Loud-loss

A perceptually motivated loss function for speech enhancement based on equal-loudness contours

Zixuan Li[1], **Xueliang Zhang**[1]*, Changjiang Zhao[1], Shuai Gao[1], Lei Miao[2], Zhipeng Yan[2], Sun Ying[2], Zhu Chong[2]

[1] College of Computer Science, Inner Mongolia University, Hohhot, China
[2] Lenovo, Beijing, China

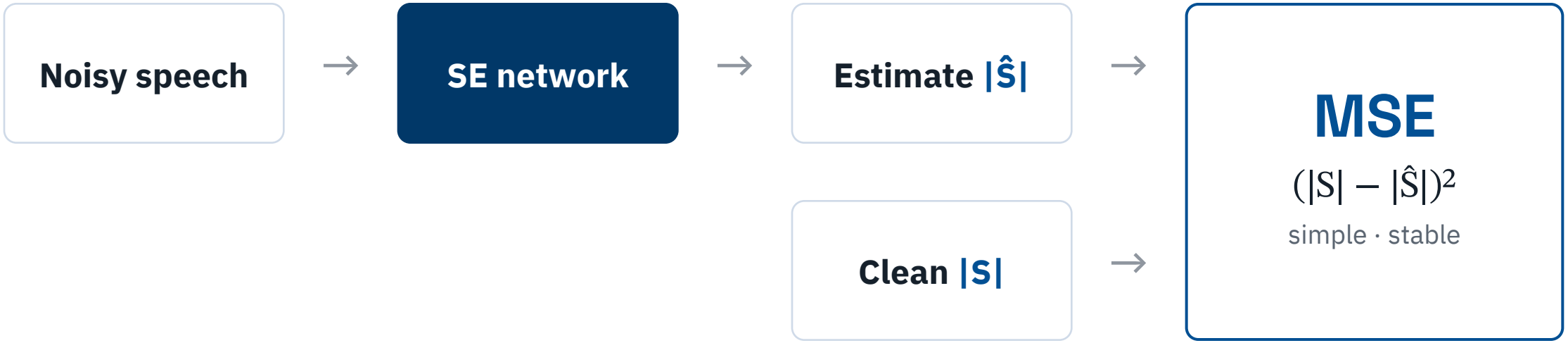
21st NCMMSC (2026)





SE models are usually trained with MSE

Background

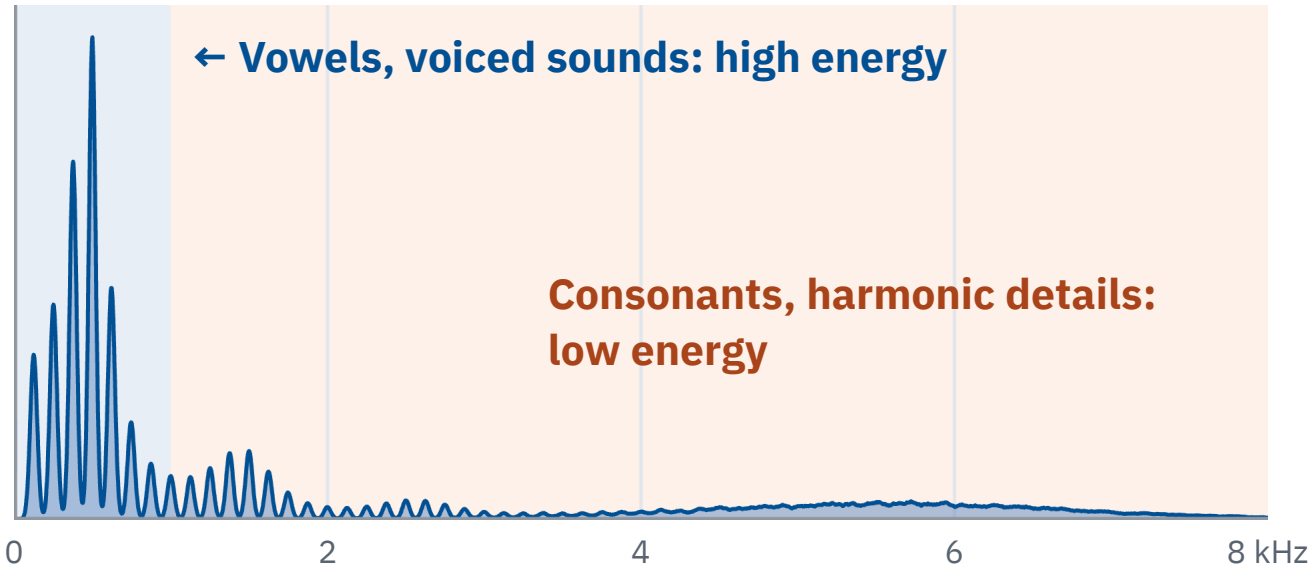


But MSE correlates imperfectly with human perception.

