A Two-stage Universal Speech Enhancement System for URGENT 2024 Challenge

Xiaobin Rong[1,2,*], Dahan Wang[1,2,*], Qinwen Hu[1,2], Jing Lu[1,2]

[1] Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China
[2] NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China

* These authors contributed equally to this work.

Contents

1. Introduction

➢ The URGENT 2024 Challenge[1]

- **Various sampling frequencies:**

  {8, 16, 22.05, 24, 32, 44.1, 48} kHz

- **Sub-tasks:**
  - Denoising
  - Dereverberation
  - Declipping
  - Bandwidth extension (BWE)

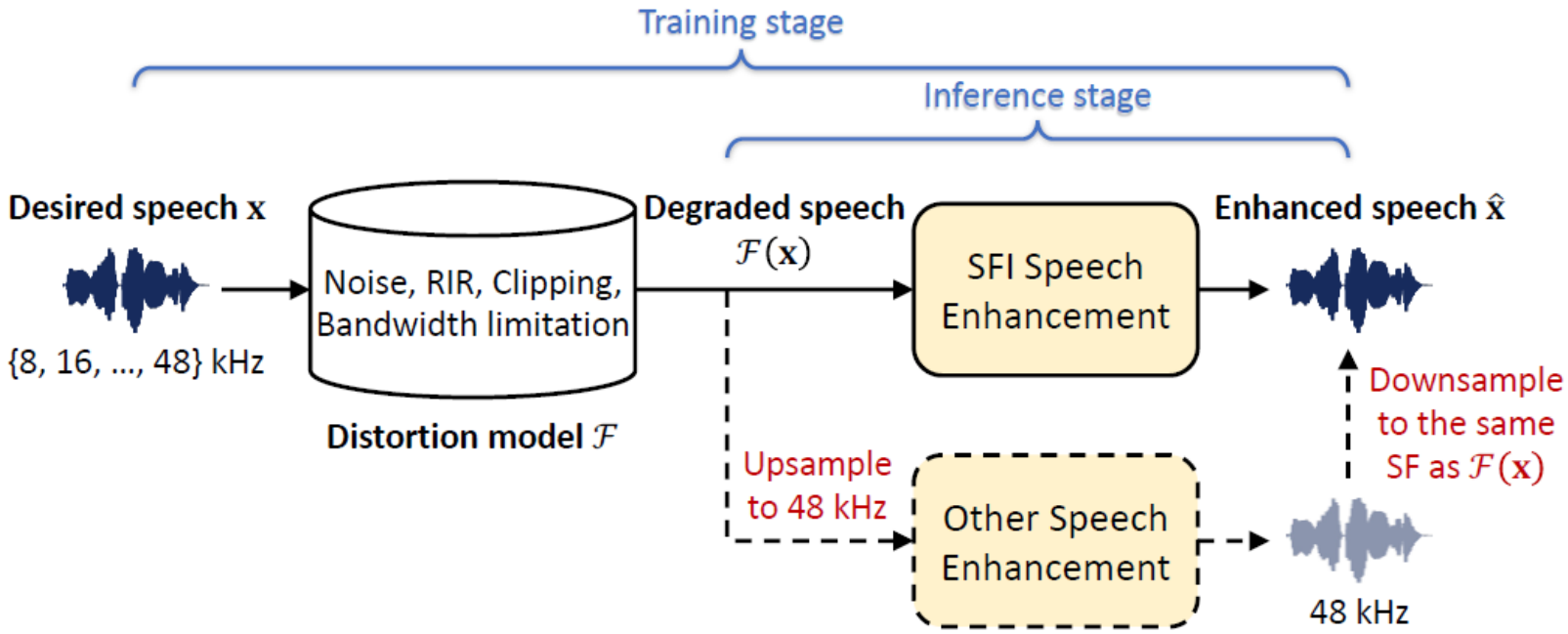


Fig. The diagram of the URGENT task.

[1] Zhang, W., Scheibler, R., Saijo, K., Cornell, S., Li, C., Ni, Z., ... & Qian, Y. (2024). URGENT Challenge: Universality, Robustness, and Generalizability For Speech Enhancement. *arXiv preprint arXiv:2406.04660*

2. Method description

➢ Overview

(1) **Denoising-Dereverberation-Declipping (D3) stage**

- Sampling-frequency-independent (SFI):
  - Using TF-GridNet[1]
  - Using fixed-duration STFT window/hop sizes
- Trained on 16 kHz data, and inferred on data with various sampling rates
- Fine-tuning the model with loss functions on log-spectral distances (LSD) and Mel-cepstral distances (MCD)

(2) **Bandwidth Extension (BWE) stage**

- Upsampling to 48 kHz → performing BWE → downsampling back to the original sampling rate
- Channel-wise subband (CWS)[2] processing: To divide the fullband complex spectrum into 3 subbands and concatenate them in the channel dimension
- Generative adversarial training:
  - Using TF-GridNet as the generator
  - Using multi-band and multi-resolution discriminators[3]

[1] Wang Z, et al. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023*(pp. 1-5). IEEE.
[2] Hao L, et al. Channel-wise subband input for better voice and accompaniment separation on high resolution music. *In Proc. Interspeech, 2020*, pp. 1241–1245.
[3] Liu W, Shi Y, Chen J, et al. Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction[J]. arXiv preprint arXiv:2306.08454, 2023.

