FINALLY

Fast and Universal Speech with Studio-like Quality

Nicholas Babaev✶*, Kirill Tamogashev✶*, Azat Saginbaev✶, Ivan Shchekotov✶,
Hanbin Bae✶, Hosang Sung✶, WonJun Lee✶, Hoon-Young Cho✶, Pavel Andreev✶*

*Equal Contribution ✶Samsung Research

Mode Collapse and Speech Enhancement

Speech Enhancement aims to restore noisy or degraded sound into the clean one

Common probabilistic formulation:

Learn a distribution of clean signals given a noisy one.

$y \sim pclean(y|x)$, x – noisy signal, y – clean signal

Mode Collapse and Speech Enhancement

Speech Enhancement aims to restore noisy or degraded sound into the clean one

Common probabilistic formulation:

Learn a distribution of clean signals given a noisy one.

$y \sim pclean(y|x)$, x – noisy signal, y – clean signal

Proposed probabilistic formulation:

Learn the most likely clean signal given a noisy one.

$y = argmaxy\ pclean(y|x)$, x – noisy signal, y – clean signal

Mode Collapse and Speech Enhancement

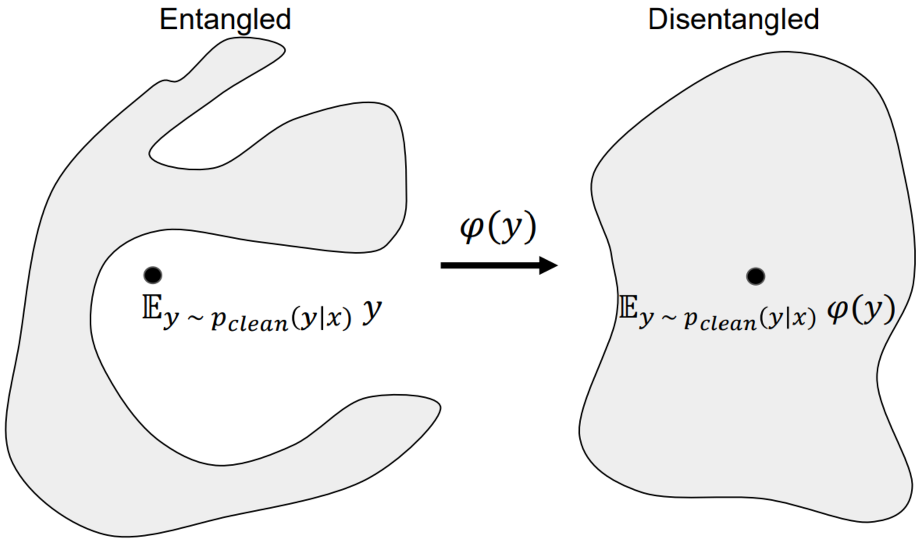
We show that GANs are a natural choice for predicting the main mode of the conditional clean speech distribution $p_{clean}(y|x)$

Proposition 1. *Let* $p_{clean}(y|x) > 0$ *be a finite and Lipschitz continuous density function with a unique global maximum and* $p_g^\xi(y|x) = \xi^n/2^n \cdot \mathbf{1}_{y-g_\theta(x)\in[-1/\xi,1/\xi]^n}$*, then*

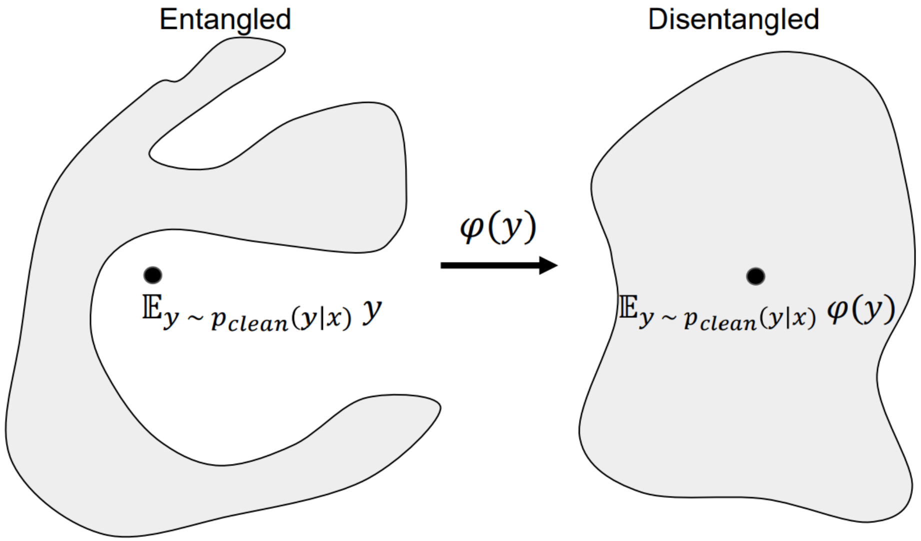
$$\lim_{\xi\to+\infty} \arg\min_{g_\theta(x)} \chi^2_{Pearson}(p_g^\xi||(p_{clean}+p_g^\xi)/2) = \arg\max_y p_{clean}(y|x)$$

