Aero: Audio Super Resolution in the Spectral Domain

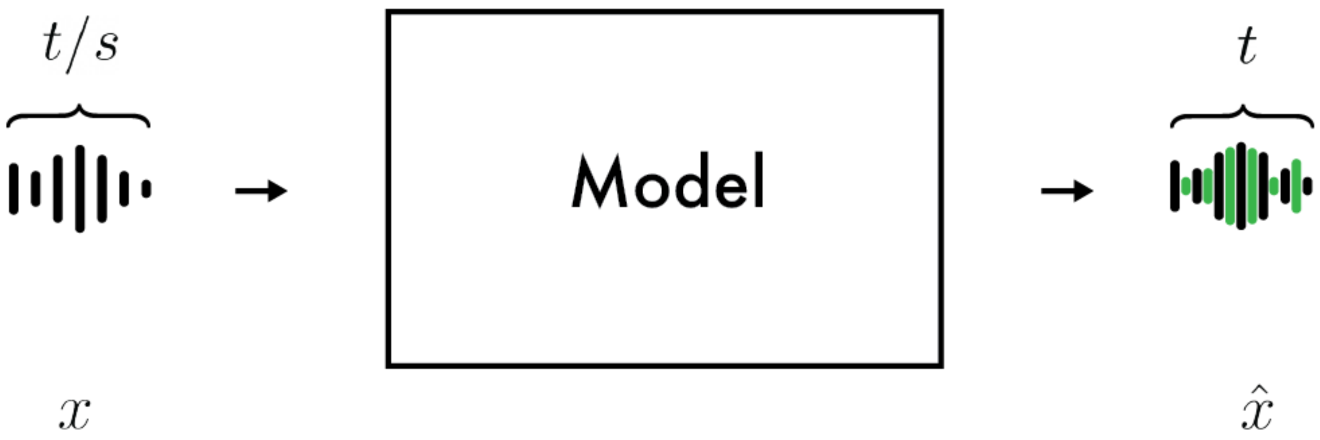
Moshe Mandel, Or Tal, Yossi Adi

Hebrew University of Jerusalem

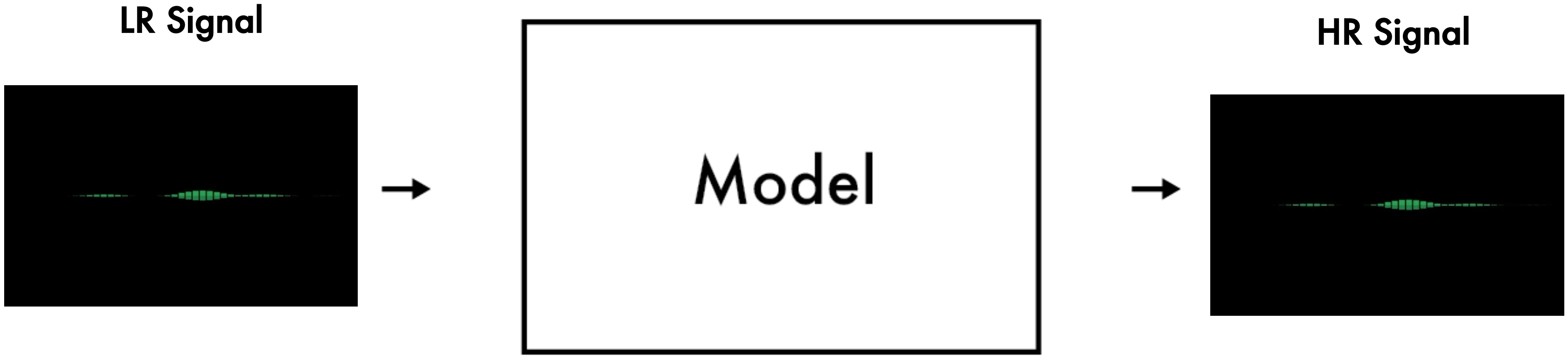
ICASSP 2023

THE HEBREW UNIVERSITY OF JERUSALEM

Audio Super Resolution

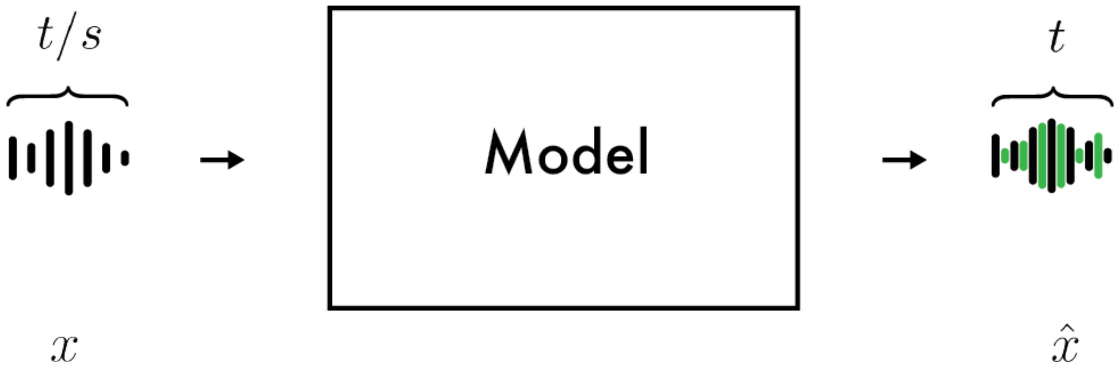


Audio Super Resolution

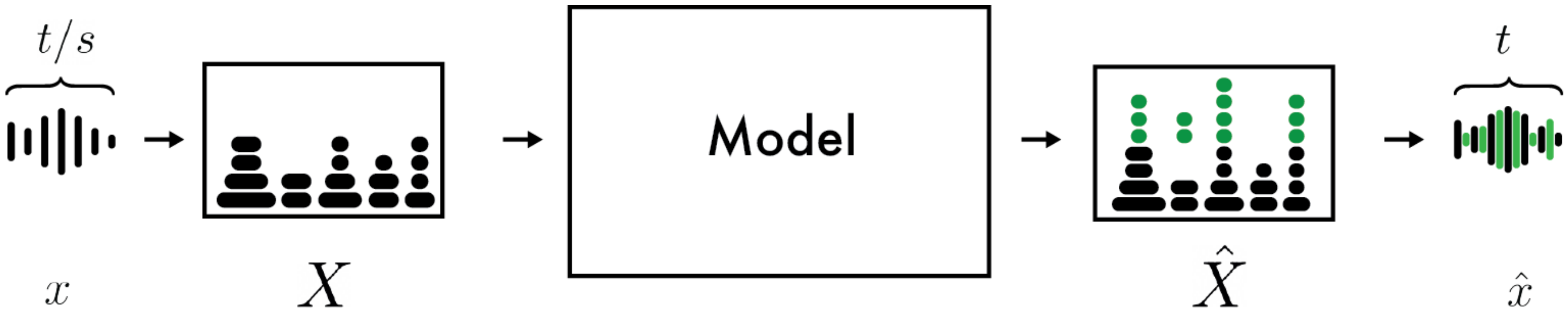


Waveform vs Spectral Methods

Waveform Methods



Spectral Methods



Prior Work

Paper	Input	Method
T-Film, Birnbaum et al. 2017	Waveform	U-Net
SeaNET, Li et al. 2020	Waveform	U-Net, Adversarial
NuWave 2, Han and Lee 2022	Waveform	Diffusion
SSR-GAN, Eskimez et al. 2019	Spectral (Magnitude)	U-Net, Adversarial