MSE has two key limitations

Problem

Limitation 1 · uniform penalty across frequency bins



Illustrative voiced frame, linear magnitude

Every bin gets the same penalty

↓

Loss falls mainly by fitting loud low frequencies

↓

Quiet, informative parts stay blurred

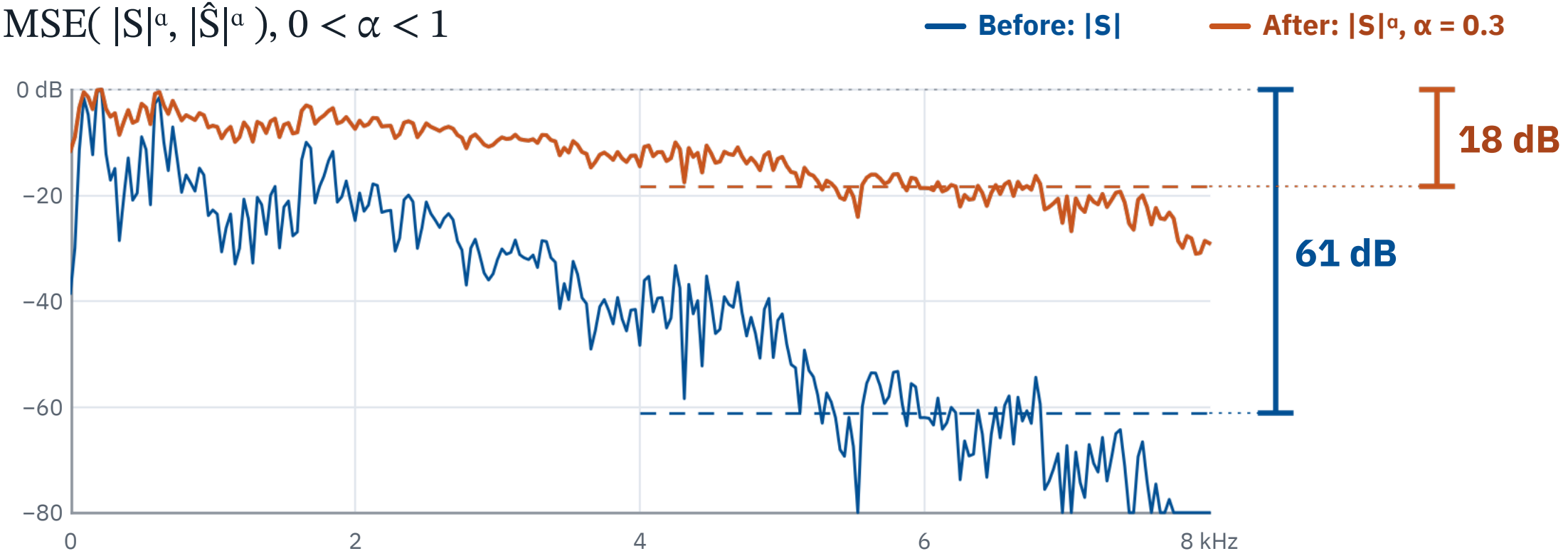
Limitation 2

Smaller spectral distance ≠ better perceived quality

Remedy 1: compress the magnitude

Related work · power-law compression

$$\mathrm{MSE}(\,|S|^{\alpha},\,|\hat{S}|^{\alpha}\,),\ 0<\alpha<1$$



One voiced frame of clean utterance p232_010 · range = frame peak to median level above 4 kHz (dashed)

