DeepFilterNet: A Low Complexity Speech Enhancement Framework based on Deep Filtering

Hendrik Schröter[1], Alberto N. Escalante-B.[2], Tobias Rosenkranz[2], Andreas Maier[1]

[1]Pattern Recognition Lab, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germay
[2]WS Audiology, Research and Development, Erlangen, Germany

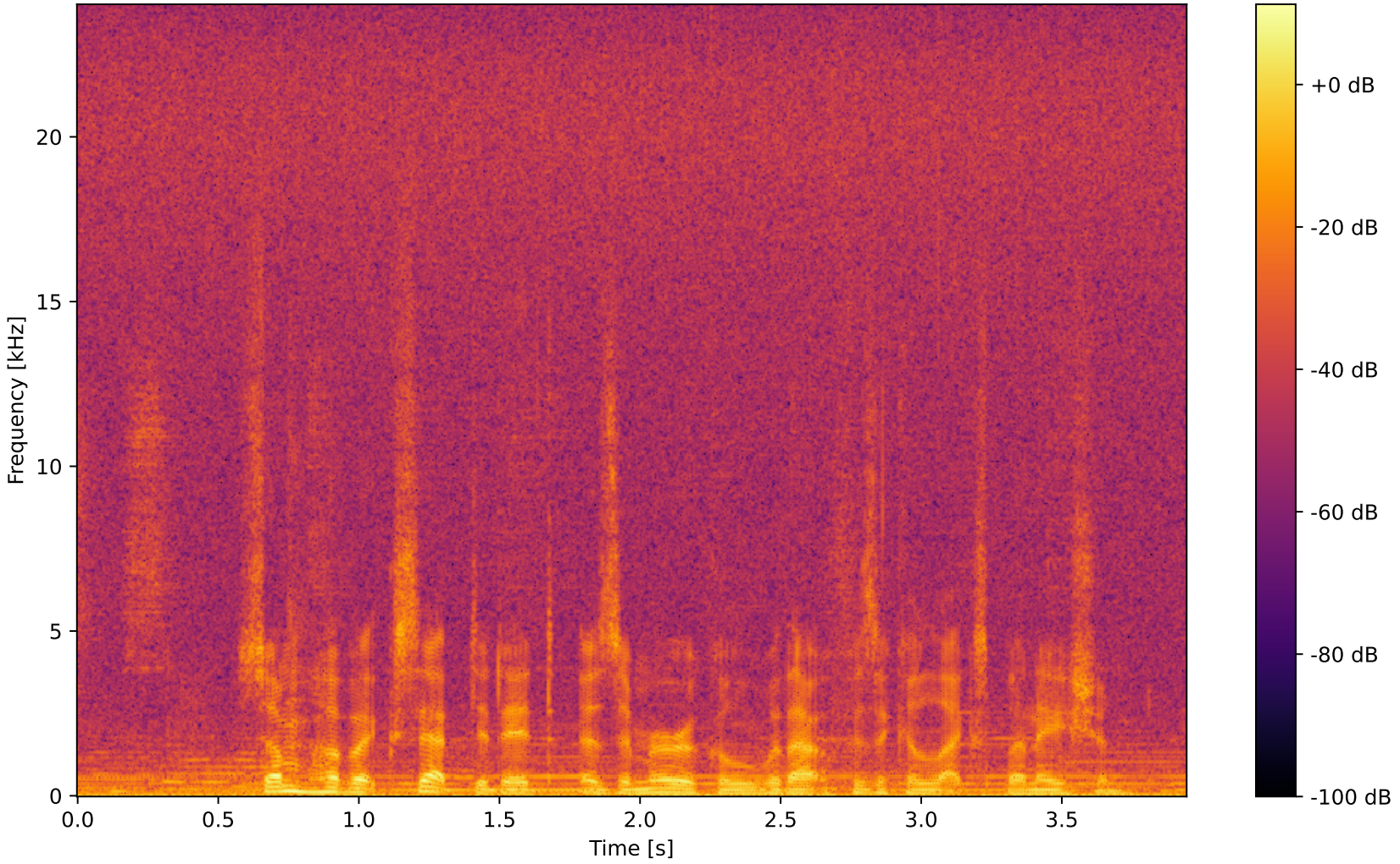
ICASSP 2022, Singapore

DeepFilterNet Contributions

- Perceptual speech enhancement approach
- Full band: 48 kHz
- Real-Time Capabilities: Supported latencies from 5 ms to 40 ms
- Low Complexity: 1.8 M Parameters, 0.35 GMACS
- Open Source: `github.com/Rikorose/DeepFilterNet`
- Demo: `https://huggingface.co/spaces/hshr/DeepFilterNet`

