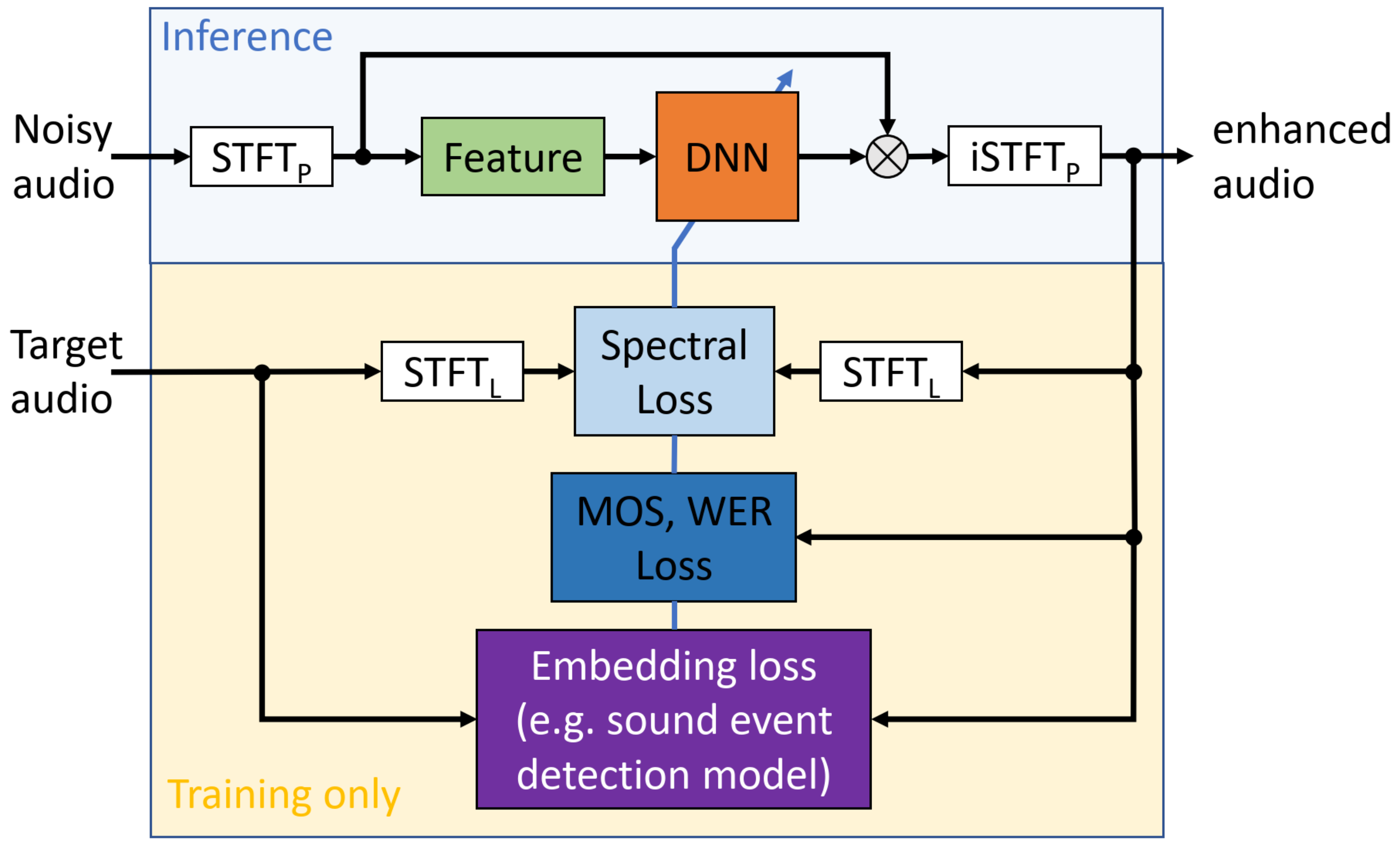
Problem description

While neural networks brought tremendous improvements to speech enhancement in terms of signal separation and suppression, especially smaller networks often introduce speech distortion artifacts, harming signal quality and ASR performance. This work's contributions:

- We show that end-to-end optimization and loss decoupling for frequency-domain networks can be beneficial.
- We reveal insights and tradeoffs of the widely used spectral complex compressed loss.
- We investigate pre-trained models on various tasks for loss improvement:
  - ➢ Reference-free MOS and WER predictors
  - ➢ Embedding distance losses using ASR and sound event detection models

System overview

- Convolutional Recurrent Speech Enhancement (CRUSE) network architecture [1] using complex compressed input features, predicting complex enhancement filter (12.8M MACs/frame, 8.4M parameters)
- End-to-end optimization by decoupling processing STFT and loss frequency analysis (using possibly different parameters).



Loss functions

Base-term (spectral complex compressed MSE):

$$\mathcal{L}_{SD} = \frac{1}{\sigma_s^c}\left(\lambda \sum_{\kappa,\eta} \left|S^c - \widehat{S}^c\right|^2 + (1-\lambda) \sum_{\kappa,\eta} \left||S|^c - |\widehat{S}|^c\right|^2\right)$$

- Additional cepstral distance term:

$$\mathcal{L}_{CD} = \beta \mathcal{L}_{SD}(s,\widehat{s}) + (1-\beta)\mathcal{L}_{CD}(s,\widehat{s})$$

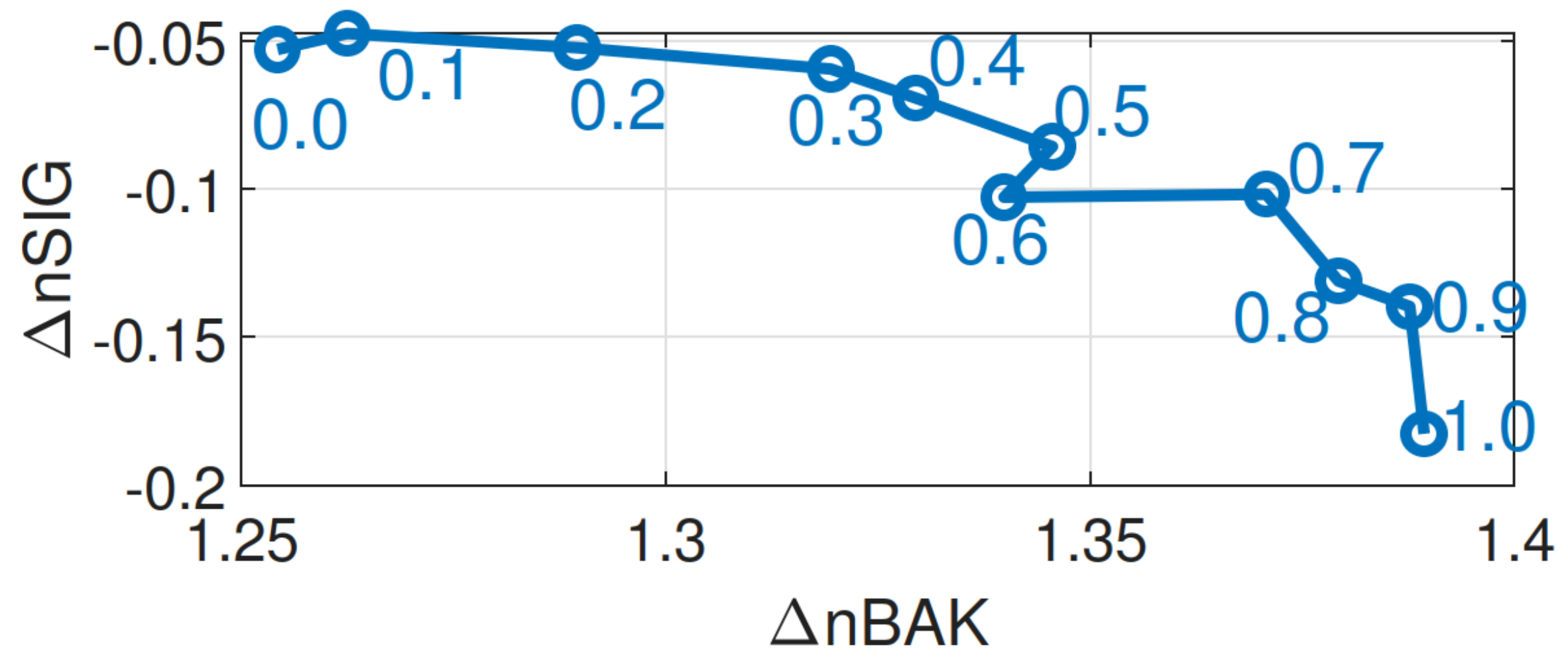
- Weighting with predictors for MOS [2] or WER [3]