2. Method description

2.1 The overall model architecture

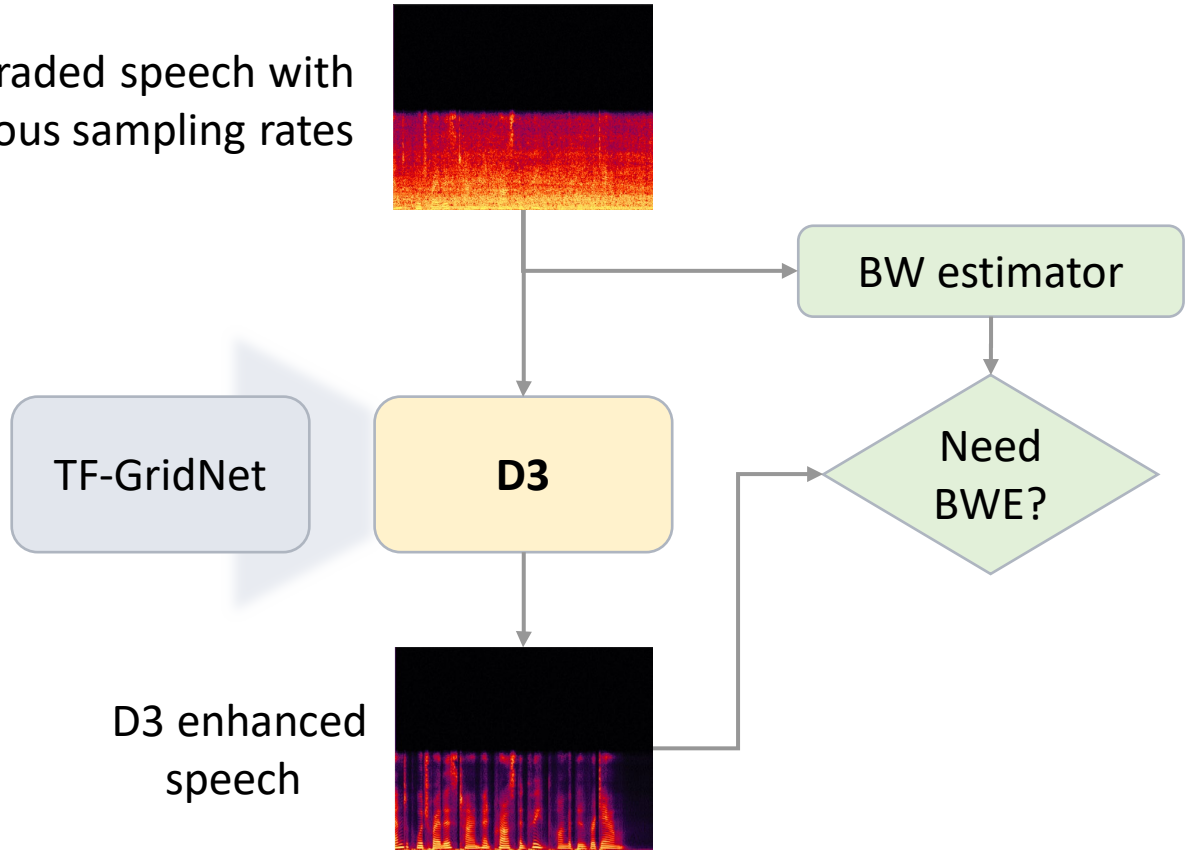


Fig. The diagram of our proposed two-stage universal speech enhancement system.

[1] Wang Z, et al. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023*(pp. 1-5). IEEE.
[2] Liu W, Shi Y, Chen J, et al. Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction[J]. arXiv preprint arXiv:2306.08454, 2023.