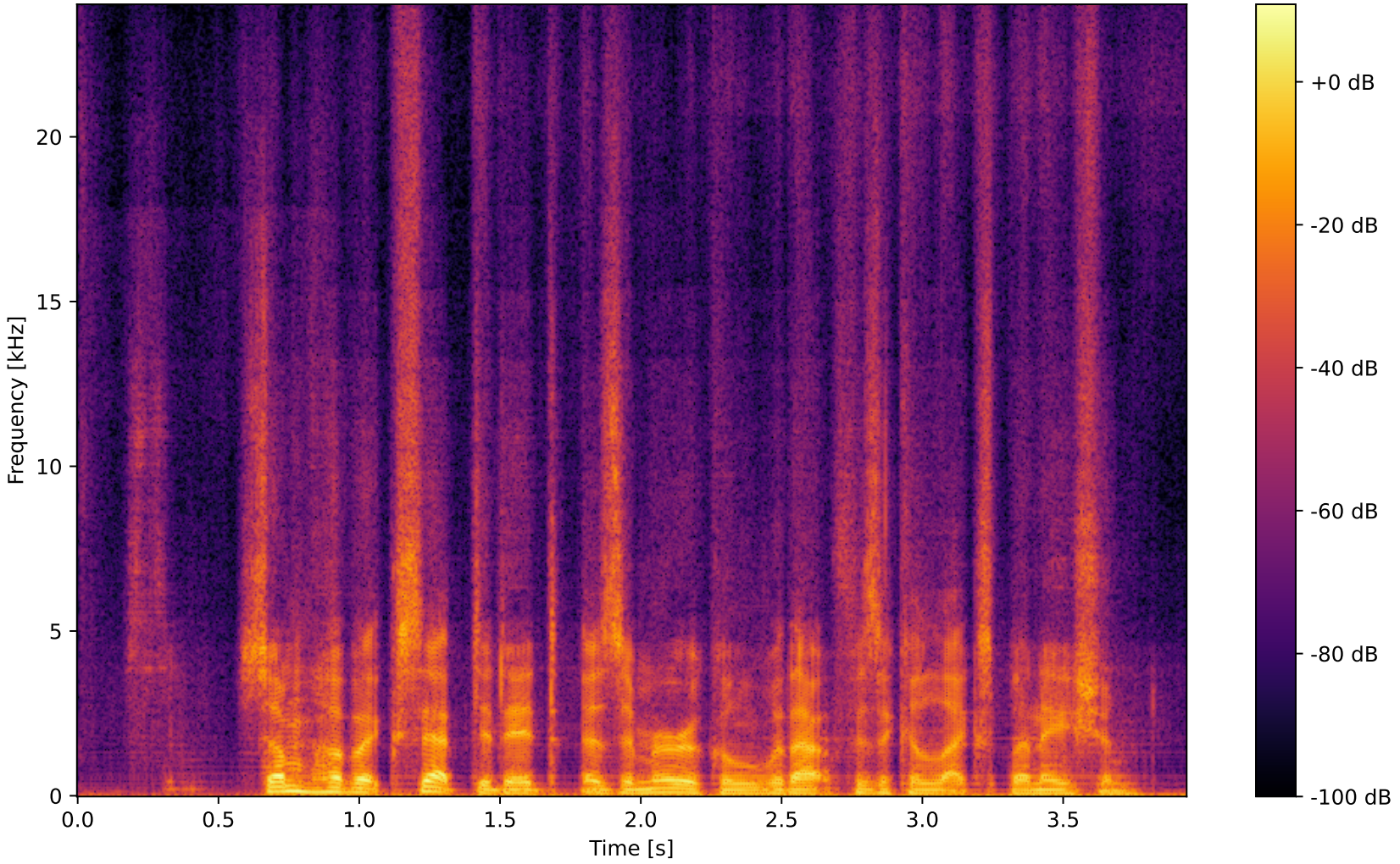
Perceptual Approach

Speech: Periodic + stochastic component



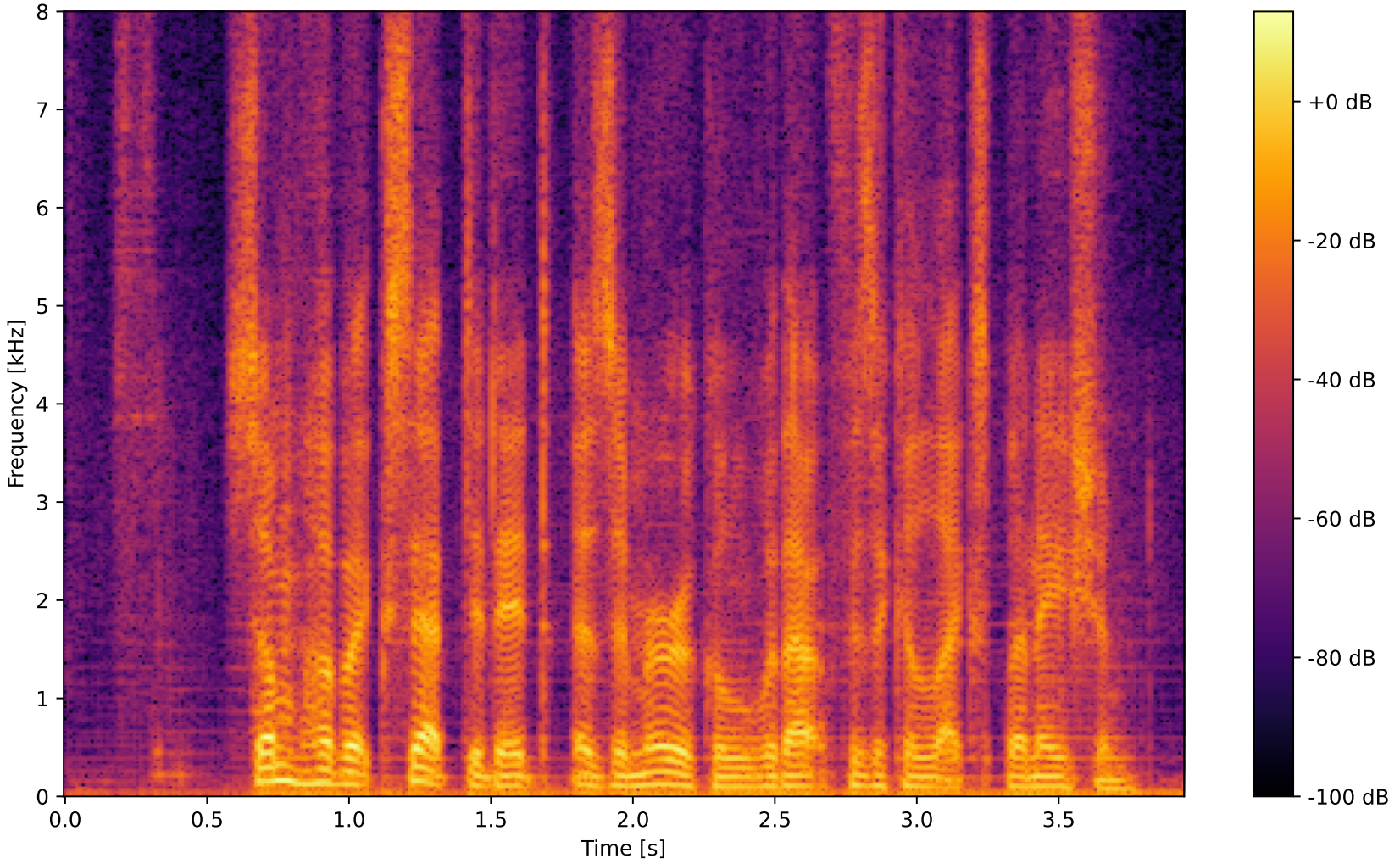
Perceptual Approach

Stage 1: Enhance speech envelope



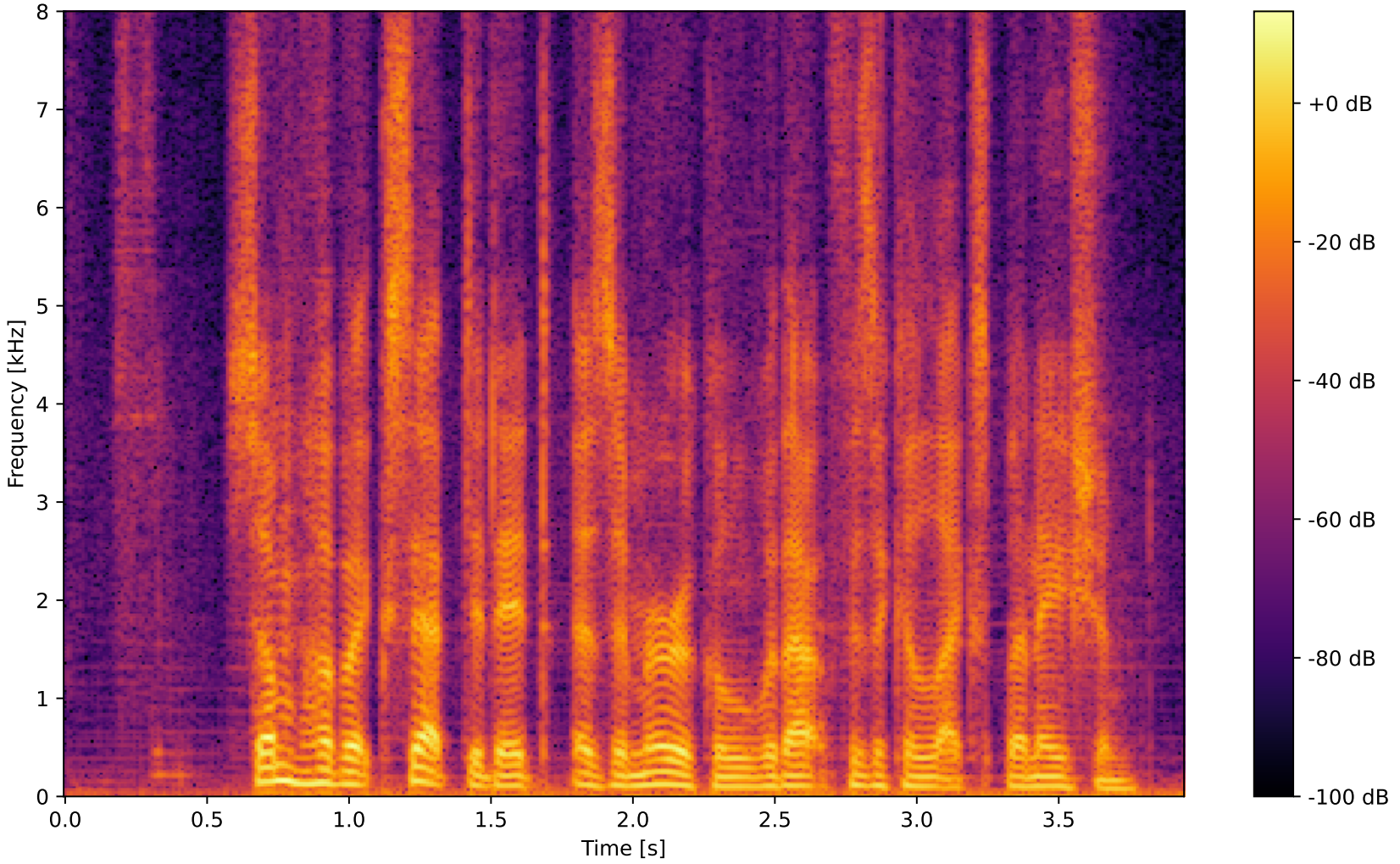
Perceptual Approach

Stage 1: Enhance speech envelope



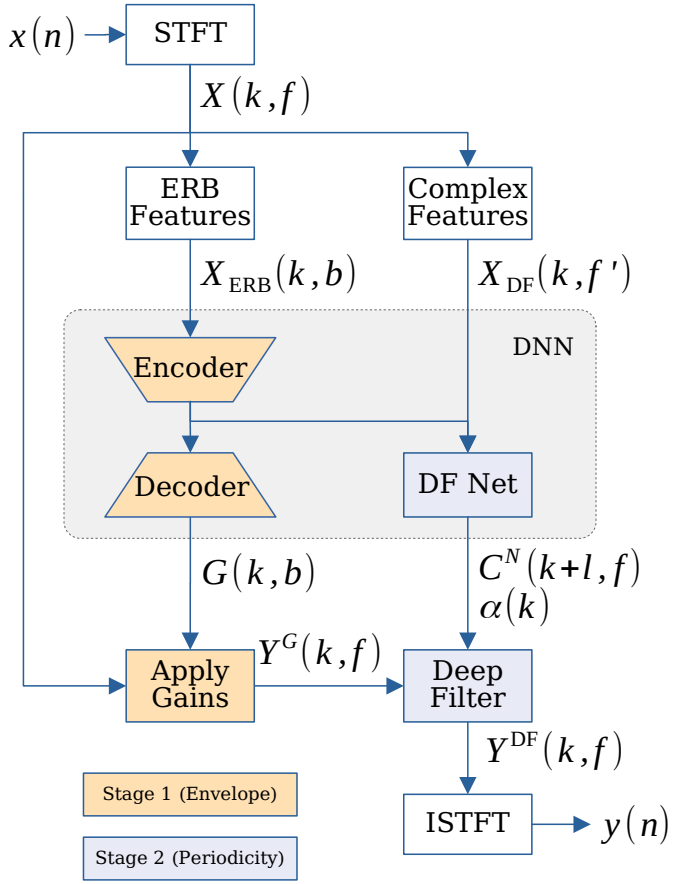
Perceptual Approach