+ Every level gap shrinks to 30%: 61 dB → 18 dB, so quiet parts count more

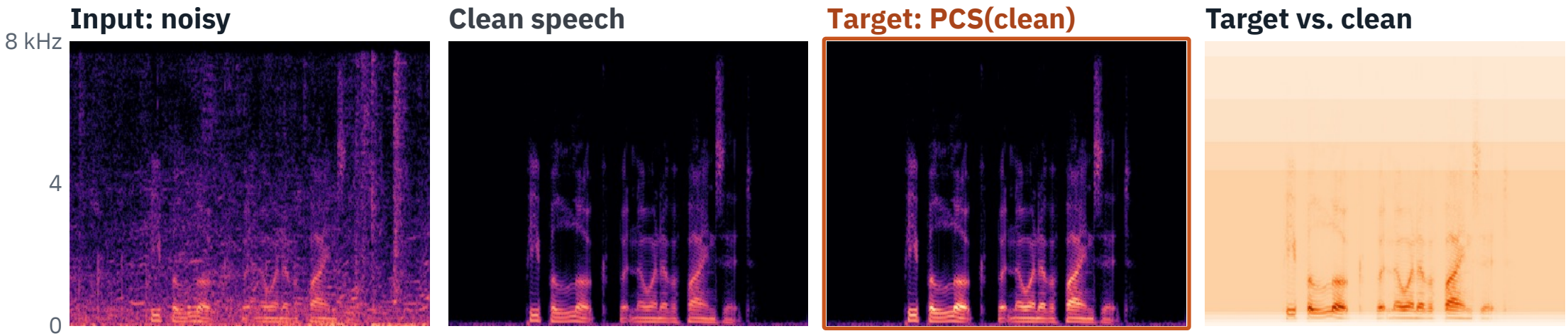
But it only rescales the amplitude, not where the ear is sensitive; α is empirical.

Remedy 2: stretch the target (PCS)[1]

Related work · perceptual contrast stretching

$$\text{Target} = \exp(\, w(f) \cdot \log(1 + |S|)\,) - 1$$

w(f): band-importance weights from 1.0 to 1.4 (largest at 0.4–4.4 kHz)



VoiceBank test utterance p232_010 (0.9 dB SNR), in dB

white 0 dB → orange +12 dB

> **Input–target mismatch:** **the network maps a natural noisy input to a stretched target, not to clean speech.**

[1] Chao, R., Yu, C., Fu, S. W., Lu, X., & Tsao, Y. (2022). Perceptual contrast stretching on target feature for speech enhancement. *arXiv preprint arXiv:2203.17152.*

Remedy 3: perceptual losses need extra networks

Related work · PFPL, MetricGAN

PFPL[1]

feature distance

$\hat{s}, s$ waveforms → **wav2vec** pretrained, frozen · 32.5 M params → **Wasserstein distance** → **loss**

MetricGAN[2]

learned metric

$\hat{s}, s$ spectra → **Discriminator D** learned in training · 0.18 M params → **predicted PESQ** → **loss**