2. Method description

2.1 The overall model architecture

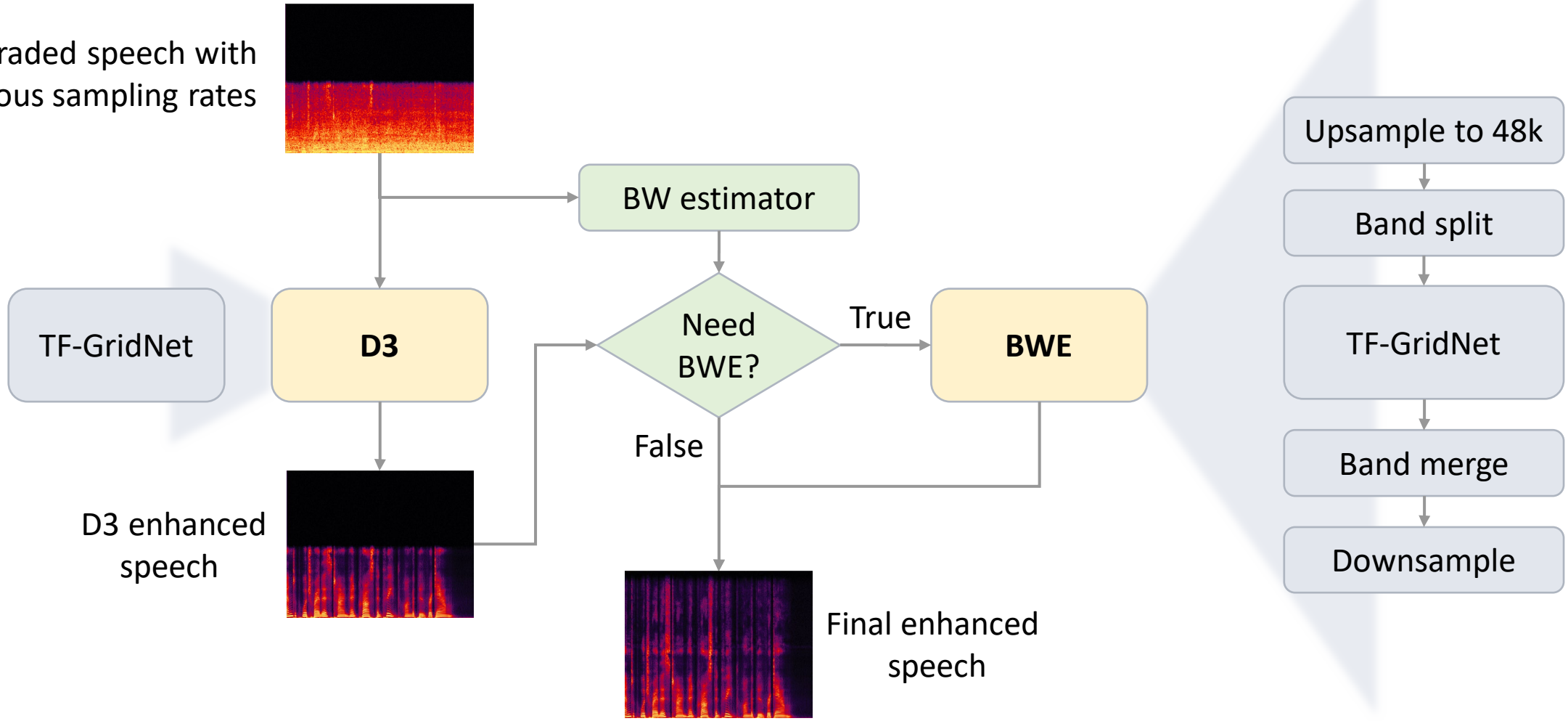


Fig. The diagram of our proposed two-stage universal speech enhancement system.

[1] Wang Z, et al. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023*(pp. 1-5). IEEE.
[2] Liu W, Shi Y, Chen J, et al. Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction[J]. arXiv preprint arXiv:2306.08454, 2023.

2. Method description

2.2 The D3 model

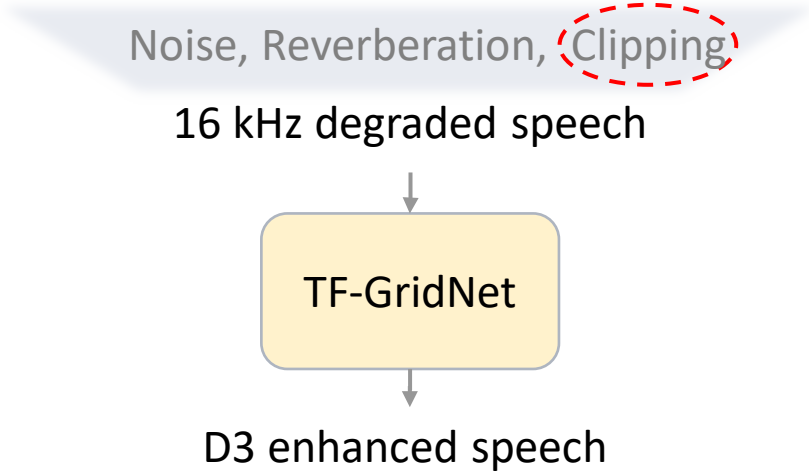


Fig. The diagram of the D3 stage.

2. Method description

2.2 The D3 model

Noise, Reverberation, Clipping

16 kHz degraded speech

TF-GridNet

D3 enhanced speech

Fig. The diagram of the D3 stage.

Declipping can be regarded as a denoising task in the STFT domain[1].

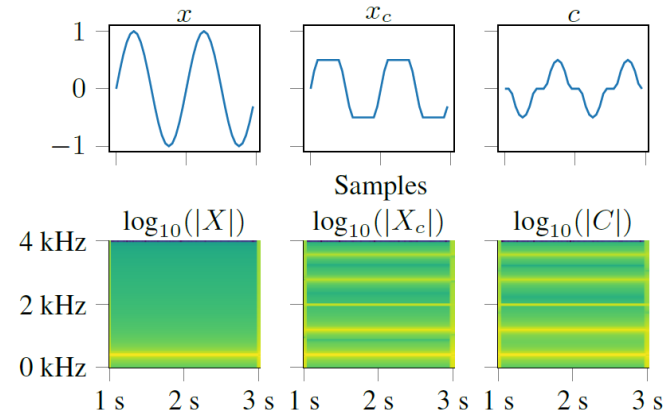


Fig. Clipping effect on a single sinusoidal signal
$c = x_c - x$, x: non-clipped signal, x_c: clipped signal.

[1] Mack, W., & Habets, E. A. Declipping speech using deep filtering. In *2019 IEEE WASPAA* (pp. 200-204). IEEE.

2. Method description

2.2 The D3 model

Table. Configuration of the TF-GridNet in D3 stage.

