APSIPA ASC 2024

SDNet: Noise-Robust Bandwidth Extension under Flexible Sampling Rates

Junkang Yang, Hongqing Liu, Lu Gan, Yi Zhou, Xing Li, Jie Jia and Jinzhuo Yao

Intelligent Speech and Audio Research Lab,
Chongqing University of Posts and Telecommunications

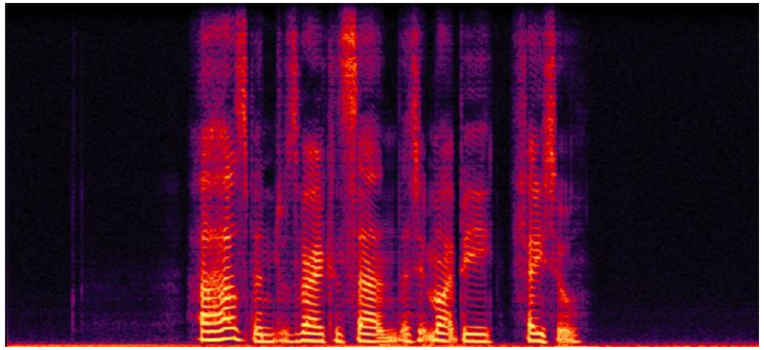
vivo AI Lab

Introduction

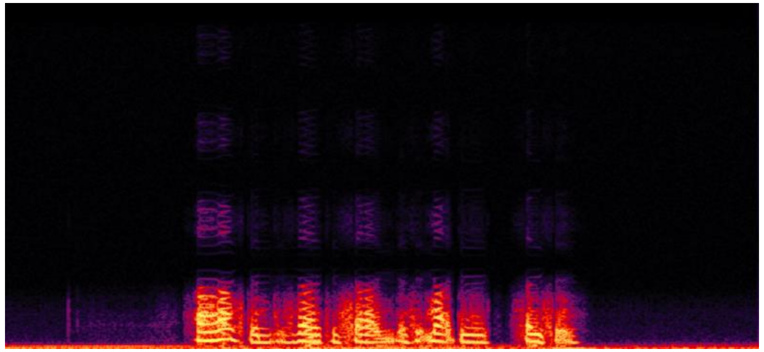
- What is Bandwidth Extension (also named as Audio Super-Resolution)?

Recovering high-resolution (HR) signals from low-resolution (LR) counterparts.

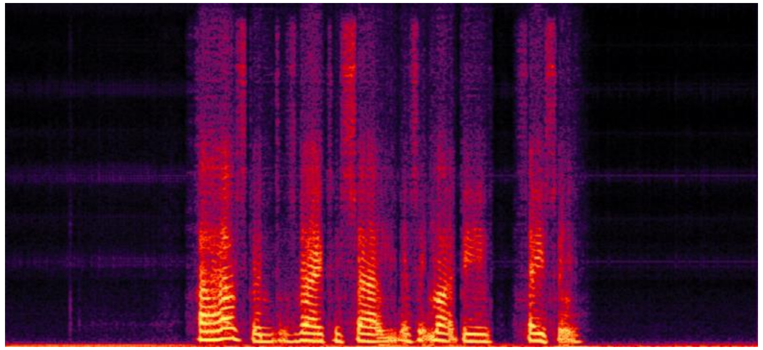
- Applications: wireless communication, speech recognition, text-to-speech.