NU-GAN, Kumar et al. 2020	Spectral (Magnitude)	U-Net, Adversarial
Phase-Aware, Hu et al. 2020	Spectral (Magnitude, Phase)	U-Net, Adversarial
BEHM-GAN, Moliner and Valimaki 2022	Spectral (Complex)	U-Net, Adversarial

Prior Work

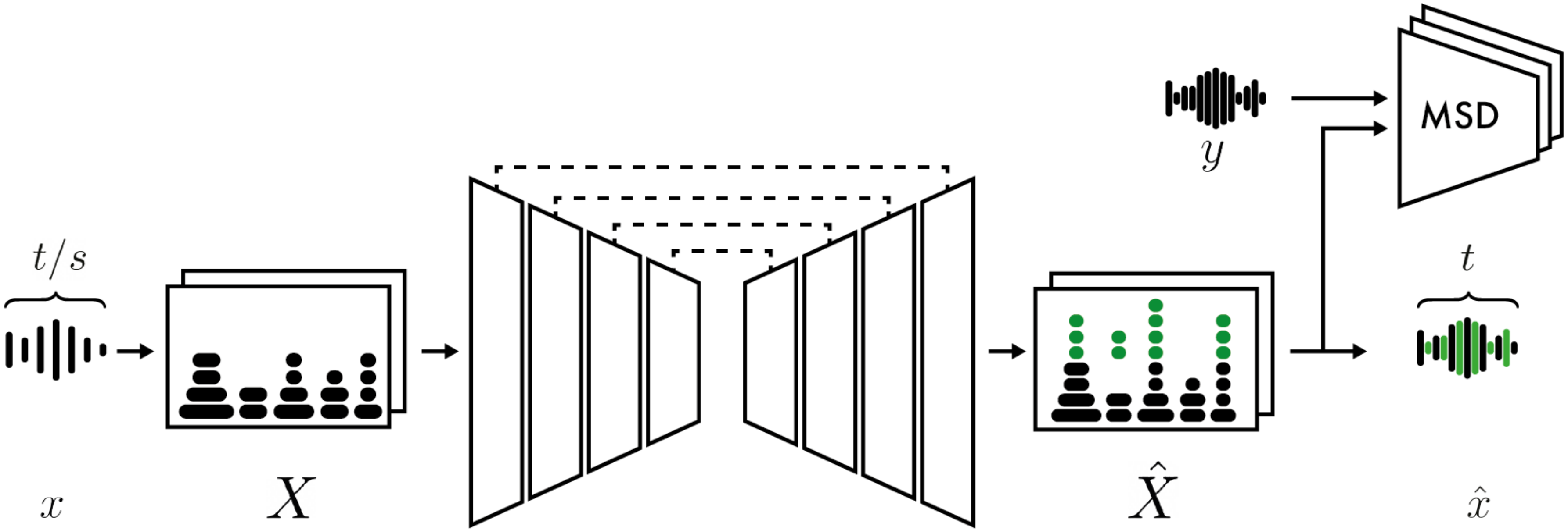
Paper	Input	Method
T-Film, Birnbaum et al. 2017	Waveform	U-Net
SeaNET, Li et al. 2020	Waveform	U-Net, Adversarial
NuWave 2, Han and Lee 2022	Waveform	Diffusion
SSR-GAN, Eskimez et al. 2019	Spectral (Magnitude)	U-Net, Adversarial
NU-GAN, Birnbaum et al. 2020	Spectral (Magnitude)	U-Net, Adversarial
Phase-Aware, Hu et al. 2020	Spectral (Magnitude, Phase)	U-Net, Adversarial
BEHM-GAN, Moliner and Valimaki 2022	Spectral (Complex)	U-Net, Adversarial