Configuration	value
n_layers	10
emb_dim	96
lstm_hidden_units	200
attn_n_head	8
attn_qk_output_channel	4

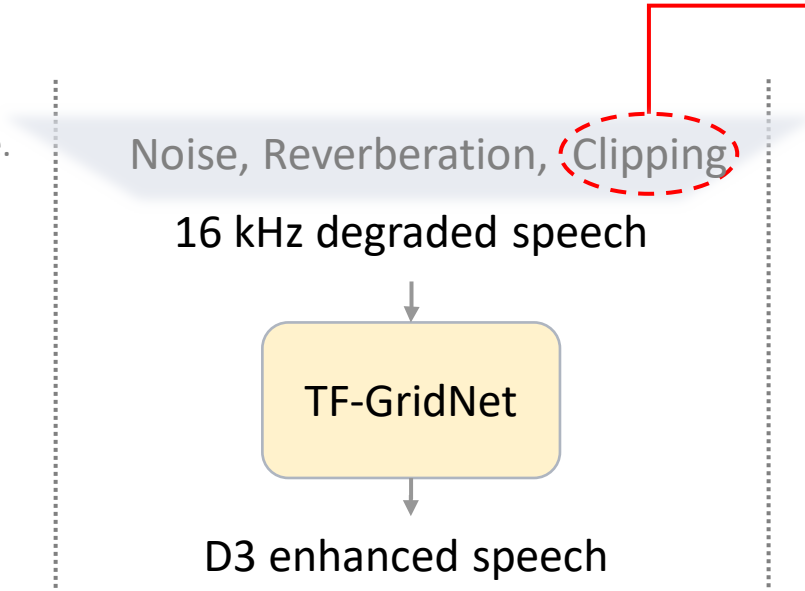


Fig. The diagram of the D3 stage.

Declipping can be regarded as a denoising task in the STFT domain[1].

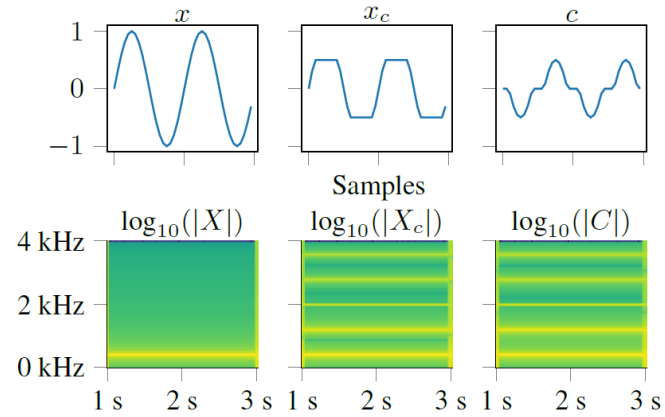


Fig. Clipping effect on a single sinusoidal signal
$c = x_c - x$, x: non-clipped signal, x_c: clipped signal.

[1] Mack, W., & Habets, E. A. Declipping speech using deep filtering. In *2019 IEEE WASPAA* (pp. 200-204). IEEE.

2. Method description

2.2 The D3 model

Table. Configuration of the TF-GridNet in D3 stage.

Configuration	value
n_layers	10
emb_dim	96
lstm_hidden_units	200
attn_n_head	8
attn_qk_output_channel	4

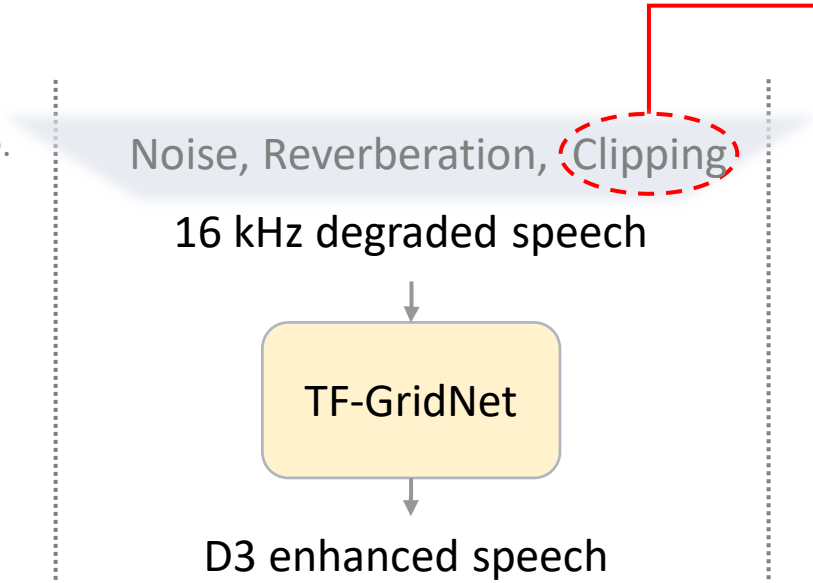


Fig. The diagram of the D3 stage.

Declipping can be regarded as a denoising task in the STFT domain[1].

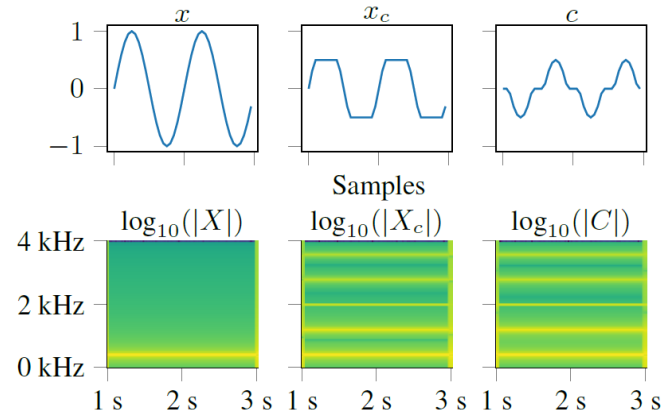


Fig. Clipping effect on a single sinusoidal signal
$c = x_c - x$, x: non-clipped signal, x_c: clipped signal.