Stage 2: Deep filtering to enhance periodic component



Framework Overview

- Framework operates in frequency domain (FD)
- Support for various FFT and hop sizes depending on latency requirements
- Stage 1 enhances overall speech envelope
  - Operates logarithmically scaled magnitude features (ERB)
  - Only 32 ERB bins allow to construct a small network
  - U-Net like convolutional encoder/decoder structure
- Stage 2 enhances speech periodicity
  - Operates in complex FD up to a frequency F_{DF}
  - Deep Filtering enhances periodic speech components



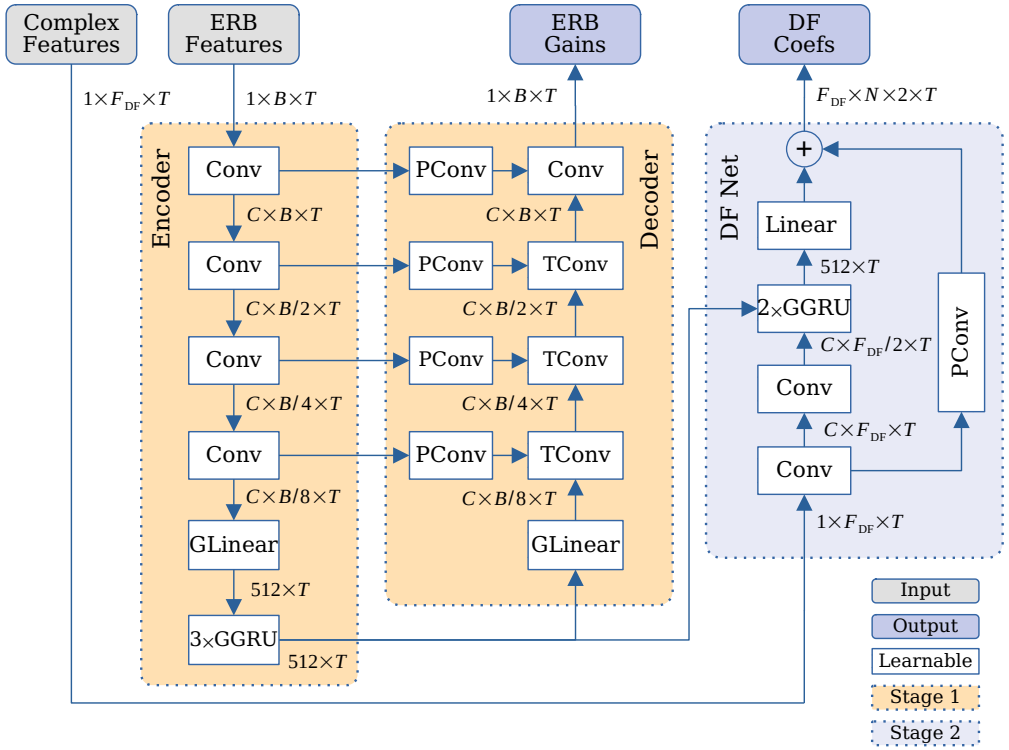
Network and Training Details

- Efficient & standard DNN layers (conv, bn, relu)
- Small DNN I/O
- DNN Sparseness via grouping [1]
- 1.8 M Parameters, 0.35 GMACS

Training Details:

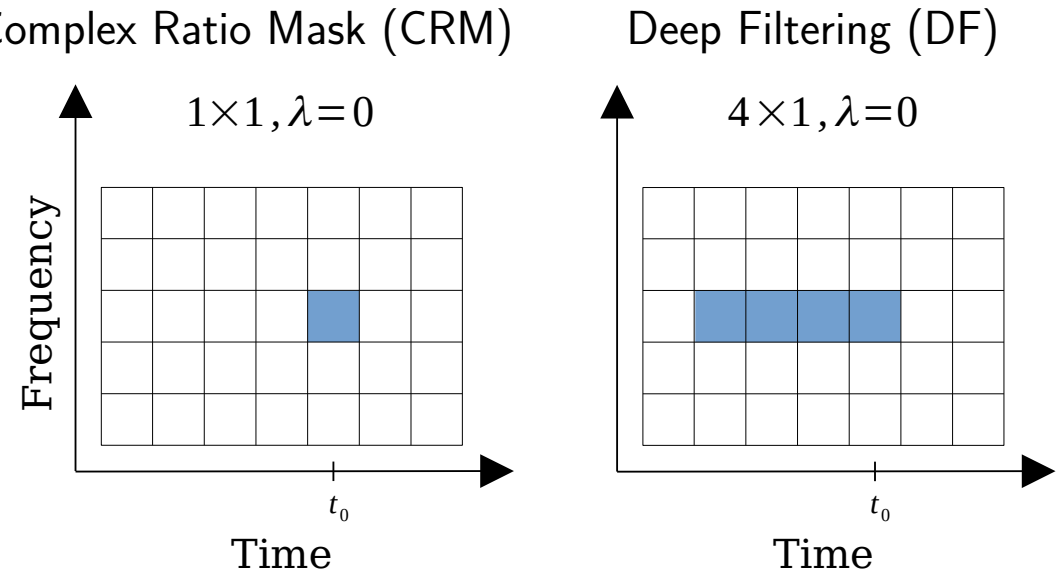
- Trained on DNS 3 dataset [2]
- 32 ERB bands for stage 1 I/O
- Upper frequency for stage 2 DF: $f_{\mathrm{DF}} = 5\,\mathrm{kHz}$
- Loss on compressed complex spectrum:

$$\mathcal{L}_{spec} = \Sigma_{k,f} \,||\, |Y|^c - |S|^c||^2 + \Sigma_{k,f} \,||\, |Y|^c e^{j\varphi_Y} - |S|^c e^{j\varphi_S}||^2$$



Deep Filtering

Originally proposed by [3], [4]

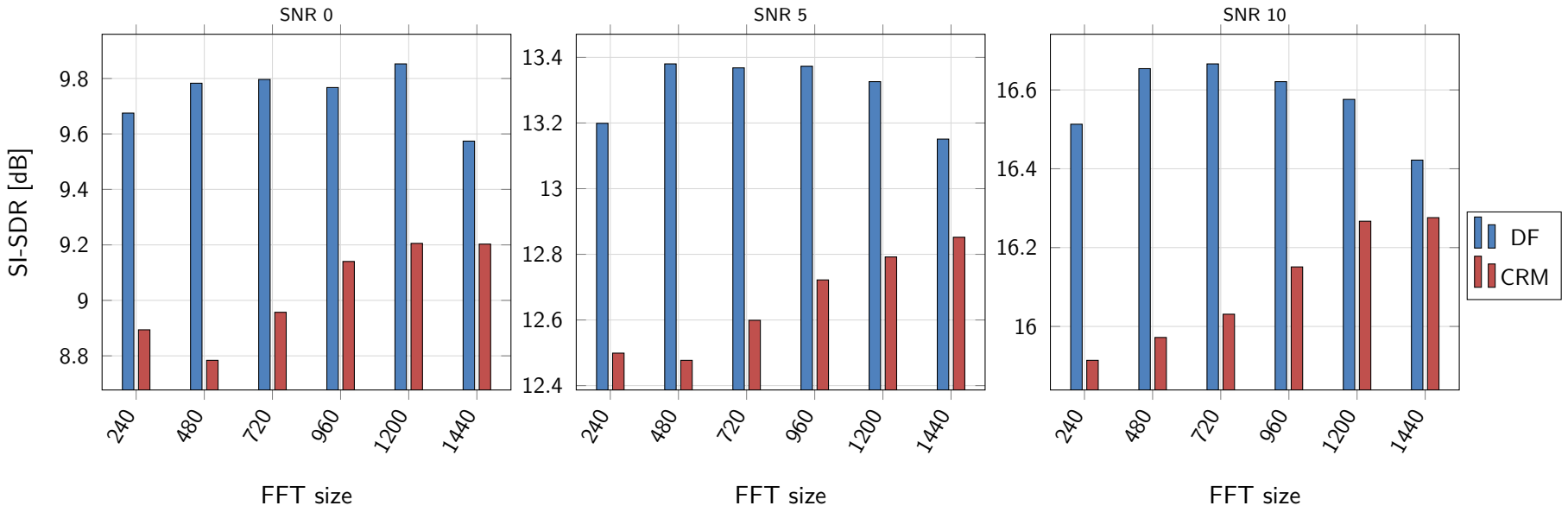


$$Y(k,f) = \sum_{i=0}^{N} C(k,i,f) \cdot X(k-i+l,f)$$

where C are the predicted FD deep filtering coefficient with filter order N.