Prior Work

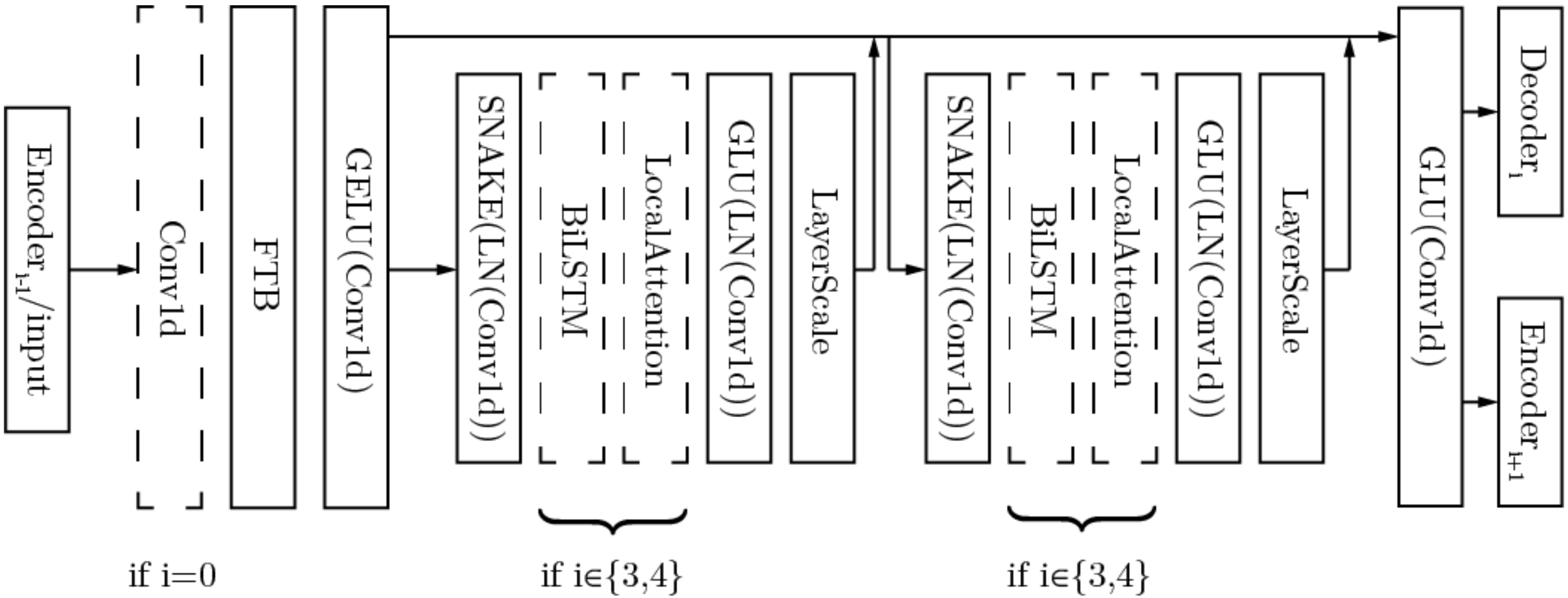
Paper	Input	Method
T-Film, Birnbaum et al. 2017	Waveform	U-Net
SeaNET, Li et al. 2020	Waveform	U-Net, Adversarial
NuWave 2, Han and Lee 2022	Waveform	Diffusion
SSR-GAN, Eskimez et al. 2019	Spectral (Magnitude)	U-Net, Adversarial
NU-GAN, Kumar et al. 2020	Spectral (Magnitude)	U-Net, Adversarial
Phase-Aware, Hu et al. 2020	Spectral (Magnitude, Phase)	U-Net, Adversarial
BEHM-GAN, Moliner and Valimaki 2022	Spectral (Complex)	U-Net, Adversarial

Proposed Model

Proposed Model



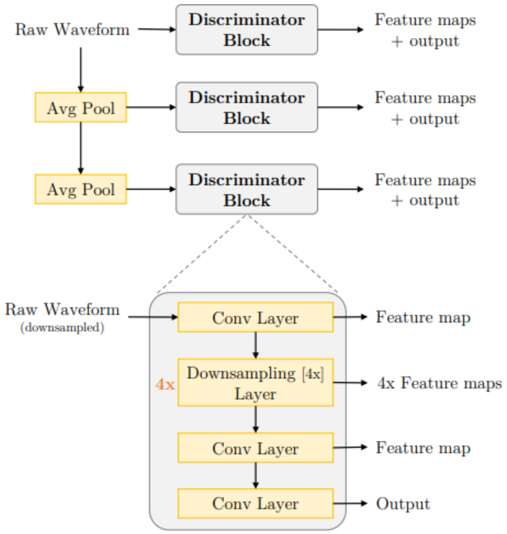
Encoder



Encoder$_{i-1}$/input
Conv1d
FTB
GELU(Conv1d)
SNAKE(LN(Conv1d))
BiLSTM
LocalAttention
GLU(LN(Conv1d))
LayerScale
SNAKE(LN(Conv1d))
BiLSTM
LocalAttention
GLU(LN(Conv1d))
LayerScale
GLU(Conv1d)
Decoder$_i$
Encoder$_{i+1}$
if i=0
if i∈{3,4}
if i∈{3,4}

Adversarial Losses

MelGAN, Kumar et al.