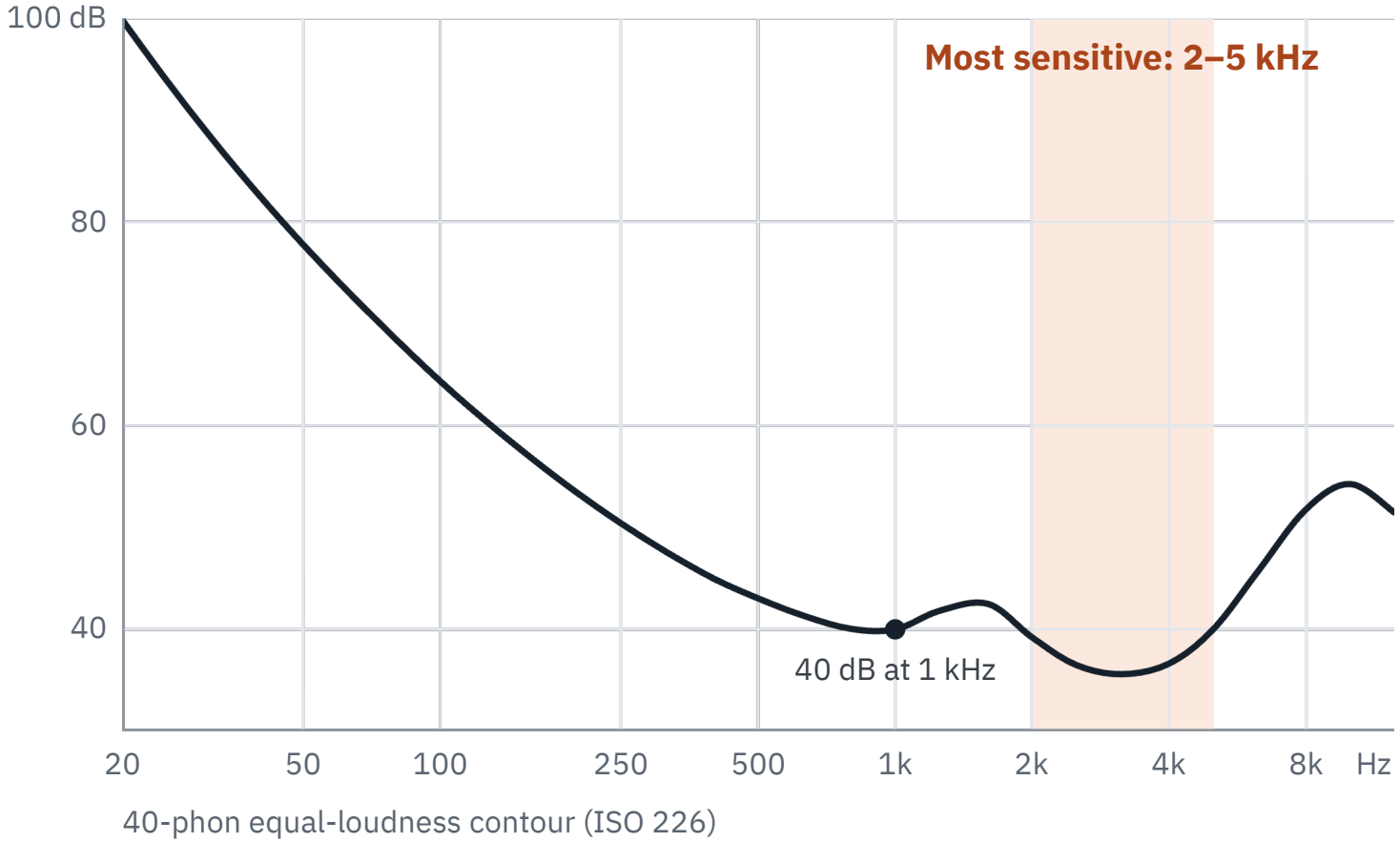
trained to match true PESQ scores

We want a loss that is grounded in hearing, keeps the clean target and needs no extra network.

[1] Hsieh, T. A., Yu, C., Fu, S. W., Lu, X., & Tsao, Y. (2020). Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement. arXiv preprint arXiv:2010.15174.
[2] Fu, S. W., Liao, C. F., Tsao, Y., & Lin, S. D. (2019, May). Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In International Conference on Machine Learning (pp. 2031-2041). PmLR.

Weight errors by the equal-loudness contour

Our proposal · Loud-loss



40-phon equal-loudness contour (ISO 226)

Lower curve = more sensitive ear

Loud-loss weights errors by this curve

Larger penalty where the ear is more sensitive

No extra network · clean target unchanged

Build the loss around how we hear

The idea

1

Loudness is perceived on a log scale

↓

Errors in dB

2

Frequency resolution is finer at low frequencies

↓

Mel sub-bands

3

Sensitivity varies with frequency

↓

Equal-loudness weights

Loud-loss is computed in four steps

Method

$|\hat{S}|, |S|$
magnitudes

→

1
Log power
$10 \log_{10} |S|^2$

→

2
Mel sub-bands
25 bands
50% overlap

→

3
Sub-band MSE
mean $(P - \hat{P})^2$

→

4
Loudness weights
$\mathcal{L} = \Sigma\, w_i\, \mathcal{L}_i$

$$w_i = \mathrm{SPL}(1\ \mathrm{kHz}) \,/\, \mathrm{SPL}(f_i)$$

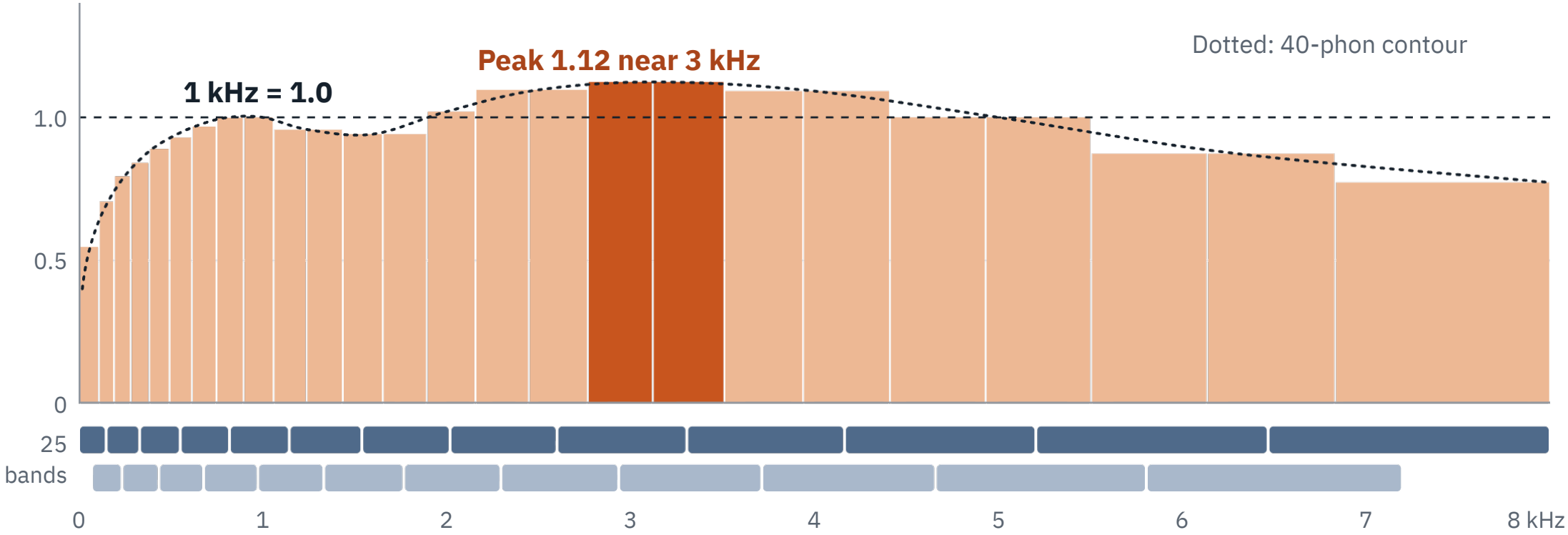
SPL read from the 40-phon equal-loudness contour