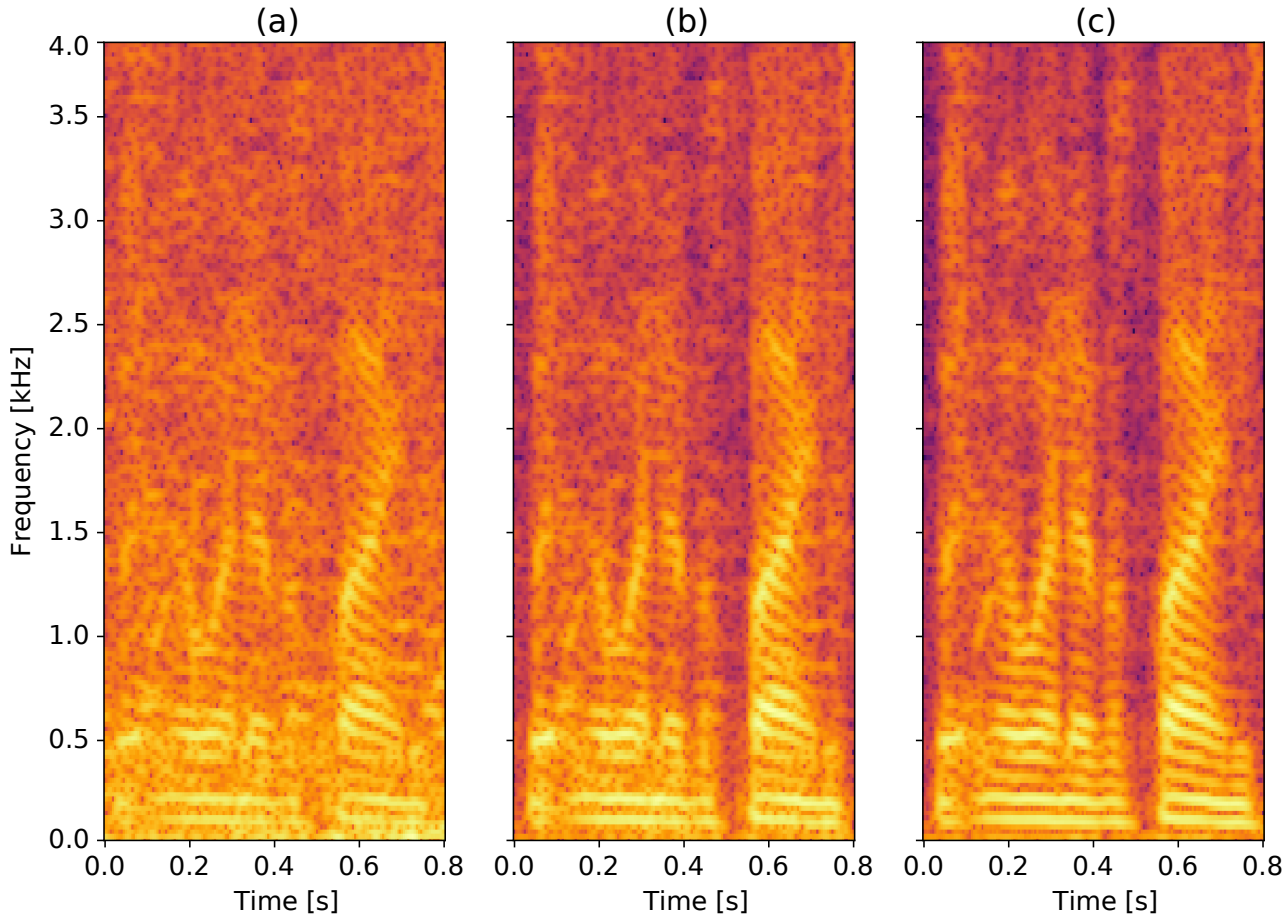
Results
Deep Filtering vs CRM

Comparison between deep filtering and a traditional complex ratio mask (CRM) based on the internal test set.



Results
Deep Filtering vs CRM

Qualitative sample from the Voicebank/DEMAND test set. Noisy (a), CRM (b), DF (c).