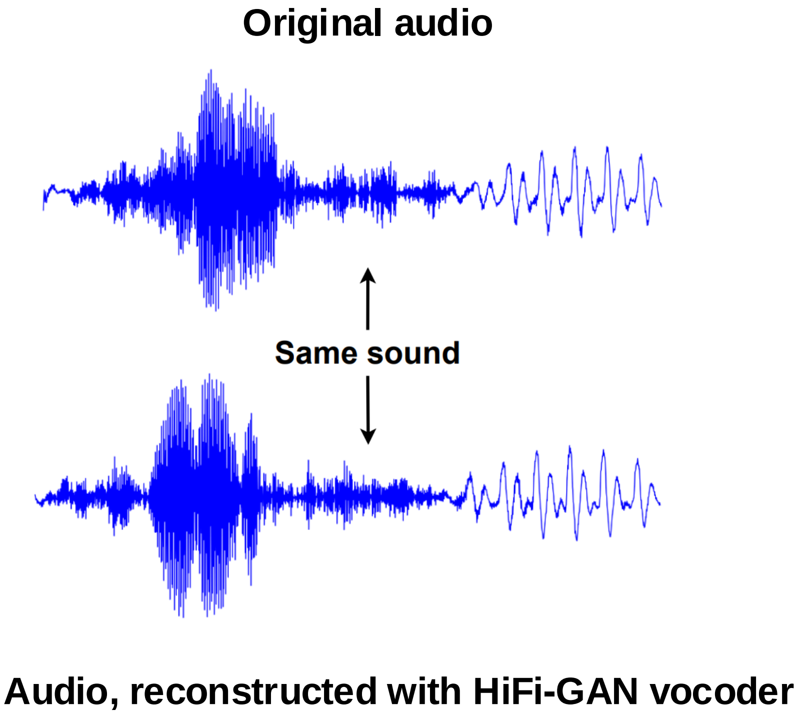
However, **adversarial training is unstable**, therefore, **auxiliary regression** losses are needed to push generator close to the desirable solution.

Practical Aspects: an issue with regression losses



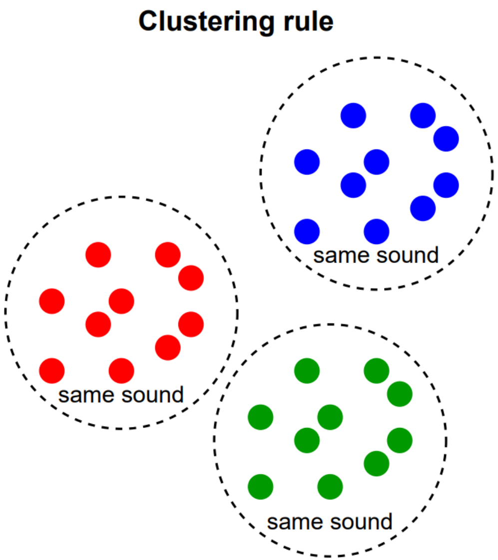
Practical Aspects: an issue with regression losses