Original



Low-Resolution



Super-Resolution

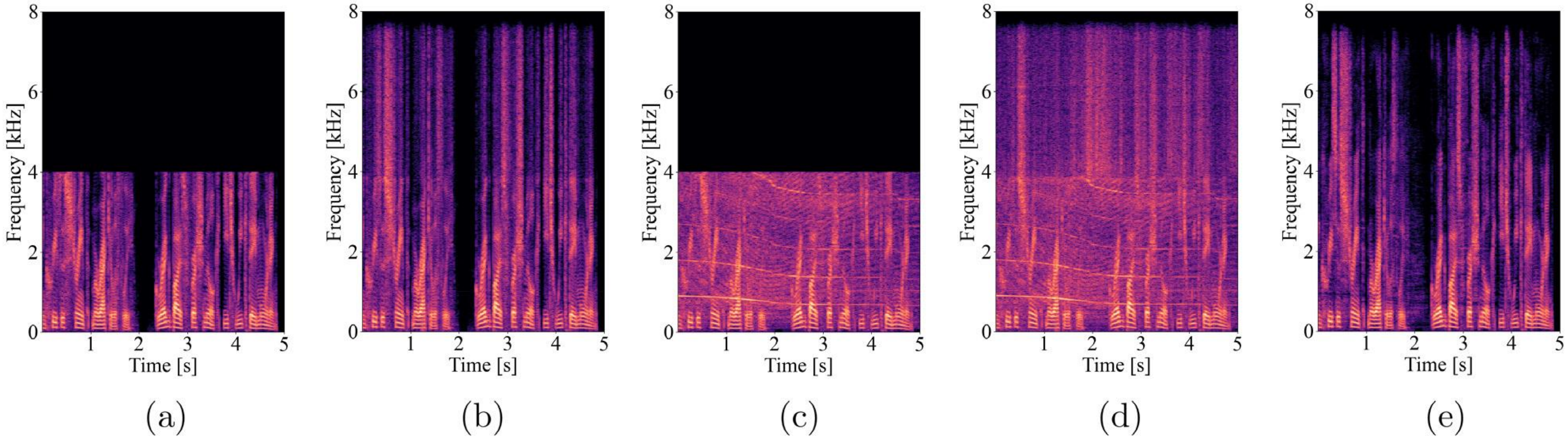
- Problem: Current models struggle with noisy environments and flexible sampling rates.

- Goal: Jointly handle noise reduction and bandwidth extension and support multiple sampling rates with a single model.

Challenges in BWE

- Ineffectiveness in noisy environments

Noise interference biases high-frequency predictions.



Challenges in BWE

- Limited flexibility: Fixed sampling rates in most models.

		SingleSpeaker			MultiSpeaker			Piano		
Ratio	Obj.	Spline	DNN	Ours	Spline	DNN	Ours	Spline	DNN	Ours
$r = 2$	SNR	20.3	20.1	21.1	19.7	19.9	20.7	29.4	29.3	30.1
	LSD	4.5	3,7	3.2	4.4	3.6	3.1	3.5	3.4	3.4
$r = 4$	SNR	14.8	15.9	17.1	13.0	14.9	16.1	22.2	23.0	23.5
	LSD	8.2	4.9	3.6	8.0	5.8	3.5	5.8	5.2	3.6
$r = 6$	SNR	10.4	n/a	14.4	9.1	n/a	10.0	15.4	n/a	16.1
	LSD	10.3	n/a	3.4	10.1	n/a	3.7	7.3	n/a	4.4

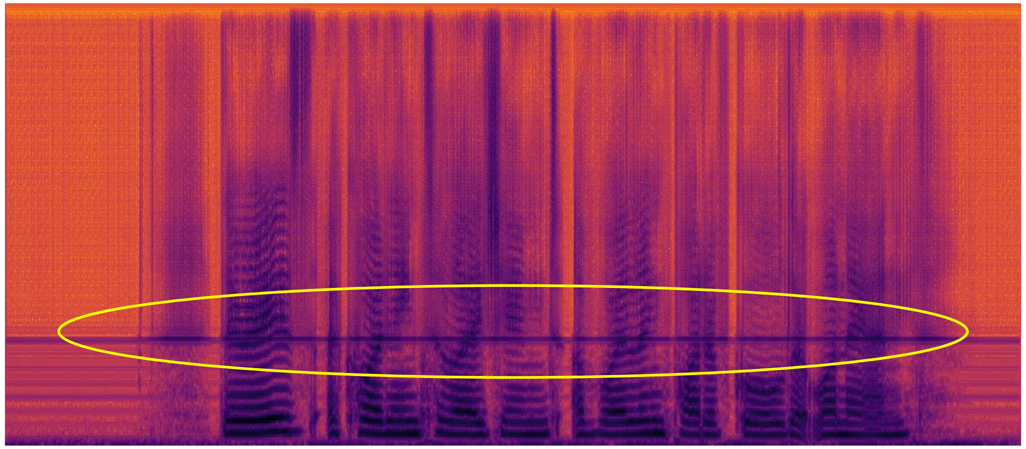
Table 2: Accuracy evaluation of audio-super resolution methods (in dB) on each of the three super-resolution tasks at upscaling ratios $r = 2, 4, 6$.