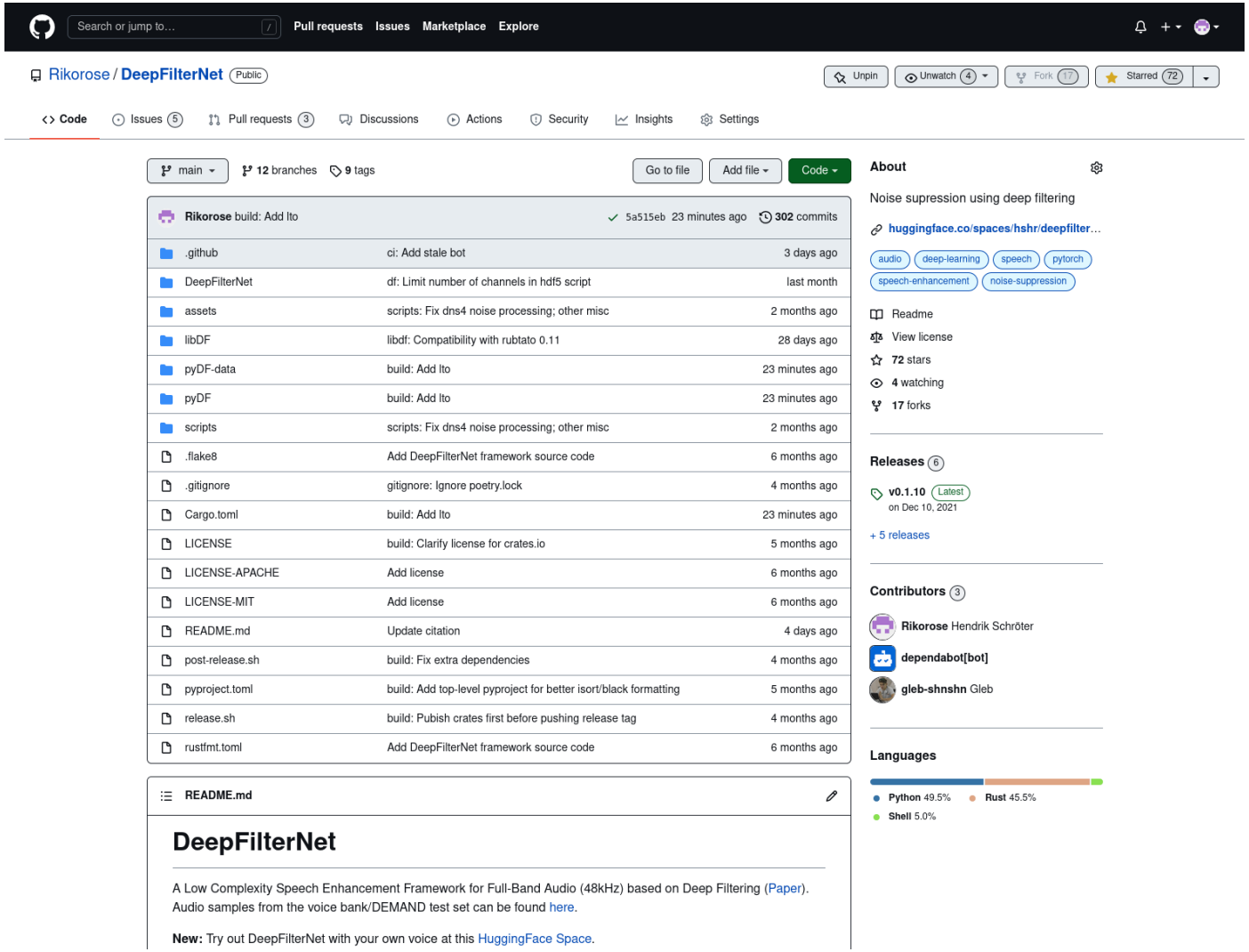
Results
Voicebank/DEMAND

Comparison with related work based on the Voicebank/DEMAND test set [5].

Model	Params [M]	MACS [G]	WB-PESQ [MOS]	SI-SDR [dB]
Noisy	-	-	1.97	8.41
PercepNet [6]	8.0	0.80	2.73	-
DCCRN [7]	3.7	14.36	2.68	-
DCCRN+ [8]	3.3	-	2.84	-
DeepFilterNet	1.8	0.35	2.81	16.63
w/o stage 2	0.9	0.25	2.57	13.81

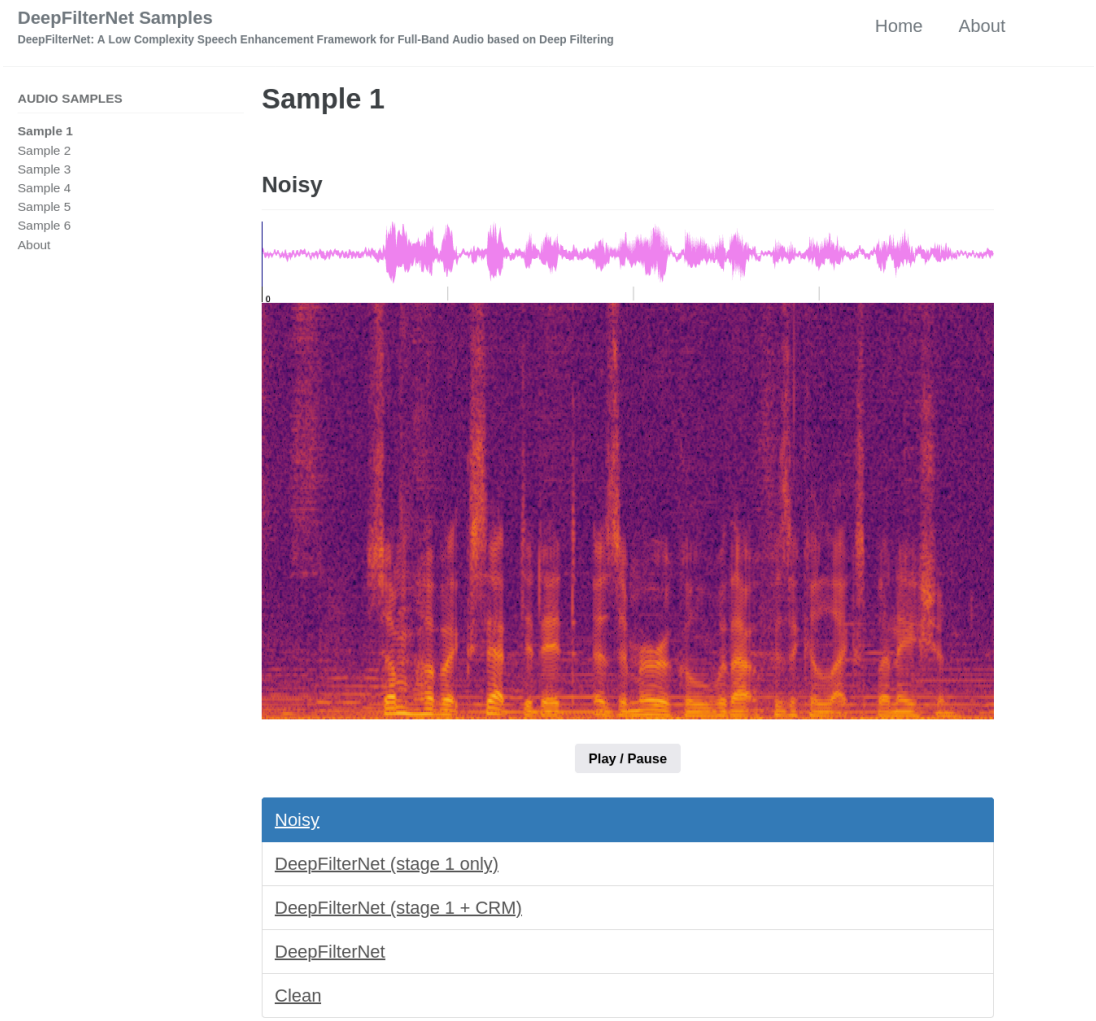
We have code!