No extra network · the clean spectrum stays the target

Weights come from the 40-phon contour

Method

$$w_i = \mathrm{SPL}(1\ \mathrm{kHz}) / \mathrm{SPL}(f_i)$$



Fixed weights from a published contour: nothing to learn.

Setup: two models, two datasets

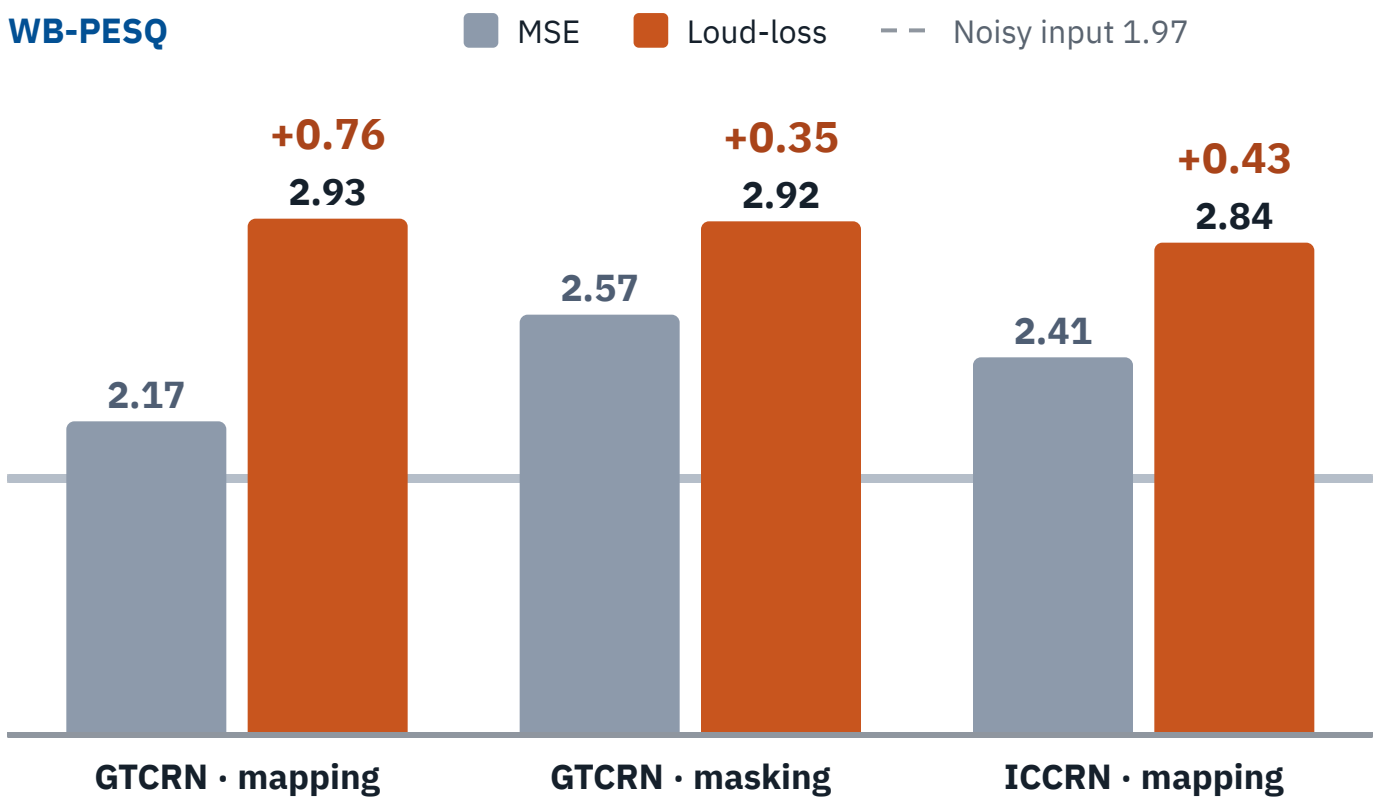
Experiments

Models	GTCRN (low complexity, no ERB) · ICCRN (medium)
Output	Magnitude by mapping or masking; noisy phase reused
Data	VoiceBank+DEMAND · DNS3 blind test
Metrics	PESQ, ESTOI, SI-SNR · DNSMOS, NISQA

Same training recipe for every loss: only the loss changes.