Table 1: L, V and M denote LSD, ViSQOL and MUSHRA respectively. MUSHRA score is specified with a ± Confidence Interval of 0.95.

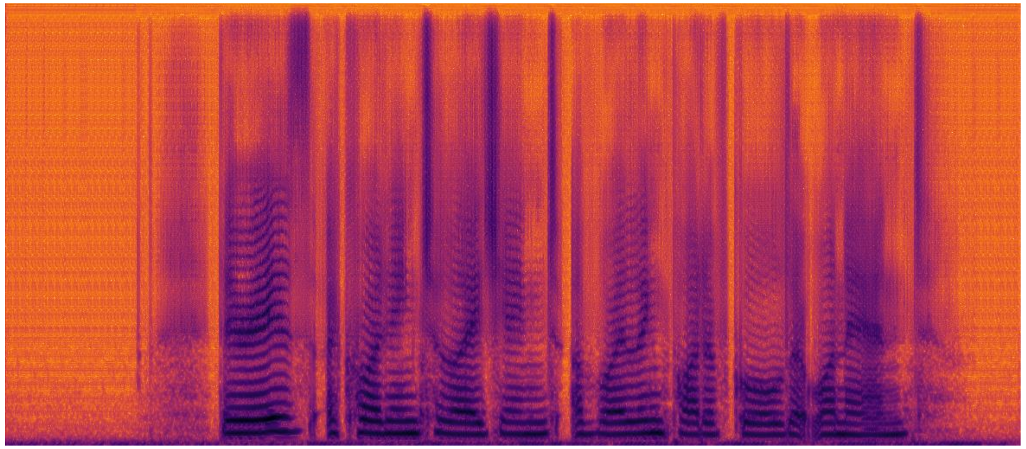
	8-16			8-24			4-16			11-44		
	L↓	V↑	M↑	L↓	V↑	M↑	L↓	V↑	M↑	L↓	V↑	M↑
Reference	-	-	96.25±1.5	-	-	97.16±1.4	-	-	96.18±1.5	-	-	95.30±2.5
Anchor	-	-	54.65±4.3	-	-	56.21±4.4	-	-	41.14±3.8	-	-	46.55±7.4
Sinc	2.32	3.41	60.13±4.7	2.96	3.41	59.49±4.8	3.59	2.27	43.03±3.9	3.91	1.97	47.61±8.0
TFiLM [4]	1.27	3.18	58.53±4.0	-	-	-	1.77	2.25	41.91±4.0	-	-	-
SEANet [5]	0.79	4.08	91.23±2.9	0.91	4.06	94.16±2.2	0.99	3.16	89.40±3.2	1.13	2.88	80.52±7.0
BEHMGAN [17]	-	-	-	-	-	-	-	-	-	1.80	2.01	46.27±8.3
Ours ($^{256}/_{512}$)	0.84	4.02	90.58±2.3	0.99	4.03	**96.40**±1.9	1.04	3.04	86.14±3.4	1.16	2.88	81.21±6.4
Ours ($^{128}/_{512}$)	0.80	4.11	92.63±2.4	0.91	4.12	95.41±2.0	0.99	3.15	**92.05**±2.7	1.16	**2.89**	81.67±6.8
Ours ($^{64}/_{512}$)	**0.77**	**4.16**	**94.64**±1.6	**0.90**	**4.17**	94.45±2.1	**0.94**	**3.28**	90.61±3.1	**1.12**	2.88	**84.18**±5.6