https://github.com/Rikorose/DeepFilterNet



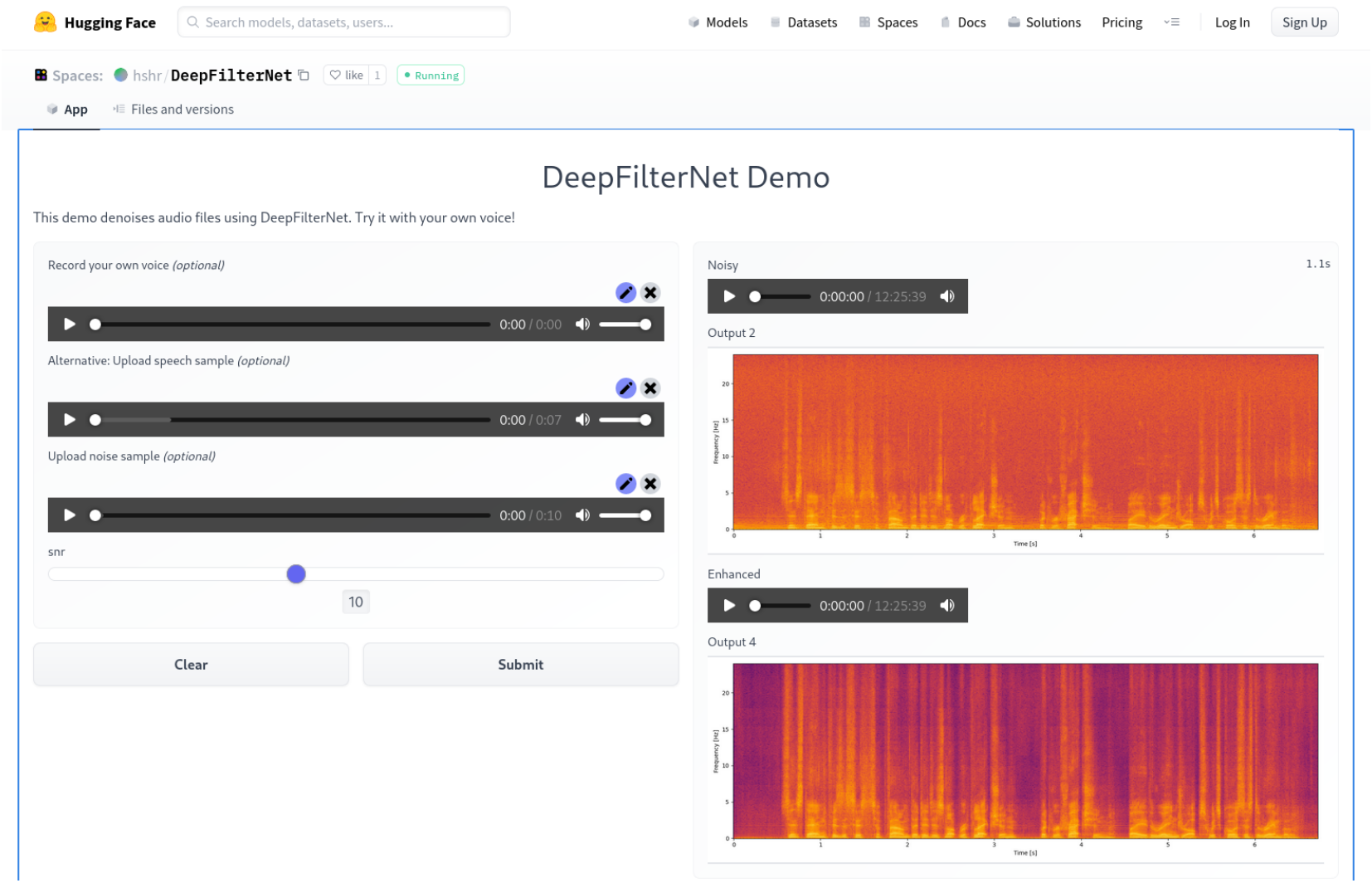
Audio Samples

https://rikorose.github.io/DeepFilterNet-Samples/



Demo

https://huggingface.co/spaces/hshr/DeepFilterNet



References I

[1] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[2] C. K. A. Reddy, E. Beyrami, H. Dubey, *et al.*, *The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework*, 2020. eprint: 2001.08662.

[3] W. Mack and E. A. Habets, "Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2020.

[4] H. Schröter, T. Rosenkranz, A. Escalante Banuelos, M. Aubreville, and A. Maier, "CLCNet: Deep learning-based noise reduction for hearing aids using complex linear coding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona), 2020. [Online]. Available: https://rikorose.github.io/CLCNet-audio-samples.github.io/.

[5] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *SSW*, 2016, pp. 146–152.

[6] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *INTERSPEECH 2020*, Oct. 25–29, 2020.

[7] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020.

[8] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *INTERSPEECH*, 2021.

[9] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6865–6869.

Thank You!

Contact

Code

Samples