(b) Discriminator

$$\mathcal{L}_G^{\text{adv}} = E_x[\frac{1}{K}\sum_{k,t}\frac{1}{T_k}\max(0, 1 - D_{k,t}(\hat{y}))]$$

$$\mathcal{L}_G^{\text{ft}} = E_x[\frac{1}{KL}\sum_{k,l}\frac{1}{T_{k,l}}\sum_t |D_{k,t}^{(l)}(y) - D_{k,t}^{(l)}(\hat{y})|]$$

$$\mathcal{L}_D = E_y[\frac{1}{K}\sum_k\frac{1}{T_k}\sum_t \max(0, 1 - D_{k,t}(y))] + E_x[\frac{1}{K}\sum_k\frac{1}{T_k}\sum_t \max(0, 1 + D_{k,t}(\hat{y}))]$$

Multi-Resolution STFT Loss

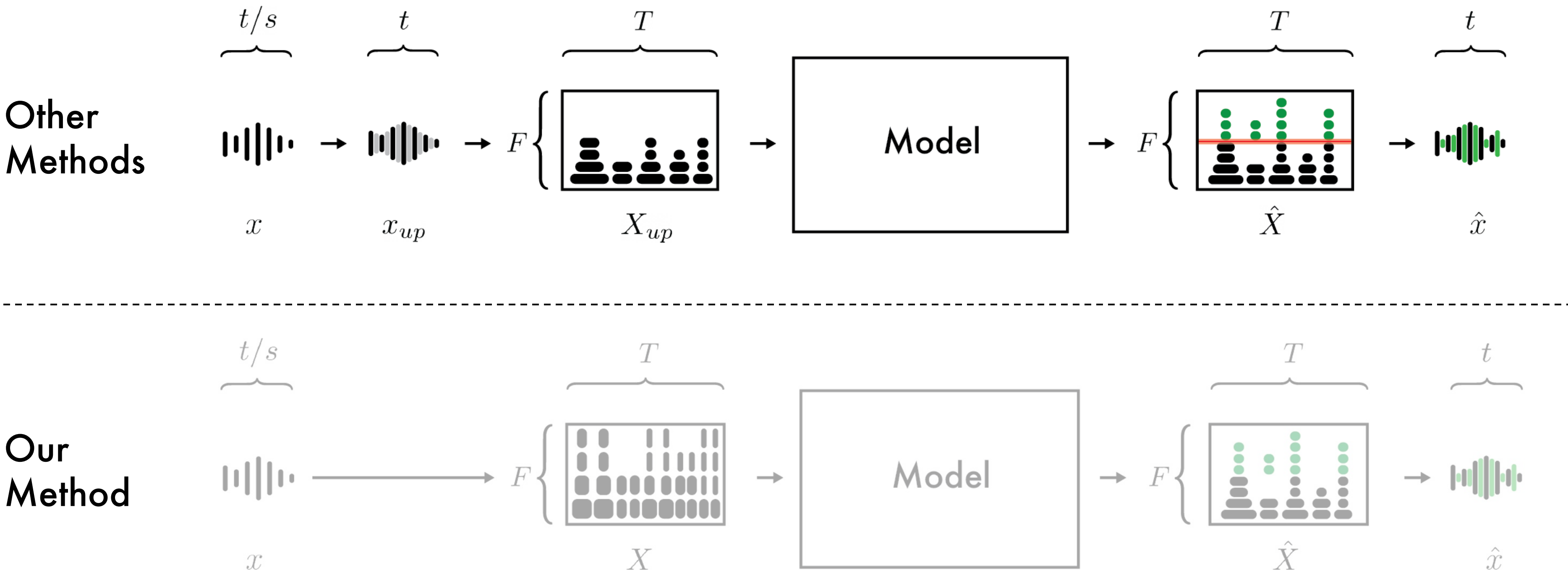
Parallel WaveGAN, Yamamoto et al.

$$\mathcal{L}_G^{\mathrm{stft}} = \sum_{i=1}^{M} L_{\mathrm{sc}}^i(y, \hat{y}) + L_{\mathrm{mag}}^i(y, \hat{y})$$