Hybrid loss: $\mathcal{L}_{\mathrm{h}} = 0.001\mathcal{L}_{\mathrm{SISNR}}(\hat{s},s) + 0.7\mathcal{L}_{\mathrm{mag}}(\hat{S},S) + 0.3\left[\mathcal{L}_{\mathrm{r}}(\hat{S},S) + \mathcal{L}_{\mathrm{i}}(\hat{S},S)\right]$

$$\mathcal{L}_{\mathrm{SISNR}} = -\mathrm{SISNR}(\hat{s},s) \qquad \mathcal{L}_{\mathrm{mag}} = \mathbb{E}\left[\left(|\hat{S}|^{0.3}\right)^2 - \left(|S|^{0.3}\right)^2\right] \qquad \mathcal{L}_{\mathrm{r/i}} = \mathbb{E}\left[\left(\frac{\hat{S}_{\mathrm{r/i}}}{|\hat{S}|^{0.7}}\right)^2, \left(\frac{S_{r/i}}{|S|^{0.7}}\right)^2\right]$$

Superscripts, subscripts and modifiers:
r: the real part of the spectrogram.
i: the imaginary part of the spectrogram.
$\hat{\cdot}$: the enhanced speech.

Fine-tuning loss: $\mathcal{L}_{\mathrm{f}} = \mathcal{L}_{\mathrm{h}} + 0.02\mathcal{L}_{\mathrm{LSD}} + 0.0005\mathcal{L}_{\mathrm{MCD}}$

[1] Mack, W., & Habets, E. A. Declipping speech using deep filtering. In *2019 IEEE WASPAA* (pp. 200-204). IEEE.

2. Method description

2.3 The BWE model

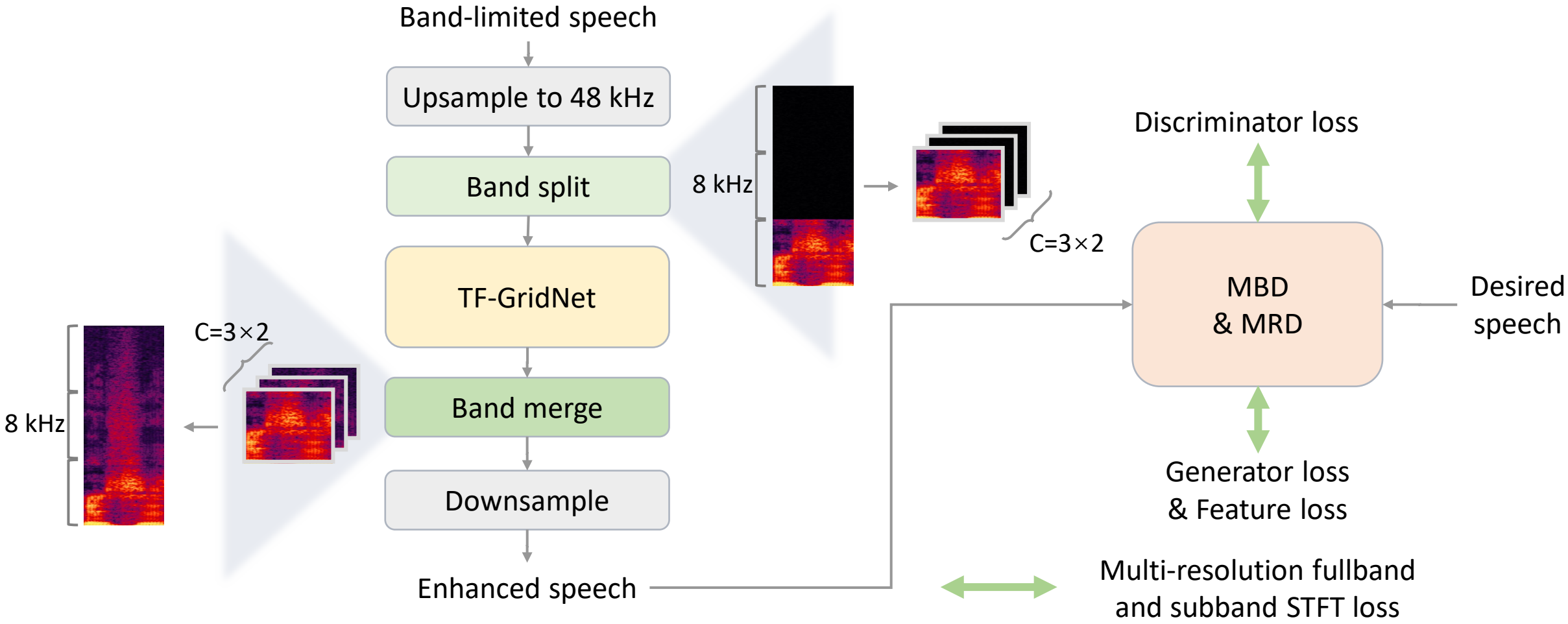


Fig. The diagram of the BWE stage.

[1] Liu W, Shi Y, Chen J, et al. Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction[J]. arXiv preprint arXiv:2306.08454, 2023.
[2] Wang Z, et al. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023*(pp. 1-5). IEEE.