Challenges in BWE

- Significant artifacts



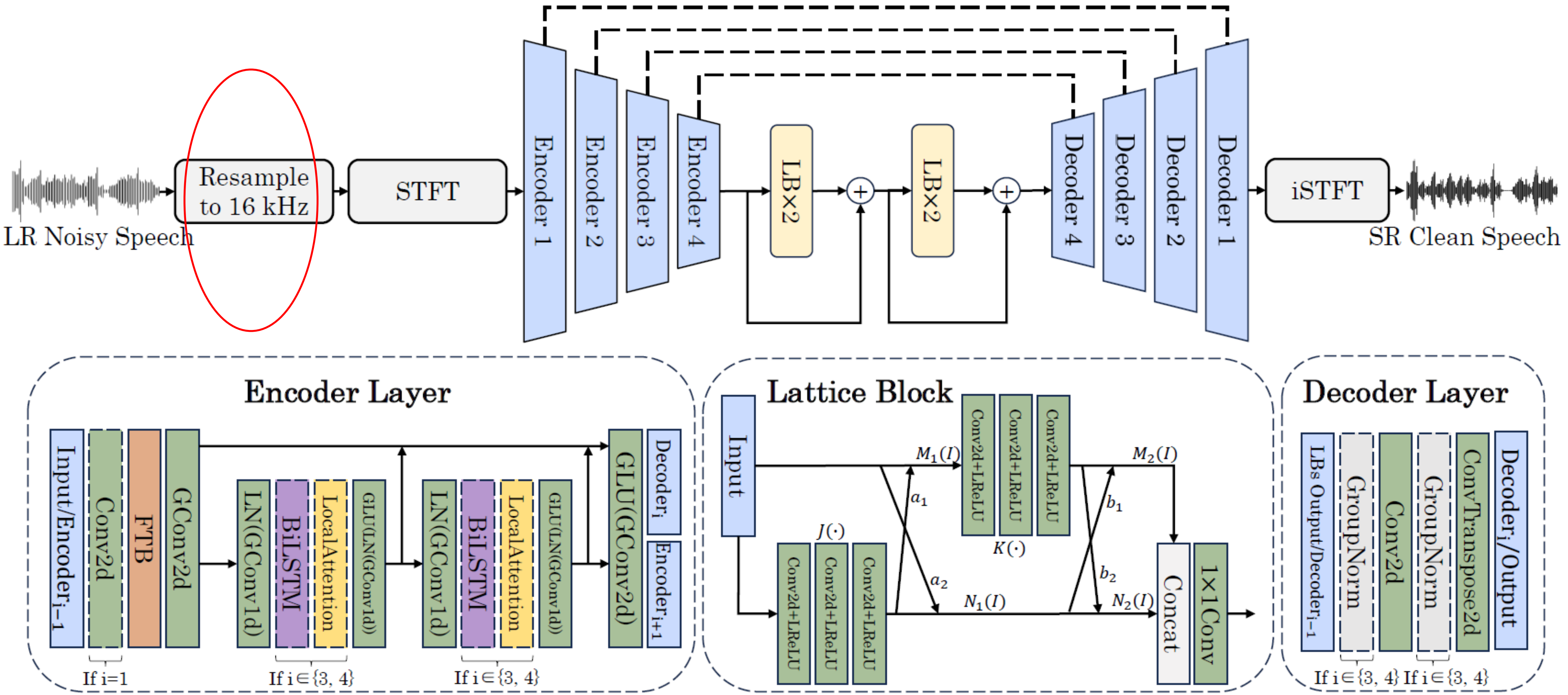
Super-Resolution Speech



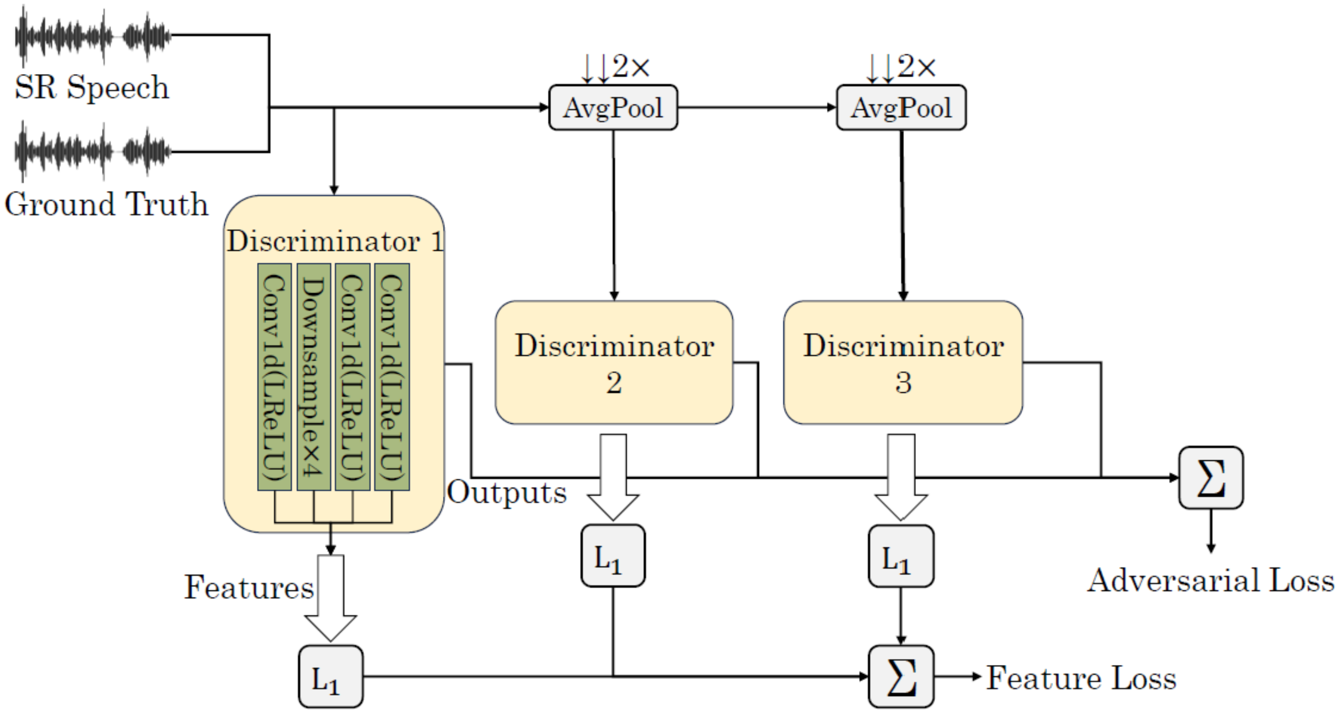
Ground Truth

- Lack of joint optimization for noise suppression and super-resolution. There are relatively few studies on noise-robust BWE compared to noise-free BWE.

Proposed Solution: SDNet



Proposed Solution: SDNet



Total Loss:

$$\mathcal{L} = \mathcal{L}_{MSTFT} + \mathcal{L}_G^{adv} + \lambda_f \mathcal{L}_f \qquad \lambda_f = 100$$

$$\mathcal{L}_{MSTFT} = E_{(x,y)\sim p_{data}} \left[\sum_{m=1}^{3} \left(\frac{||s(y,\theta_m) - s(x,\theta_m)||_F}{||s(y,\theta_m)||_F} + \frac{1}{N} ||log \frac{s(y,\theta_m)}{s(x,\theta_m)}|| \right) \right]$$

$||\cdot||_F$ and $||\cdot||_1$ are Frobenius and ℓ1-norms, N is the number of elements in the magnitude.