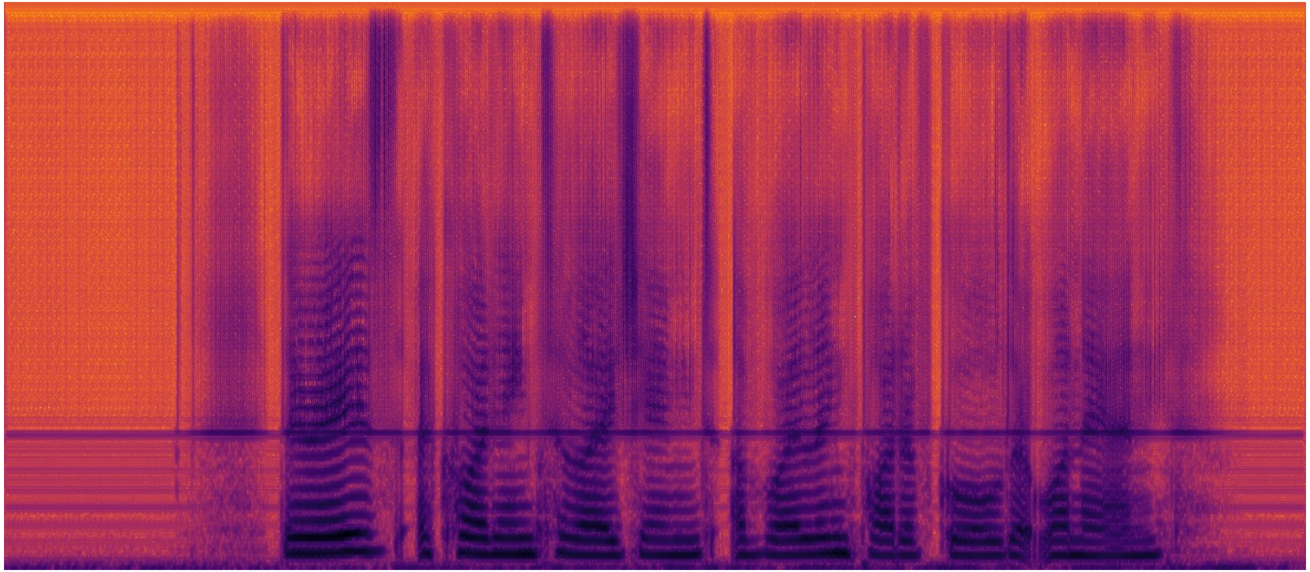
$$L_{\mathrm{sc}}(y, \hat{y}) = \frac{\||STFT(y)| - |STFT(\hat{y})|\|_F}{\|STFT(y)\|_F}$$

$$L_{\mathrm{mag}}(y, \hat{y}) = \frac{1}{T}\| \log |STFT(y)| - \log |STFT(\hat{y})|\|_1$$