Loud-loss beats MSE on every model

Results · VoiceBank+DEMAND



Bars start at 1.0

+0.76

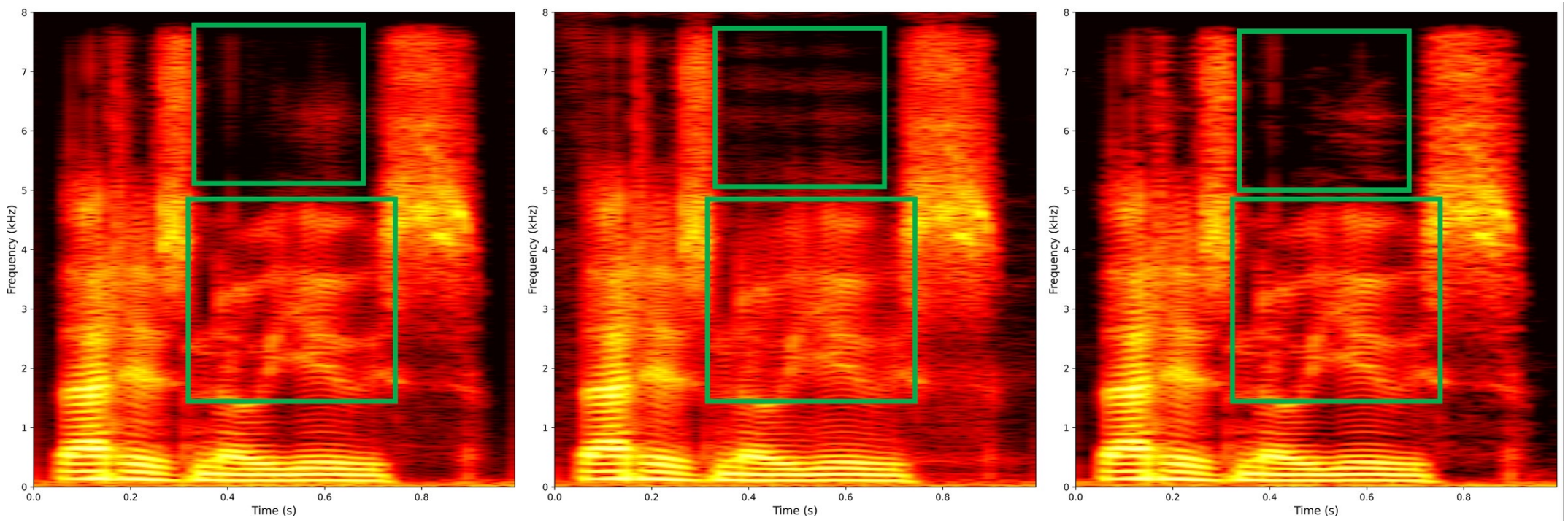
WB-PESQ 2.17 → 2.93

GTCRN · mapping, same network and recipe

High-frequency detail is better preserved

Results · VoiceBank+DEMAND

Clean reference **MSE** **Loud-loss**



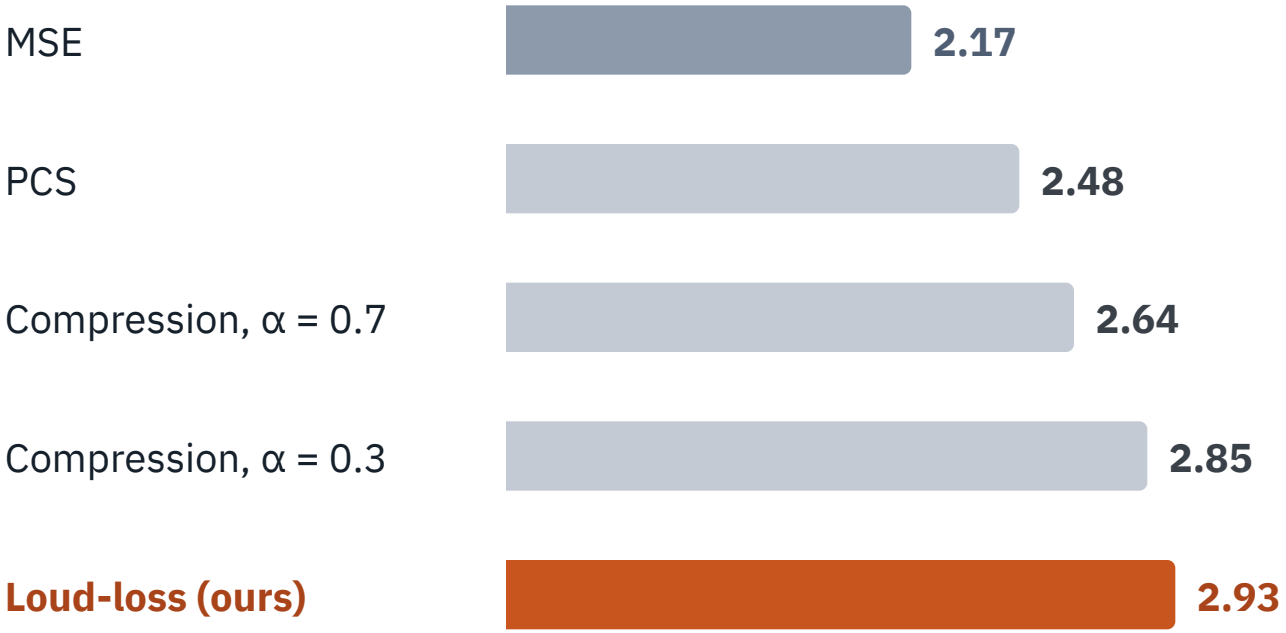