2. Method description

2.3 The BWE model

Generator loss[1]:

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{S}}\left(\hat{S}, S\right) + \mathcal{L}_{\mathrm{S}}\left(\hat{S}^{sub}, S^{sub}\right) + \mathcal{L}_{\mathrm{adv}}\left(\hat{s}\right) + 0.1\mathcal{L}_{\mathrm{feat}}\left(\hat{s}, s\right)$$

$$\mathcal{L}_{\mathrm{S}}\left(\hat{S}, S\right) = \sum_{R}\left[\left\|\log \hat{S}_R - \log S_R\right\|_1 + \frac{\left\|S_R - \hat{S}_R\right\|_2}{\left\|\hat{S}_R\right\|_2}\right]$$

$$\mathcal{L}_{\mathrm{adv}}\left(\hat{s}\right) = \mathbb{E}\left[\left(1 - D\left(\hat{s}\right)\right)^2\right]$$

$$\mathcal{L}_{\mathrm{feat}}\left(\hat{s}, s\right) = \mathbb{E}\left[\frac{1}{L}\sum_{l=0}^{L-1}\left|D^l\left(s\right) - D^l\left(\hat{s}\right)\right|\right]$$

Discriminator loss:

$$\mathcal{L}_{\mathrm{D}}\left(\hat{s}, s\right) = \mathbb{E}\left[\left(D\left(s\right) - 1\right)^2 + \left(D\left(\hat{s}\right)\right)^2\right]$$

Table. Configuration of the TF-GridNet in BWE stage.

Configuration	value
n_layers	5
emb_dim	48
lstm_hidden_units	100
attn_n_head	4
attn_qk_output_channel	2

Superscripts, subscripts and modifiers:
L: the number of the discriminator's layer.
R: the resolution of STFT.
D: the discriminator.

[1] Liu W, Shi Y, Chen J, et al. Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction[J]. arXiv preprint arXiv:2306.08454, 2023.

3. Experiments and Results

➢ Experiments configurations

Configuration		value
Frame length		32 ms
Hop length		16 ms
Learning rate	D3/BWE	Max: 5e-4; Min: 5e-6
	Fine-tuning of D3	Max: 1e-4; Min: 1e-6
Warmup steps	D3/BWE	Warmup (warmup steps: 20 K)
	Fine-tuning of D3	Warmup (warmup steps: 20 K)
Training steps	D3/BWE	200 K
	Fine-tuning of D3	3 K
Training segment length	D3	4 s
	BWE	2 s
Batch size	D3	8
	BWE	24
Model size	D3	22.09 M
	BWE	2.77 M

3. Experiments and Results

➢ Experimental results

Table. Experimental results on the validation set.

Method	Non-intrusive		Intrusive					Downstream-task-independent		Downstream-task-dependent	
	DNSMOS	NISQA	PESQ	ESTOI	SDR	MCD	LSD	BERT	LPS	Spksim	WAcc
D3	3.05	3.68	2.92	0.86	16.58	3.67	3.51	0.863	0.888	0.750	88.39
+ FT with LSD & MCD	3.07	3.57	2.89	0.86	16.44	3.32	3.12	0.868	0.885	0.760	88.17
+ BWE	3.07	3.65	2.86	0.86	16.26	3.07	2.36	0.868	0.881	0.784	88.07

3. Experiments and Results

➢ **Experimental results**

Table. Experimental results on the blind test set.

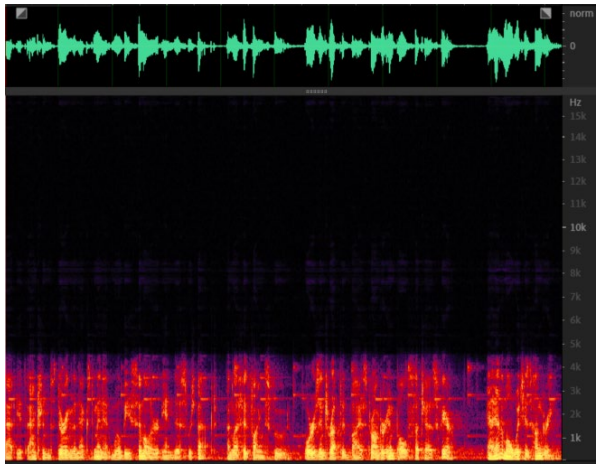
Non-intrusive		Intrusive						Downstream-task-independent		Downstream-task-dependent		Subjective	Overall ranking score
DNSMOS	NISQA	PESQ	ESTOI	SDR	MCD	LSD	POLQA	BERT	LPS	Spksim	WAcc	MOS	
2.95 (11)	3.35 (11)	2.66 (2)	0.86 (3)	13.54 (3)	3.14 (2)	2.70 (1)	3.45 (3)	0.85 (2)	0.83 (3)	0.81 (2)	73.10 (3)	3.40 (7)	5.07

3. Experiments and Results

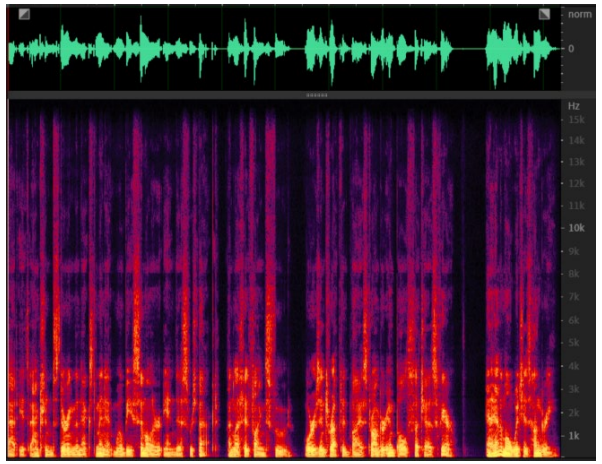
➢ Audio samples

Samples from the validation set:

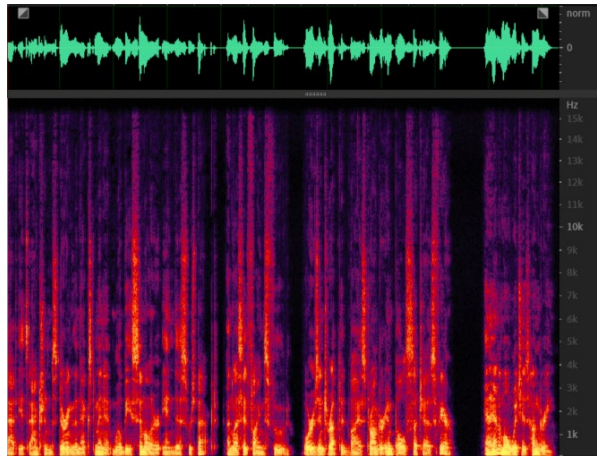
File id 7037: a sample with noise and band-limitation distortion



(a) noisy



(b) enhanced



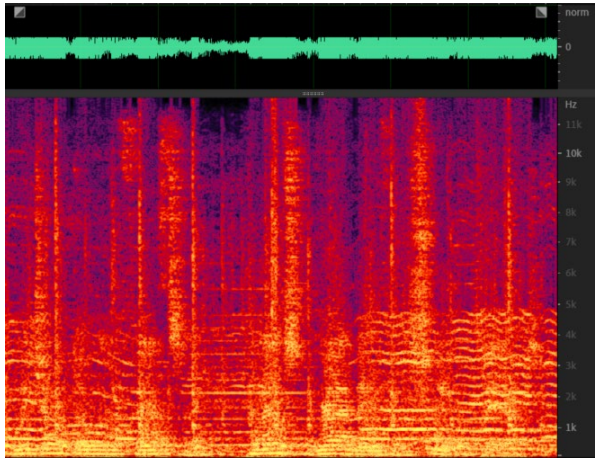
(c) clean

3. Experiments and Results

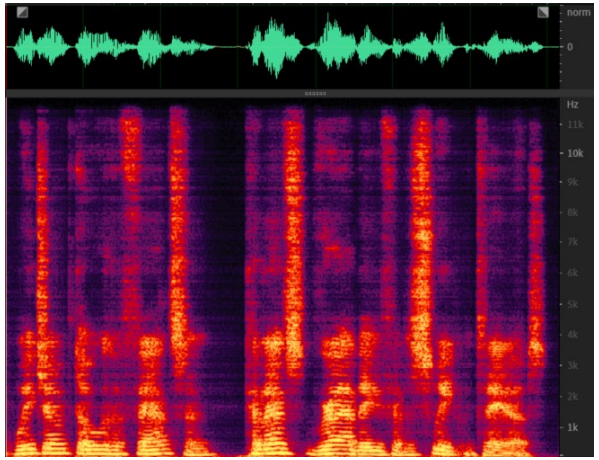
➢ Audio samples

Samples from the validation set:

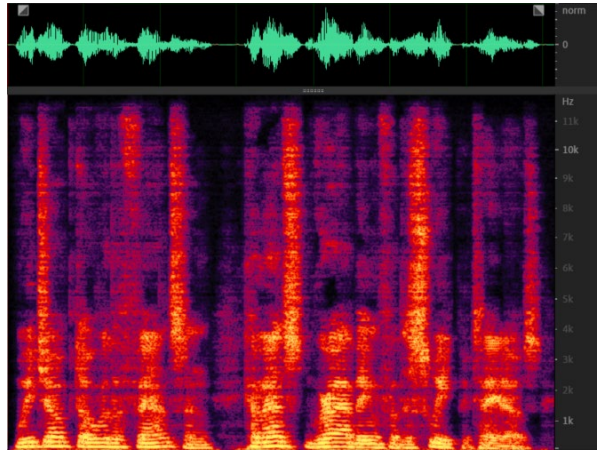
File id 20322: a sample with noise and clipping distortion



(a) noisy



(b) enhanced



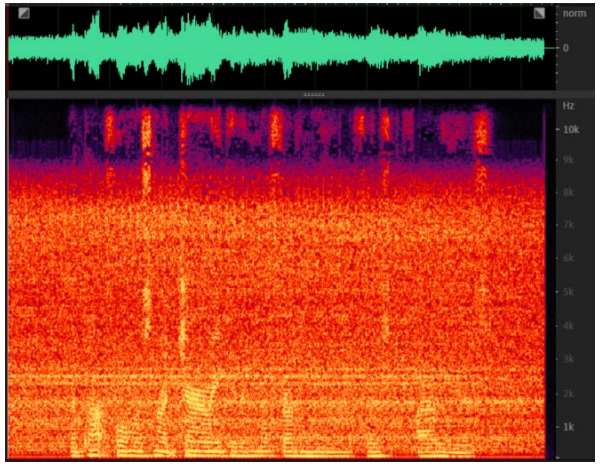
(c) clean

3. Experiments and Results

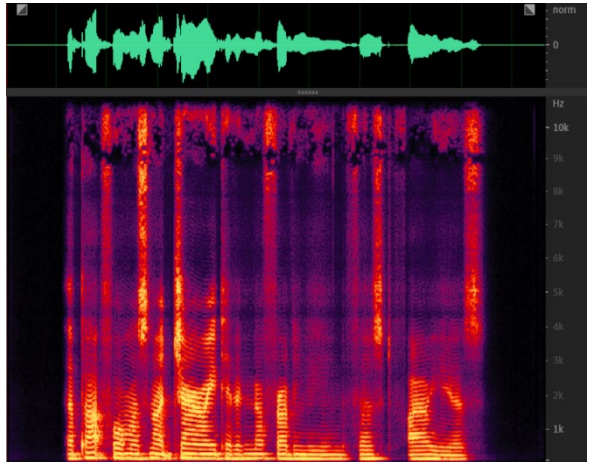
➢ Audio samples

Samples from the validation set:

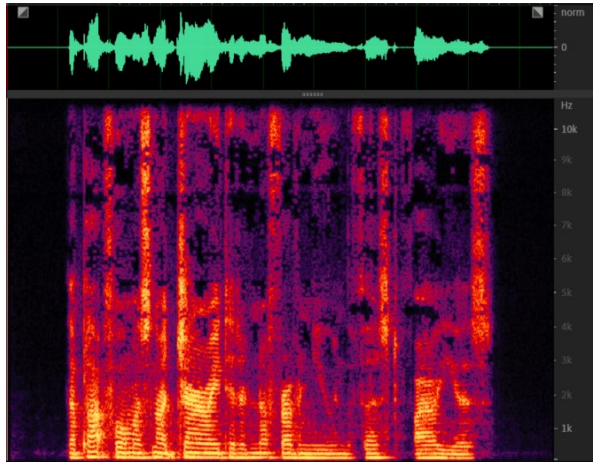
File id 22999: a sample with severe noise



(a) noisy



(b) enhanced

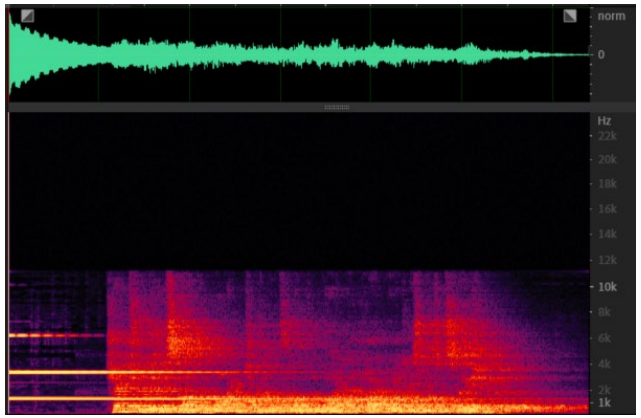


(c) clean

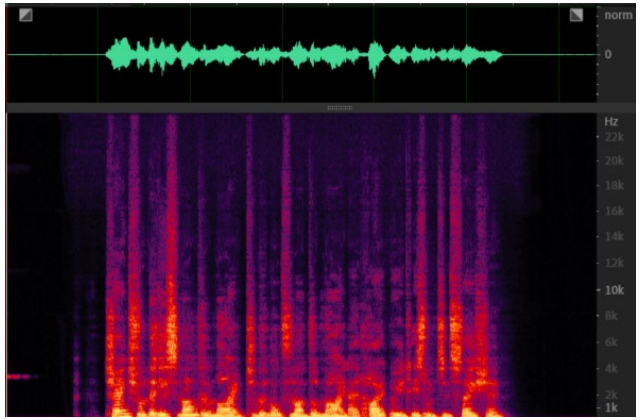
3. Experiments and Results

➢ Audio samples

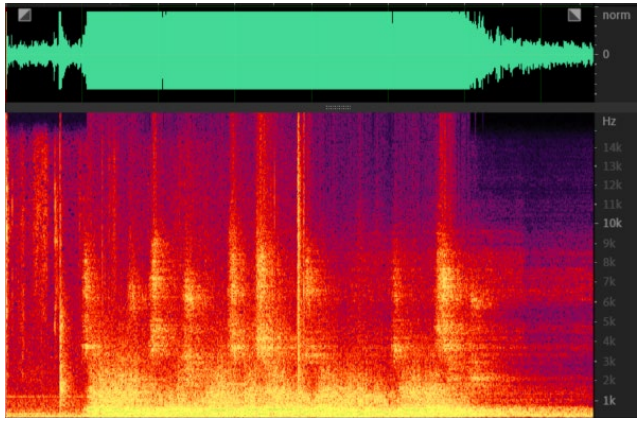
Samples from the blind test set:



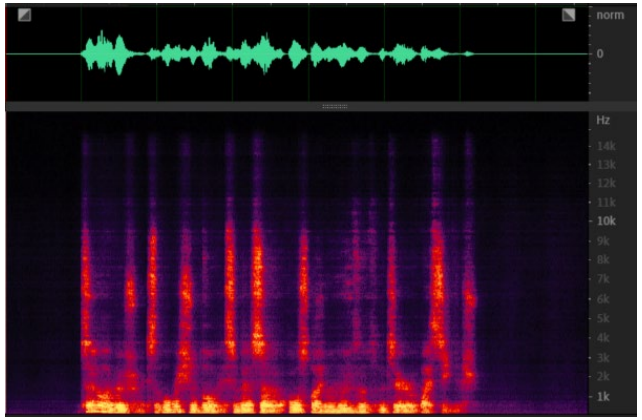
(a) File id 229: noisy



(b) File id 229: enhanced



(c) File id 63: noisy



(d) File id 63: enhanced

Thanks for your attention!