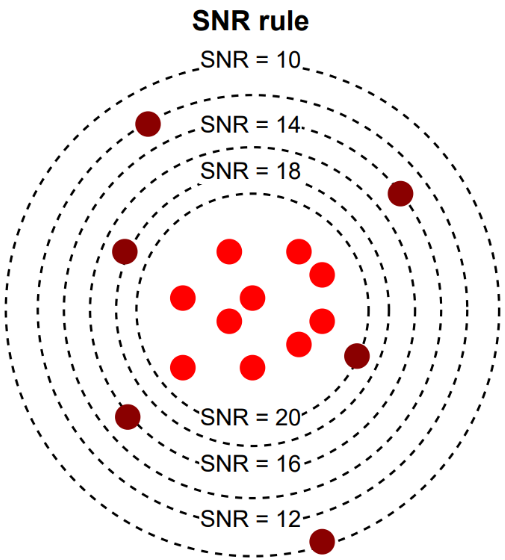




Practical Aspects: rules for an ideal regression loss

1. **Clustering rule**: same speech sounds should be closer to each other, while different speech sounds should be separated.
2. **SNR rule**: the higher the noise level on the noisy signal, the further it should be from its clean version.





Practical Aspects: choice of the mapping ϕ

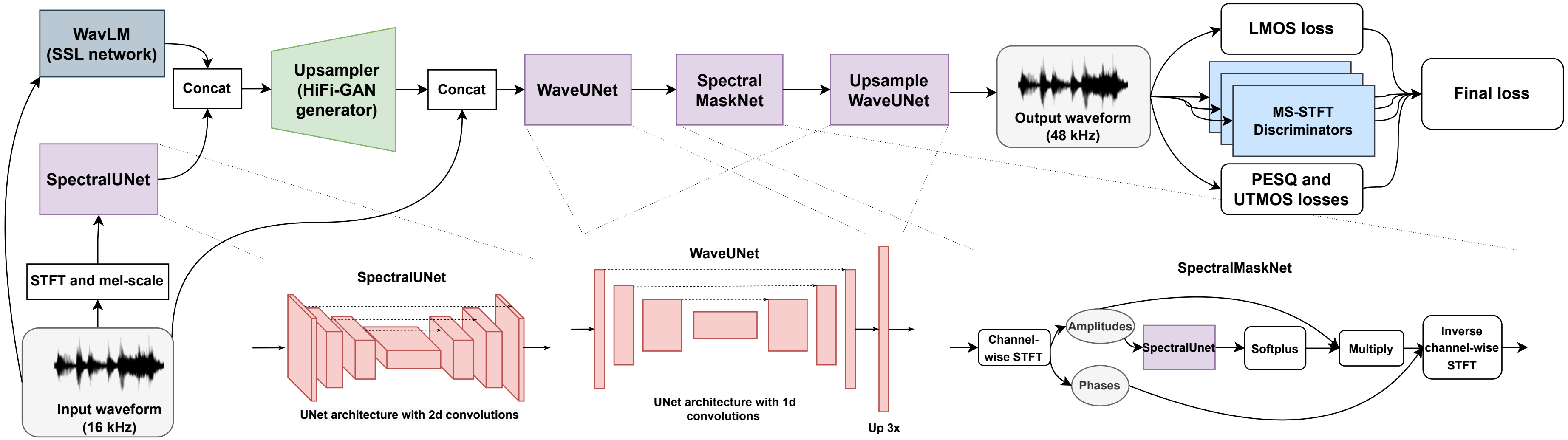
Feature space	Rand score (↑) (Clustering rule)	Negative correlation (↑) (SNR rule)	MOS (↑) (Vocoding)
Waveform	0.00 ± 0.00	0.31 ± 0.02	Failed
Spectrogram	0.00 ± 0.00	0.08 ± 0.03	1.78 ± 0.08
Wav2Vec 2.0	0.25 ± 0.03	0.19 ± 0.03	1.65 ± 0.08
Wav2Vec 2.0-conv	0.94 ± 0.01	0.78 ± 0.02	2.23 ± 0.09
WavLM	0.46 ± 0.05	0.46 ± 0.03	1.71 ± 0.07
WavLM-conv	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{3.27 \pm 0.10}$
EnCodec	0.55 ± 0.03	0.67 ± 0.03	1.80 ± 0.08
CDPAM	0.00 ± 0.00	0.17 ± 0.03	Failed