FFT bins ∈ {512, 1024, 2048} and hop length ∈ {50, 120, 240}. The window lengths are {240, 600, 1200}.

$$\mathcal{L}_G^{adv} = E_{x\sim p_{data}} \left[\frac{1}{K} \sum_k max(0, 1 - D_k(G(x))) \right],$$

$$\mathcal{L}_f = E_{(x,y)\sim p_{data}} \left[\frac{1}{KL} \sum_{k,l} ||D_k^l(y) - D_k^l(G(x))||_1 \right],$$

Where *k* is the number of discriminators, *l* is the number of layer in one discriminator.

Proposed Solution: SDNet

Training Data Augmentation:

Data: $\boldsymbol{y} \in \mathbb{Y}$
Result: The high-quality speech $\boldsymbol{y}$ and its downsampled version $\boldsymbol{x}$
$\boldsymbol{x} = \boldsymbol{s}$;
type $=$ random type (Chebyshev, Elliptic, Butterworth, Boxcar);
$f_{cut} \sim U(C_{low}, C_{high})$;
$order \sim U(O_{low}, O_{high})$;
$\boldsymbol{x} = \boldsymbol{x} * Filter(type, f_{cut}, order)$;
if resample, **then**
$\boldsymbol{x} = Resample(Resample(\boldsymbol{x}, 16000, f_{cut} \times 2), f_{cut} \times 2, 16000)$;
end if

We use a filter with random parameters when doing downsampling, the types include *Chebyshev, Elliptic, Butterworth* and *Boxcar*, the order is a random integer from 2 to 10, the cutoff frequency is an integer from 2000 to 8000 Hz. SNR: [-5, 20] dB.

Datasets(noise & speech):
DNS Challenge dataset, Valentini-Botinhao dataset.

Handle uncertain low sampling rate inputs and artifacts.

Experiment Results

TABLE I

TEST RESULTS OF NOISE-ROBUST BWE MODELS ON VALENTINI-BOTINHAO NOISY TEST SET DOWNSAMPLED TO 8 KHZ.

Method	**PESQ-WB↑**	**STOI (%)↑**	**CSIG↑**	**CBAK↑**	**COVL↑**	**LSD↓**
UEE [15]	2.23	93	2.27	2.39	2.17	2.72
MTL-MBE [8]	2.55	94	2.64	3.21	2.46	2.29
EP-WUN [9]	2.25	92	**3.50**	2.94	2.86	1.23
AFiLM + I-DTLN [4]	2.54	90	2.63	2.87	2.18	1.54
Ours	**2.67**	**95**	3.29	**3.32**	**2.92**	**1.16**