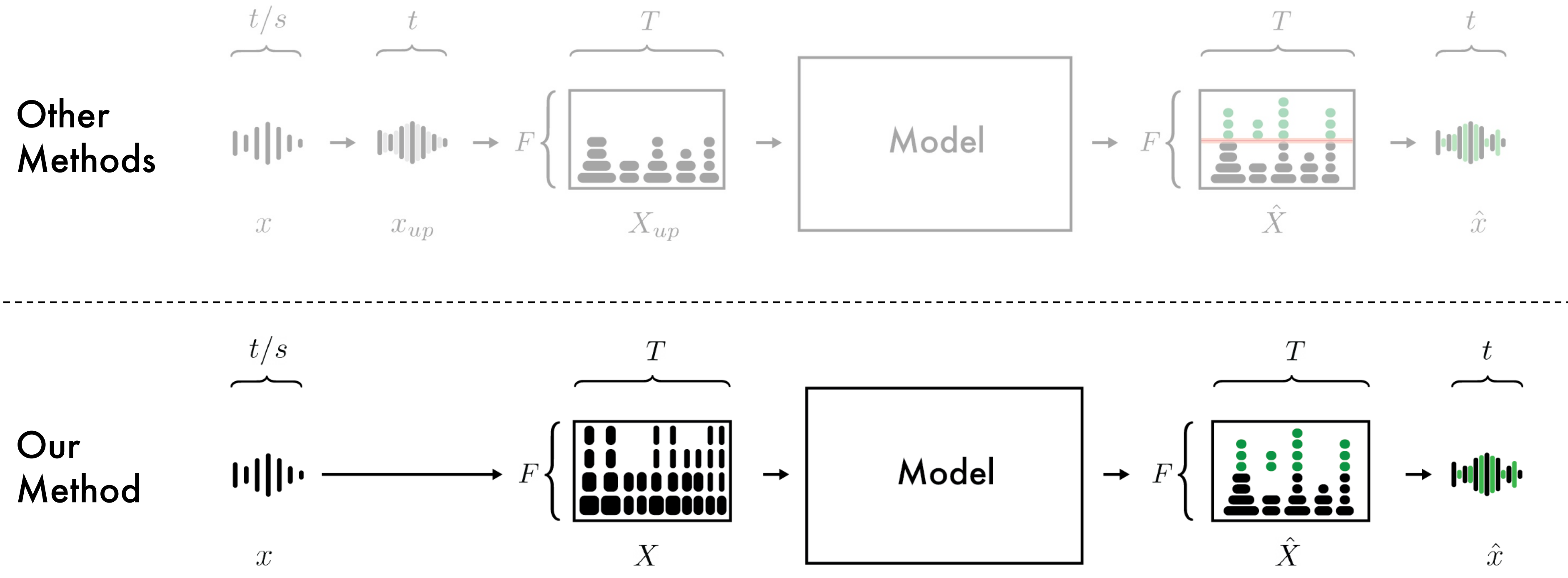
Spectral Upsampling Method



Artifact At Verge



Spectral Upsampling Method



Datasets, Metrics, and Baselines

Datasets

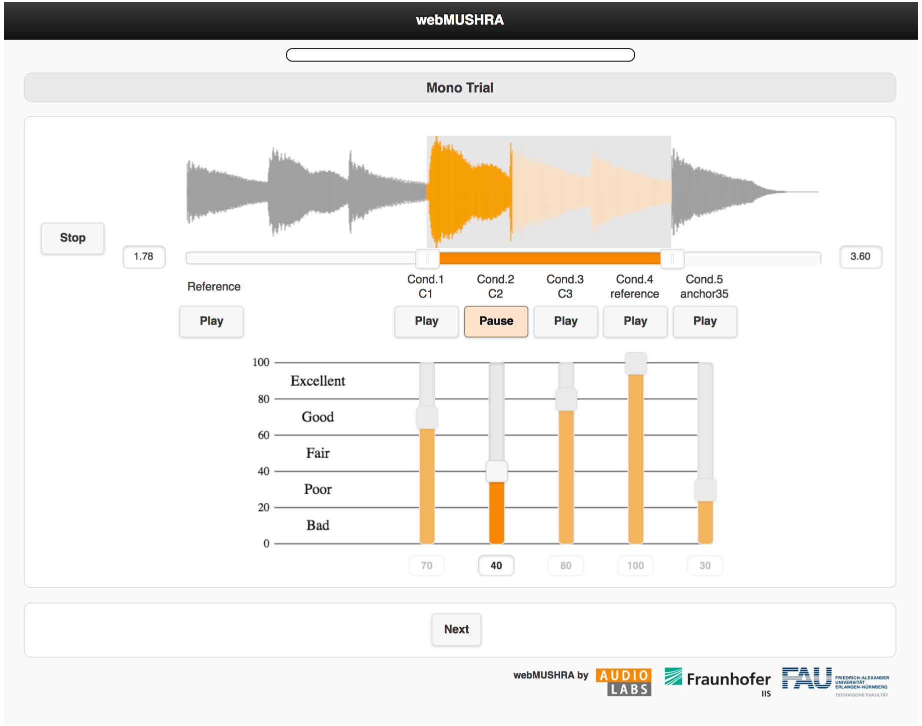
Dataset	Domain	Source to Target Sample Rate (kHz)
VCTK	Speech	8 → 16
		8 → 24
		4 → 16
		12 → 48
MusDB	Music	11 → 44

Metrics

- Log Spectral Distance (LSD)

$$LSD(\hat{y}, y) = \frac{1}{T}\sum_{\tau=1}^{T}\sqrt{\frac{1}{K}\sum_{\kappa=1}^{K}(\hat{Y}(\tau,\kappa) - Y(\tau,\kappa))^2}$$

- Virtual Speech Quality Objective Listener (ViSQOL)
- MUSHRA



Baselines

Baseline	Domain	Input	Method
SeaNET, Li et al.	Speech, Music	Waveform	U-Net, Adversarial
T-Film, Birnbaum et al.	Speech	Waveform	U-Net
NuWave 2, Han and Lee	Speech	Waveform	Diffusion
BEHM-GAN, Moliner and Valimaki	Music	Spectral	U-Net, Adversarial

Results

Results: Speech

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			12-48		
	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	-	-	98.47±0.9
Anchor	-	-	54.65±4.3	-	-	56.21±4.4	-	-	41.14±3.8	-	-	67.76±4.1
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	3.36	4.33	69.77±4.3
TFiLM	1.27	3.18	58.53±4.0	-	-	-	1.77	2.25	41.91±4.0	-	-	-
SEANet	0.79	4.08	91.23±2.9	0.91	4.06	94.16±2.2	0.99	3.16	89.40±3.2	**0.86**	**4.71**	96.17±1.6
NuWave2	-	-	-	-	-	-	-	-	-	1.34	4.42	84.87±4.5
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	0.92	4.67	**96.71**±1.8
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	-	-	-
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	-	-	-

Results: Speech

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

	8-16			8-24			4-16			12-48		
	L↓	V↑	M↑	L↓	V↑	M↑	L↓	V↑	M↑	L↓	V↑	M↑
Reference	-	-	96.25$\pm$1.5	-	-	97.16$\pm$1.4	-	-	96.18$\pm$1.5	-	-	98.47$\pm$0.9
Anchor	-	-	54.65$\pm$4.3	-	-	56.21$\pm$4.4	-	-	41.14$\pm$3.8	-	-	67.76$\pm$4.1
Sinc	2.32	3.41	60.13$\pm$4.7	2.96	3.41	59.49$\pm$4.8	3.59	2.27	43.03$\pm$3.9	3.36	4.33	69.77$\pm$4.3
TFiLM	1.27	3.18	58.53$\pm$4.0	-	-	-	1.77	2.25	41.91$\pm$4.0	-	-	-
SEANet	0.79	4.08	91.23$\pm$2.9	0.91	4.06	94.16$\pm$2.2	0.99	3.16	89.40$\pm$3.2	**0.86**	**4.71**	96.17$\pm$1.6
NuWave2	-	-	-	-	-	-	-	-	-	1.34	4.42	84.87$\pm$4.5
Ours ($^{256}/_{512}$)	0.84	4.02	90.58$\pm$2.3	0.99	4.03	**96.40**$\pm$1.9	1.04	3.04	86.14$\pm$3.4	0.92	4.67	**96.71**$\pm$1.8
Ours ($^{128}/_{512}$)	0.80	4.11	92.63$\pm$2.4	0.91	4.12	95.41$\pm$2.0	0.99	3.15	**92.05**$\pm$**2.7**	-	-	-
Ours ($^{64}/_{512}$)	**0.77**	**4.16**	**94.64**$\pm$1.6	**0.90**	**4.17**	94.45$\pm$2.1	**0.94**	**3.28**	90.61$\pm$3.1	-	-	-