Practical Aspects: choice of the mapping ϕ

Feature space	Rand score (↑) (Clustering rule)	Negative correlation (↑) (SNR rule)	MOS (↑) (Vocoding)
Waveform	0.00 ± 0.00	0.31 ± 0.02	Failed
Spectrogram	0.00 ± 0.00	0.08 ± 0.03	1.78 ± 0.08
Wav2Vec 2.0	0.25 ± 0.03	0.19 ± 0.03	1.65 ± 0.08
Wav2Vec 2.0-conv	0.94 ± 0.01	0.78 ± 0.02	2.23 ± 0.09
WavLM	0.46 ± 0.05	0.46 ± 0.03	1.71 ± 0.07
WavLM-conv	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{3.27 \pm 0.10}$
EnCodec	0.55 ± 0.03	0.67 ± 0.03	1.80 ± 0.08
CDPAM	0.00 ± 0.00	0.17 ± 0.03	Failed

$$\mathcal{L}_{\mathrm{LMOS}}(\theta) = \mathop{\mathbb{E}}_{x,y\sim p(x,y)} \left[100 \cdot \|\phi(y) - \phi(g_\theta(x))\|_2^2 + \||\mathrm{STFT}(y)| - |\mathrm{STFT}(g_\theta(x))|\|_1\right]$$

FINALLY Architecture



Training in 3 stages

1. Training in 16kHz without Upsample WaveUNet and with regression losses only.

$$\mathcal{L}_{\text{gen}}(\theta) = \underbrace{\lambda_{\text{LMOS}} \cdot \mathcal{L}_{\text{LMOS}}(\theta)}_{\text{1st stage (16 kHz)}}$$

Training in 3 stages

1. Training in 16kHz without Upsample WaveUNet and with regression losses only
2. Adversarial training in 16 kHz without Upsample WaveUNet with GAN loss and features matching loss.

$$\mathcal{L}_{\text{gen}}(\theta) = \overbrace{\underbrace{\lambda_{\text{LMOS}} \cdot \mathcal{L}_{\text{LMOS}}(\theta)}_{\text{1st stage (16 kHz)}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN-gen}}(\theta) + \lambda_{\text{FM}} \cdot \mathcal{L}_{\text{FM}}(\theta)}^{\text{2nd stage (16 kHz)}}$$

$$\mathcal{L}_{\text{disc}}(\varphi_i) = \mathcal{L}_{\text{GAN-disc}}(\varphi_i), \quad i = 1, \ldots, k.$$

Training in 3 stages

1. Training in 16kHz without Upsample WaveUNet and with regression losses only.
2. Adversarial training in 16 kHz without Upsample WaveUNet with GAN loss and features matching loss.
3. Adversarial Training with Upsample WaveUNet in 48 kHz using all previous losses and also Human Feedback losses.

$$\mathcal{L}_{\text{gen}}(\theta) = \underbrace{\overbrace{\underbrace{\lambda_{\text{LMOS}} \cdot \mathcal{L}_{\text{LMOS}}(\theta)}_{\text{1st stage (16 kHz)}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN-gen}}(\theta) + \lambda_{\text{FM}} \cdot \mathcal{L}_{\text{FM}}(\theta)}^{\text{2nd stage (16 kHz)}} + \lambda_{\text{HF}} \cdot \mathcal{L}_{\text{HF}}(\theta)}_{\text{3rd stage (48 kHz)}},$$

$$\mathcal{L}_{\text{disc}}(\varphi_i) = \mathcal{L}_{\text{GAN-disc}}(\varphi_i), \quad i = 1, \ldots, k.$$

Evaluations and Results

Dataset	Model	MOS (↑)	RTF (↓)
VoxCeleb	Input	3.46 ± 0.07	-
	HiFi-GAN-2 (by Adobe)	4.47 ± 0.05	0.5
	Ours	$\mathbf{4.63 \pm 0.04}$	**0.03**
UNIVERSE (validation set)	Input	2.87 ± 0.05	-
	UNIVERSE (by Dolby)	4.10 ± 0.07	0.5
	Ours	$\mathbf{4.23 \pm 0.07}$	**0.03**
LibriTTS	Input	3.59 ± 0.07	-
	MIIPHER (by Google)	4.18 ± 0.06	N/A
	Ours	$\mathbf{4.54 \pm 0.05}$	**0.03**