1.5–5 kHz: MSE smears the harmonics

5–8 kHz: MSE leaves streaky residue

Loud-loss beats other spectral remedies

Results · VoiceBank+DEMAND

WB-PESQ, GTCRN with mapping



Bars start at 1.0

Adding an SI-SNR term

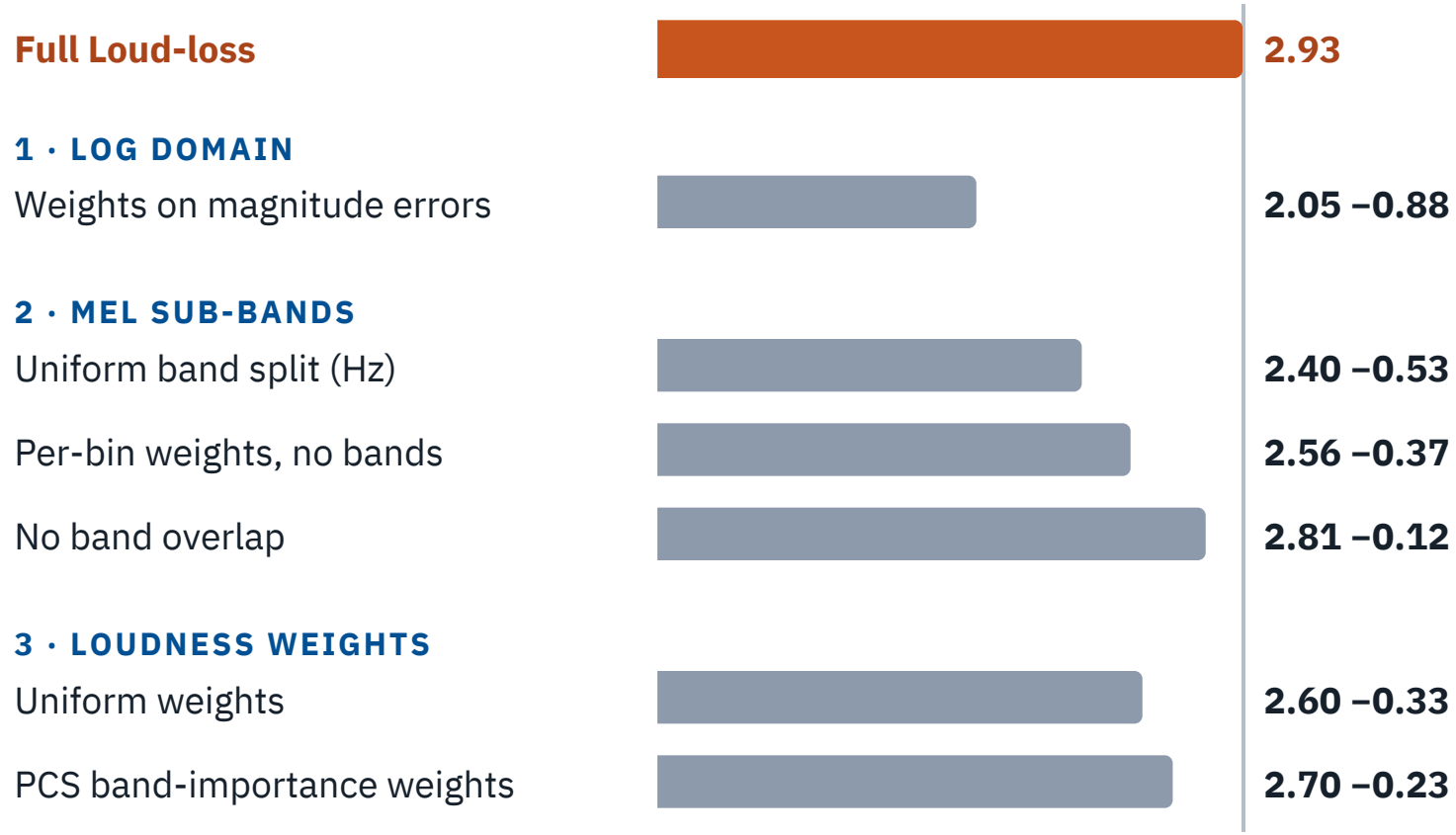
Loss	PESQ	SI-SNR
MSE	2.17	17.24
Loud-loss	2.93	14.41
+ SI-SNR	2.92	18.44