Results: Speech

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			12-48		
	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	-	-	98.47±0.9
Anchor	-	-	54.65±4.3	-	-	56.21±4.4	-	-	41.14±3.8	-	-	67.76±4.1
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	3.36	4.33	69.77±4.3
TFiLM	1.27	3.18	58.53±4.0	-	-	-	1.77	2.25	41.91±4.0	-	-	-
SEANet	0.79	4.08	91.23±2.9	0.91	4.06	94.16±2.2	0.99	3.16	89.40±3.2	**0.86**	**4.71**	96.17±1.6
NuWave2	-	-	-	-	-	-	-	-	-	1.34	4.42	84.87±4.5
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	0.92	4.67	**96.71**±1.8
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	-	-	-
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	-	-	-

Results: Music

	11-44		
	L↓	V↑	M↑
Reference	-	-	95.30±2.5
Anchor	-	-	46.55±7.4
Sinc	3.91	1.97	47.61±8.0
TFiLM [4]	-	-	-
SEANet [5]	1.13	2.88	80.52±7.0
BEHMGAN [17]	1.80	2.01	46.27±8.3
Ours ($^{256}/_{512}$)	1.16	2.88	81.21±6.4
Ours ($^{128}/_{512}$)	1.16	**2.89**	81.67±6.8
Ours ($^{64}/_{512}$)	**1.12**	2.88	**84.18±5.6**

Ablation Study

We evaluate the following:

- Discriminators
- Component study
  - Activation function (ReLU/Snake)
  - Upsampling (Time/Spectral)
  - Frequency Transformer Block (FTB)
- Input size
  - Hop size and window length

Experiments: Ablation Study - Discriminators

Setting	LSD $\downarrow$	VISQOL $\uparrow$	MUSHRA $\uparrow$
Reference	-	-	92.49$\pm$2.2
Anchor	-	-	32.34$\pm$3.3
No disc.	**0.8793**	**3.363**	32.46$\pm$3.5
1 MSD	0.978	3.202	**85.79$\pm$3.0**
3 MSD	0.943	3.275	85.57$\pm$2.9
Only feat. loss	0.986	3.253	77.64$\pm$3.7
Only adv. loss	1.012	3.018	73.96$\pm$4.0

Experiments: Ablation Study - Components

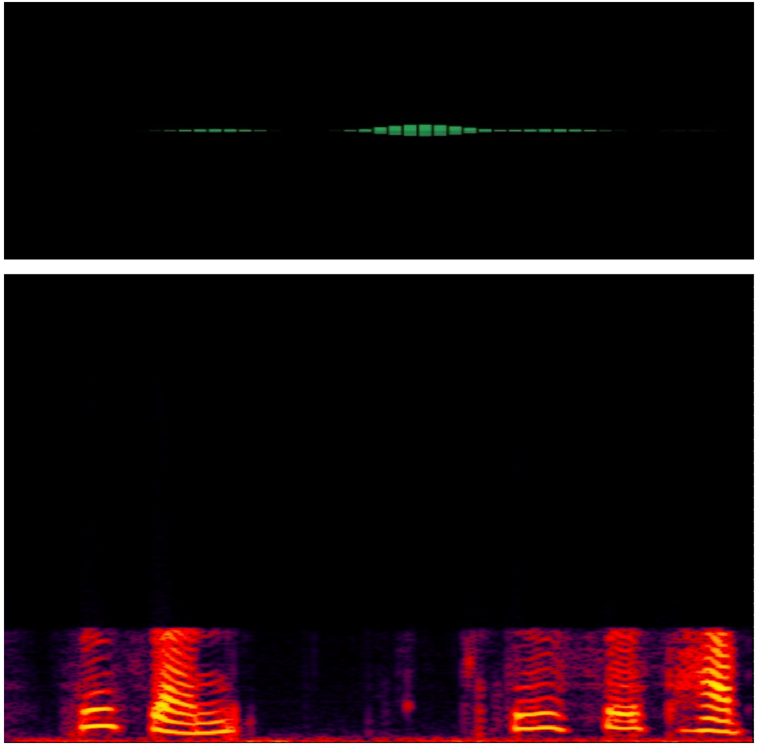
	Activation	Upsampling	FTB	LSD ↓	VISQOL ↑
1	ReLU	spec.	yes	0.945	3.262
2	ReLU	spec.	no	0.952	3.273
3	ReLU	time	yes	0.957	3.263
4	ReLU	time	no	0.948	3.249
5	Snake	spec.	yes	**0.943**	**3.275**
6	Snake	spec.	no	0.958	3.243
7	Snake	time	yes	0.947	3.267
8	Snake	time	no	0.977	3.245