Experiment Results

TABLE II

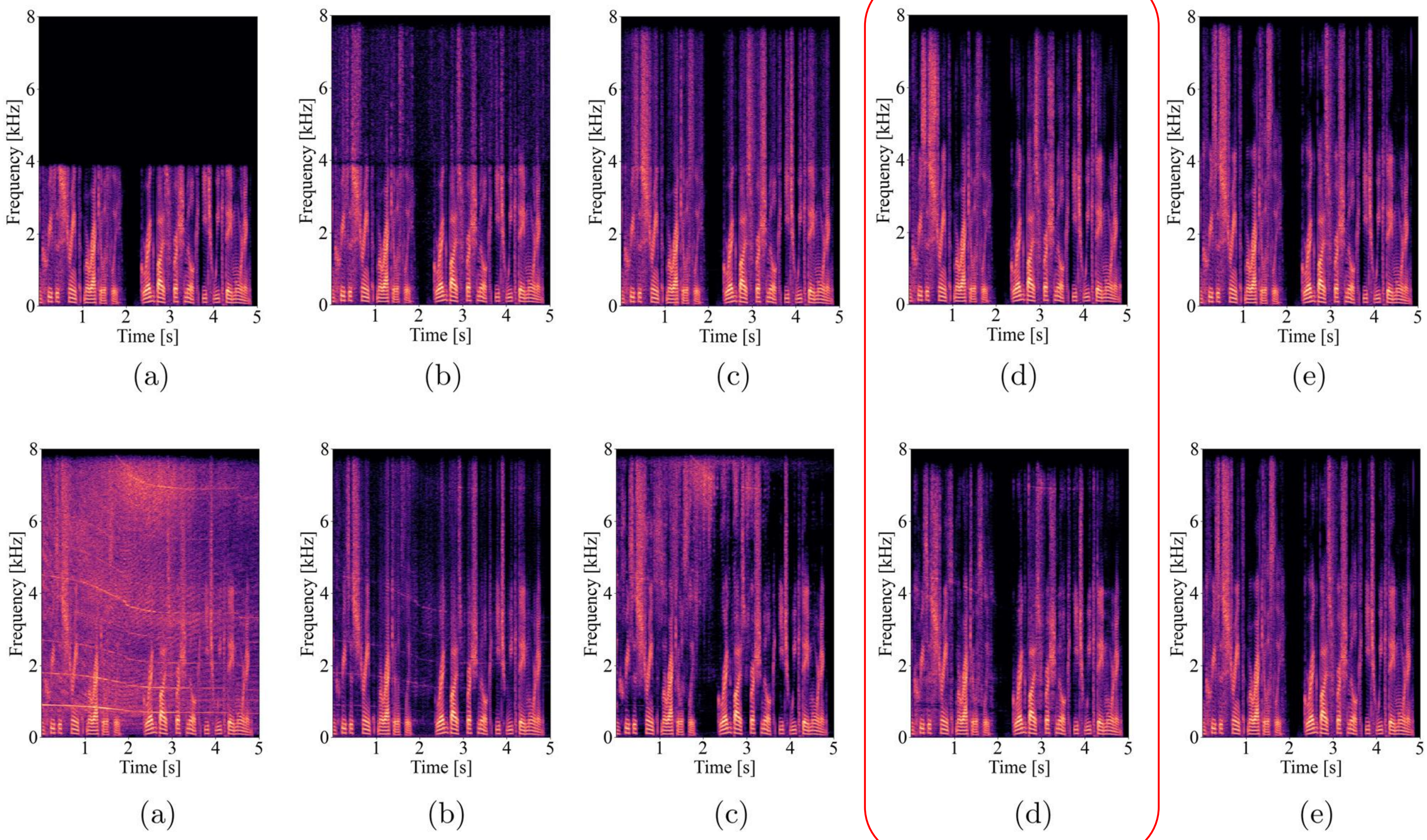
TEST RESULTS FOR DIFFERENT TASK ON DNS-CHALLENGE NO-REVERB TEST. "B" IS NOISE-FREE BWE, "D" IS DENOISE, AND "RB" IS NOISE-ROBUST BWE. "SOURCE" AND "NOISE" REPRESENT THE SAMPLING RATE OF INPUTS AND THE CASE WHETHER THE INPUTS CONTAIN NOISES.

Method	Task	Source	Noise	PESQ-NB↑	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD	MOS↑
WSRGlow	B	8 kHz	×	4.365	2.811	**99.4**	3.946	4.068	3.433	0.929	4.21
NU-Wave 2	B	8 kHz	×	4.353	2.646	**99.4**	3.663	2.869	3.209	1.328	4.08
AERO	B	8 kHz	×	4.369	3.295	98.5	**4.287**	4.273	3.844	0.802	4.27
Ours	B	8 kHz	×	**4.377**	**3.661**	98.6	4.103	**4.553**	**3.935**	**0.783**	**4.55**
DCCRN	D	16 kHz	✓	3.17	2.64	92.9	—	—	—	—	—
FullSubNet	D	16 kHz	✓	3.28	2.72	95.3	—	—	—	—	—
DPT-FSNet	D	16 kHz	✓	3.28	2.72	95.3	—	—	—	—	—
Ours	D	16 kHz	✓	**3.29**	**2.80**	**96.0**	—	—	—	—	—
VoiceFixer	RB	8 kHz	✓	2.535	1.679	84.0	2.532	1.914	2.043	1.323	3.83
Ours	RB	8 kHz	✓	**3.554**	**2.777**	**97.1**	**3.313**	**3.532**	**3.063**	**1.218**	**4.38**
VoiceFixer	RB	4-16 kHz	✓	2.540	1.822	84.2	2.737	1.984	2.222	1.280	3.89
Ours	RB	4-16 kHz	✓	**3.550**	**3.013**	**97.3**	**3.657**	**3.726**	**3.355**	**1.112**	**4.43**