$$\mathcal{L}_{MOS,WER} = \sum_b \frac{nWER(\widehat{s}_b)}{nMOS(\widehat{s}_b)} \mathcal{L}_{SD}(s_b,\widehat{s}_b)$$

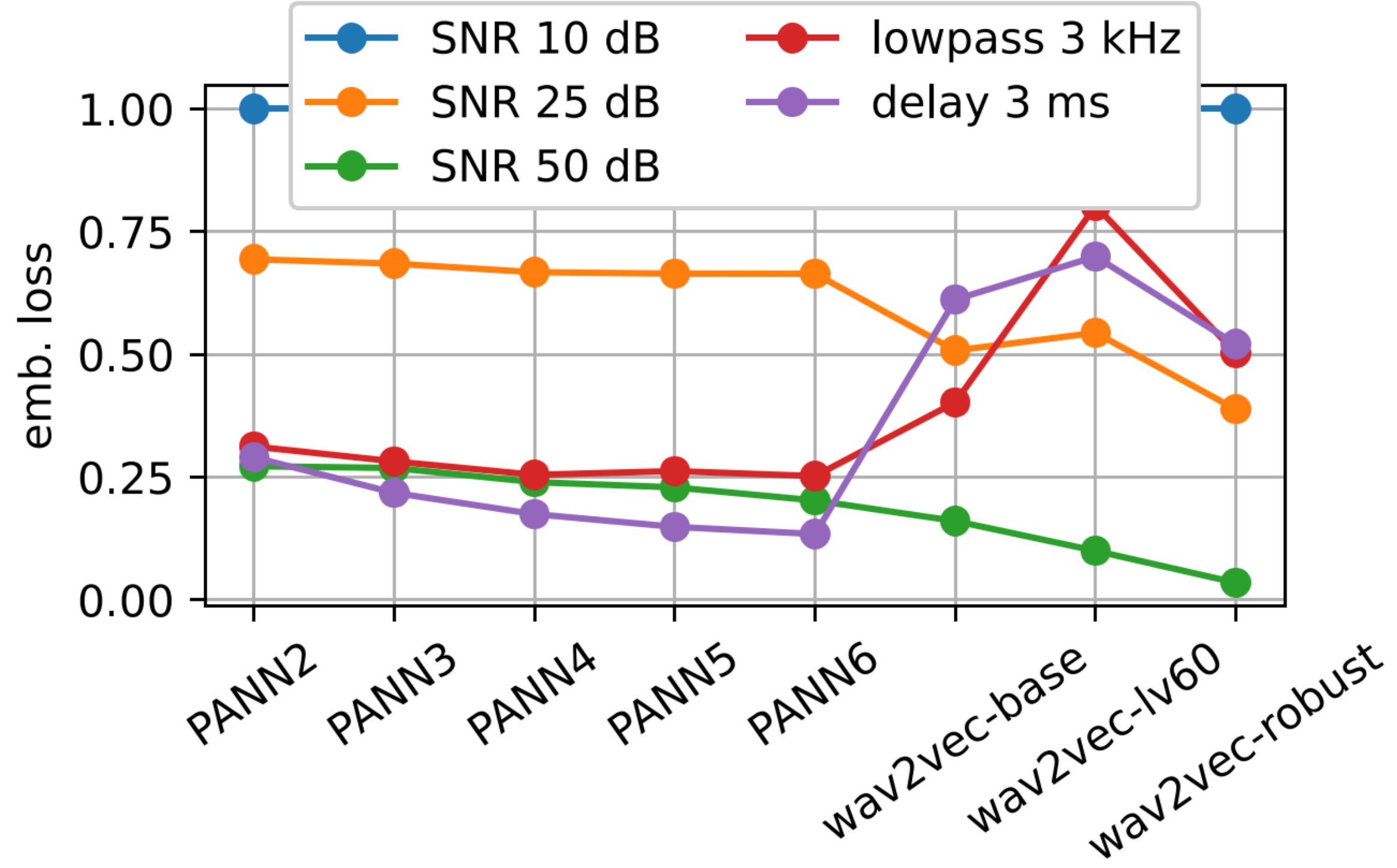
- Additional distance loss using embeddings $\mathbf{u}$ pretrained on ASR (wav2vec) or sound event detection (PANN [4]):