Experiments: Training/Inference Durations

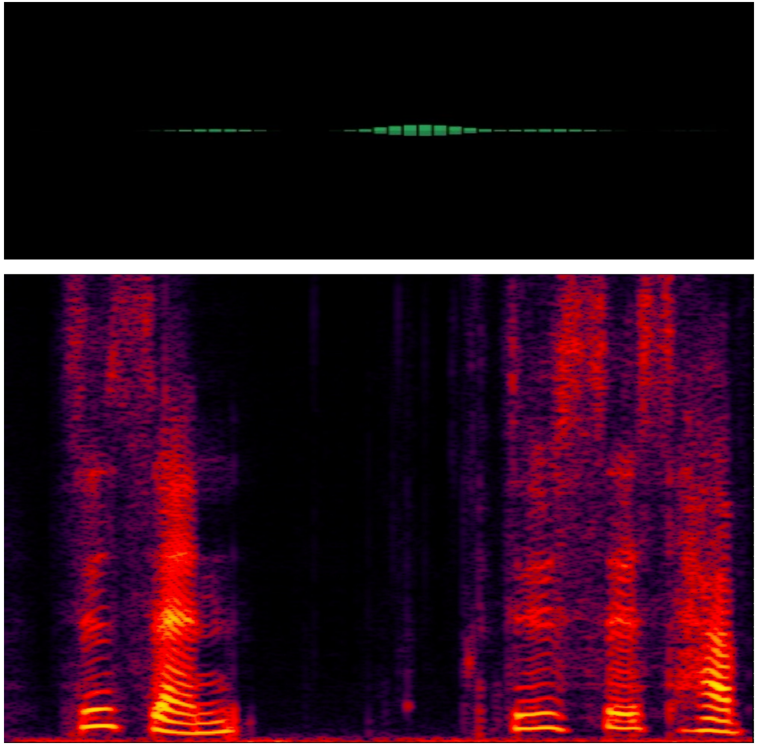
$^{hop}/_{window}$	Training Duration Per Epoch (HH:MM)	Inference Duration (Sec.)
$^{256}/_{512}$	00:35	0.178
$^{128}/_{512}$	00:50	0.449
$^{64}/_{512}$	01:11	1.508

Examples

Low Resolution

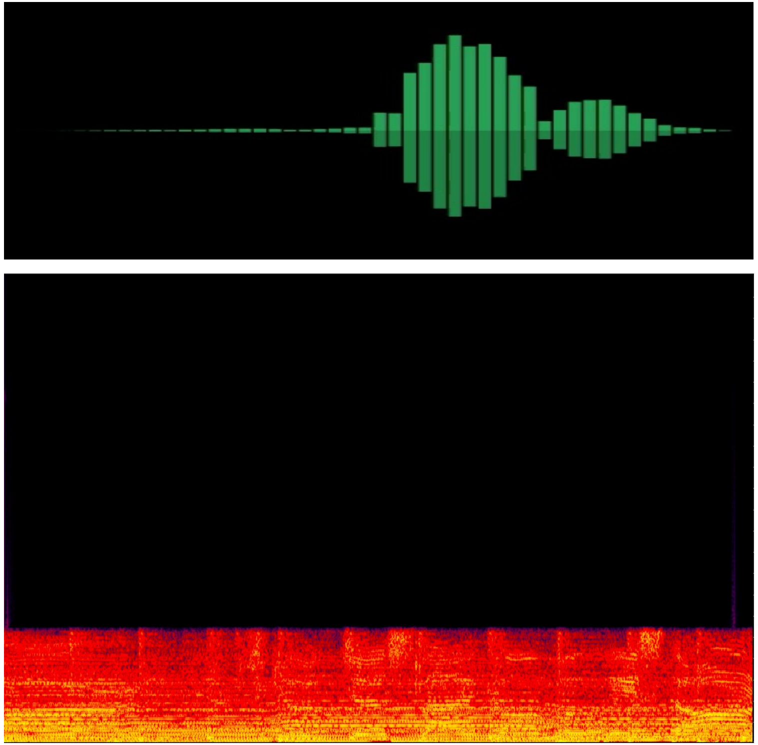


Enhanced

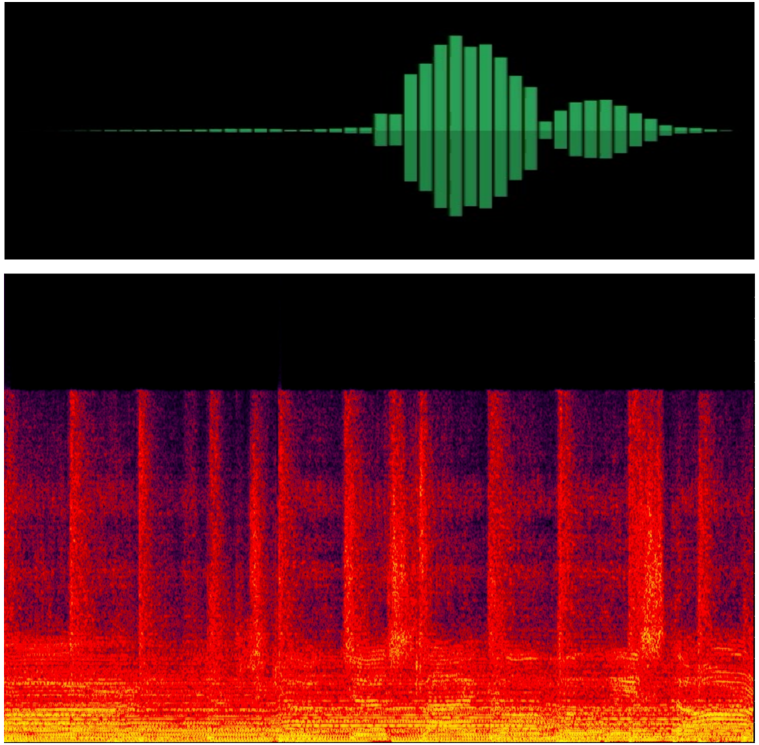


Examples

Low Resolution



Enhanced



Open Source

https://github.com/slp-rl/aero

Conclusions

- Variety of sampling rates and domains.
- State of the art.
- Component ablation study.
- A novel pre-processing approach.

Future Work

- Multiple sample rates in a single model.
- Noisy environments.
- Cross domain generalization.
- Real time inference.
- Simultaneous speech enhancement tasks.

Thank you

See our paper and samples at: https://pages.cs.huji.ac.il/adiyoss-lab/aero/