Experiment Results

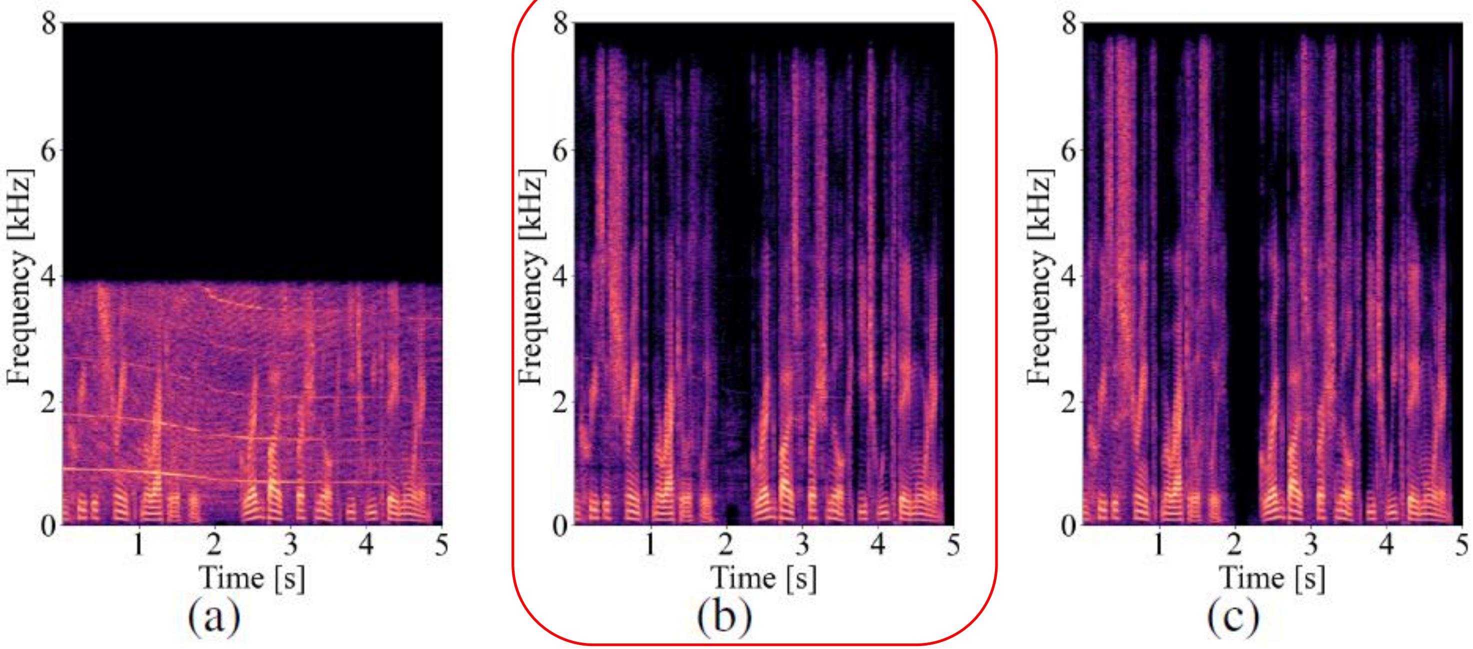
BWE:

Denoise:



Experiment Results

Noise-Robust BWE:



More samples in our demo page:

https://sdnetdemo.github.io/

Conclusions

SDNet Contributions:
- First noise-robust BWE supporting flexible sampling rates.
- Joint optimization for noise reduction and super-resolution.
- Superior performance across diverse scenarios.

Limitations:
- Challenges with higher resolution (e.g., 48 kHz).

Future Work:
- Extend to music datasets and higher resolutions.

E-Mail: s220101187@stu.cqupt.edu.cn
Lab Website: https://hong-qing-liu.github.io/