$$\mathcal{L}_{\text{emb}} = \sum_b \mathcal{L}_{SD}(s_b,\widehat{s}_b) + \gamma \frac{\|\mathbf{u}(s_b) - \mathbf{u}(\widehat{s}_b)\|_p}{\|\mathbf{u}(s_b)\|_p}$$

Analysis



Reducing speech distortion with the magnitude loss term (λ→0)



Sensitivity of embedding distances to various degradations.

Results

- Training data: Large-scale DNS challenge data with various online augmentations (>9 years).
- Testsets: 3rd DNS challenge (real recordings), additional High-Quality (HQ) and meeting recordings for ASR evaluation.
- Metrics: P.835 DNSMOS (nSIG, nBAK, nOVL) and WER of using commercial ASR system.

loss	nSIG	nBAK	nOVL	WER (%)		
dataset		DNS		DNS	HQ	meet
noisy	3.87	3.05	3.11	27.9	5.7	16.5
$\mathcal{L}_{SD}^{20}$ (20 ms, 50%)	3.77	4.23	3.50	31.0	5.9	18.7
$\mathcal{L}_{SD}^{32}$ (32 ms, 50%)	3.79	4.26	3.53	30.6	5.9	18.6
$\mathcal{L}_{SD}^{64}$ (64 ms, 75%)	**3.79**	**4.28**	**3.54**	**30.1**	5.9	18.4
$\mathcal{L}_{SD}^{64}$-ERB	3.73	4.22	3.46	31.9	6.0	18.6
$\mathcal{L}_{SD}^{64}$-CD (5)	**3.79**	4.26	**3.53**	30.4	**5.8**	**18.1**
$\mathcal{L}_{SD}^{64}$-MOS (6)	3.78	**4.27**	**3.53**	30.2	6.0	**18.0**
$\mathcal{L}_{SD}^{64}$-WER (6)	**3.79**	**4.27**	**3.53**	30.5	**5.8**	18.2
$\mathcal{L}_{SD}^{64}$-MOS-WER (6)	**3.79**	4.26	**3.53**	**30.1**	**5.8**	18.4
$\mathcal{L}_{SD}^{64}$-PANN$_4$ (7)	**3.79**	**4.27**	**3.54**	30.4	**5.8**	18.5
$\mathcal{L}_{SD}^{64}$-wav2vec (7)	**3.79**	4.26	**3.53**	30.3	**5.9**	18.6

Conclusions

- The magnitude term in the spectral complex compressed loss trades off noise reduction for less speech distortion.
- Loss decoupling boosts performance by increasing spectral loss frequency resolution
- Not all pre-trained embeddings provide useful additional information to an already strong spectral loss.
- Our experiments on large scale training data, real-world test data and strongly indicative metrics did not show significant benefits from using additional pre-trained loss terms.

[1] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," IEEE ICASSP 2021
[2] H. Gamper, C.K.A. Reddy, R. Cutler, I. Tashev, J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in 2019 IEEE WASPAA, 2019.
[3] H. Gamper, D. Emmanilidou, S. Braun, and I. Tashev, "Predicting word error rate for reverberant speech," in Proc. IEEE ICASSP 2020
[4] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE TASLP vol. 28, 2020.