Evaluations and Results

VoxCeleb (HiFi-GAN-2 validation set, real data)						
Model	MOS (↑)	UTMOS (↑)	WV-MOS (↑)	DNSMOS (↑)	-	RTF (↓)
Input	3.46 ± 0.07	2.76 ± 0.13	2.90 ± 0.16	2.72 ± 0.11	-	-
VoiceFixer	3.41 ± 0.07	2.60 ± 0.09	2.79 ± 0.09	3.08 ± 0.06	-	0.02
DEMUCS	3.79 ± 0.07	3.51 ± 0.08	3.72 ± 0.08	3.27 ± 0.04	-	0.08
STORM	3.75 ± 0.06	3.29 ± 0.08	3.54 ± 0.09	3.17 ± 0.04	-	1.05
BBED	3.97 ± 0.06	3.30 ± 0.10	3.47 ± 0.08	3.23 ± 0.04	-	0.43
HiFi-GAN-2	4.47 ± 0.05	3.67 ± 0.09	$\mathbf{3.96 \pm 0.06}$	$\mathbf{3.32 \pm 0.03}$	-	0.50
Ours	$\mathbf{4.63 \pm 0.04}$	$\mathbf{4.05 \pm 0.07}$	$\mathbf{3.98 \pm 0.06}$	$\mathbf{3.31 \pm 0.04}$	-	**0.03**

Model	MOS (↑)	UTMOS (↑)	WV-MOS (↑)	DNSMOS (↑)	PESQ (↑)	STOI (↑)	SI-SDR (↑)	WER (↓)
Input	3.18 ± 0.07	3.06 ± 0.14	2.99 ± 0.24	2.53 ± 0.10	1.98 ± 0.17	0.92 ± 0.01	8.4 ± 1.2	0.09 ± 0.03
MetricGAN+	3.75 ± 0.06	3.62 ± 0.09	3.89 ± 0.10	2.95 ± 0.05	3.14 ± 0.10	0.93 ± 0.01	8.6 ± 0.7	0.10 ± 0.04
DEMUCS	3.95 ± 0.06	3.95 ± 0.05	4.37 ± 0.06	3.14 ± 0.04	3.04 ± 0.12	$\mathbf{0.95 \pm 0.01}$	18.5 ± 0.6	$\mathbf{0.07 \pm 0.03}$
HiFi++	4.08 ± 0.05	3.89 ± 0.06	4.36 ± 0.06	3.10 ± 0.04	2.90 ± 0.12	$\mathbf{0.95 \pm 0.01}$	17.9 ± 0.6	0.08 ± 0.03
HiFi-GAN-2	4.13 ± 0.05	3.99 ± 0.05	4.26 ± 0.05	3.12 ± 0.05	3.14 ± 0.12	$\mathbf{0.95 \pm 0.01}$	18.6 ± 0.6	$\mathbf{0.07 \pm 0.03}$
DB-AIAT	4.22 ± 0.05	4.02 ± 0.05	4.38 ± 0.06	3.18 ± 0.04	$\mathbf{3.26 \pm 0.12}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{19.3 \pm 0.8}$	$\mathbf{0.07 \pm 0.03}$
Ours (16 kHz)	$\mathbf{4.41 \pm 0.04}$	$\mathbf{4.32 \pm 0.02}$	$\mathbf{4.87 \pm 0.05}$	$\mathbf{3.22 \pm 0.04}$	2.94 ± 0.10	0.92 ± 0.01	4.6 ± 0.3	$\mathbf{0.07 \pm 0.03}$
Ours (48 kHz)	$\mathbf{4.66 \pm 0.04}$	$\mathbf{4.32 \pm 0.02}$	$\mathbf{4.87 \pm 0.05}$	$\mathbf{3.22 \pm 0.04}$	2.94 ± 0.10	0.92 ± 0.01	4.6 ± 0.3	$\mathbf{0.07 \pm 0.03}$
GT (16 kHz)	4.26 ± 0.05	4.07 ± 0.04	4.52 ± 0.04	3.16 ± 0.04	–	–	–	–
GT (48 kHz)	4.56 ± 0.03	4.07 ± 0.04	4.52 ± 0.04	3.16 ± 0.04	–	–	–	–

For more in formation, please, visit our demo.