Best SI-SNR of all systems, with PESQ kept

Compression depends on a hand-tuned α; Loud-loss needs no such tuning.

Every design choice contributes

Ablation · GTCRN, WB-PESQ



Bars start at 1.0

Gains hold on the DNS3 blind test

Generalization · DNS3 blind test

	DNSMOS				NISQA	
Loss	**P.808**	**SIG**	**BAK**	**OVRL**	**MOS**	**Noise**
MSE	3.0477	2.9139	3.0105	2.3416	2.3468	2.0555
PFPL	3.2226	2.9757	3.4343	2.5276	2.4191	2.8315
MetricGAN	3.2266	2.9424	3.4958	2.5285	2.5918	2.9450
Loud-loss	3.2272	2.8972	3.6656	2.5500	2.5189	3.0390

Orange = best in column. NISQA Coloration, Discontinuity and Loudness are in the paper.

Best P.808, BAK, OVRL and NISQA Noise, with no auxiliary network.

Training cost stays at the MSE level

Training cost

Method	Extra network	Training-step time vs. MSE
MSE	None	**1.00×**
PFPL	Frozen wav2vec, 32.6 M	**3.07×**
MetricGAN	Discriminator, 0.21 M	**8.84×**
Loud-loss	None	**1.05×**

MSE and Loud-loss: 0.028 M parameters, 313.6 MB peak memory. 10-s utterance, batch 1, one RTX 5090.

1.05× the step time of MSE, with no extra network.

Summary and outlook

Conclusion

Problem	MSE over-emphasizes high-energy low frequencies
Method	Log-power errors in Mel sub-bands, weighted by the equal-loudness contour
Result	WB-PESQ **2.17 → 2.93** on GTCRN, at 1.05× the cost of MSE

github.com/zx1292982431/Loud-Loss

Scan for the code

> **Drop-in replacement for MSE: the inference model is untouched.**

Next: other loudness levels, adaptive contour selection

Thank you

Questions and discussion are welcome.

Zixuan Li · cslzx@mail.imu.edu.cn
Xueliang Zhang (corresponding author) · cszxl@imu.edu.cn

Loud-loss: A Perceptually Motivated Loss Function for Speech Enhancement Based on Equal-Loudness Contours · NCMMSC 2026

ACKNOWLEDGMENTS

Supported by the Inner Mongolia Natural Science Foundation (2025LHMS06005), the Hohhot R&D Investment Incentive Program (2025-Gui-Yan-1), the Inner Mongolia Postgraduate Research Project (KC2025006S), and the CCF-Lenovo Research Fund (No. 20240203).

INNER MONGOLIA UNIVERSITY

Lenovo

CODE · OPEN SOURCE

github.com/zx1292982431/Loud-Loss

Scan to open the repository