Dolby

RF-GML: Reference-Free Generative Machine Listener

—

ARIJIT BISWAS AND GUANXIN JIANG

ICASSP, HYDERABAD, INDIA, 06-11 APRIL 2025

Motivation – addressing current limitations

Goal: Evaluate coded audio (48 kHz sample rate; mono/stereo/binaural) without unencoded reference.

Limitations of existing reference-free (RF) metrics:

- Primarily speech-centric (designed mostly for sample rates below 48 kHz).
- Lack of accuracy for diverse content types and coded audio.
- Complexity of metrics using *non-matching** reference signals.

Our solution: Leverages transfer learning from a state-of-the-art full-reference (FR) Generative Machine Listener (GML)**.

*A. Ragano, J. Skoglund, and A. Hines, "NOMAD: Unsupervised Learning of Perceptual Embeddings for Speech Enhancement and Non-Matching Reference Audio Quality Assessment," *ICASSP*, 2024.
G. Jiang, L. Villemoes, and A. Biswas, "Generative Machine Listener**," *155th AES Convention*, 2023.

RF-GML

Aims at simulating scores s of an arbitrary number of listeners.

We use a two-parameter model of $p(s|y)$ and train with maximum likelihood.

Pros

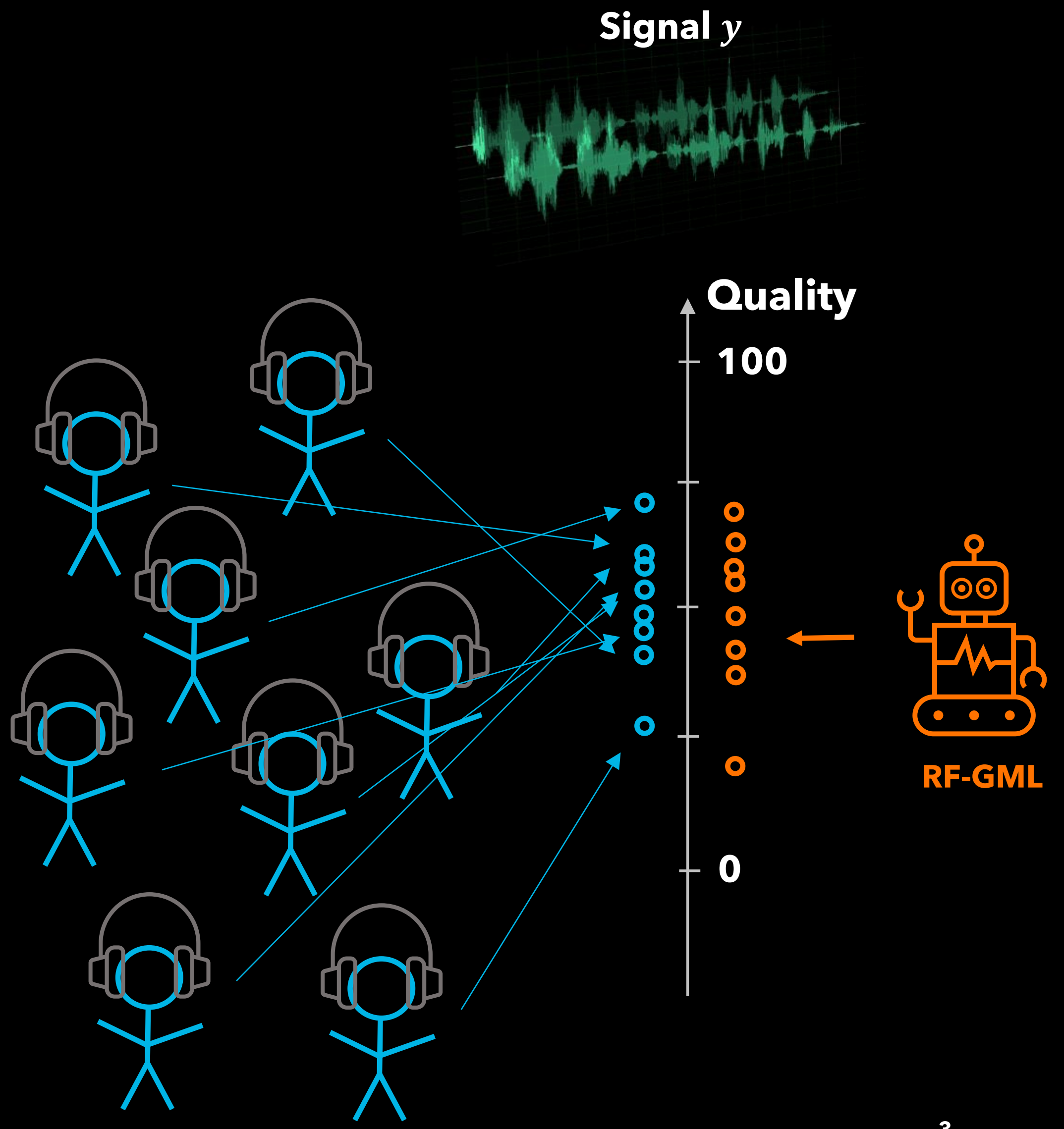
✓ Uses individual quality scores from the dataset

✓ Can capture confidence intervals (not evaluated in this work).

Con

➢ More demanding to train than the mean score regression.

How good is the quality of signal y?



MODEL

GML to RF-GML

- ❖ Removed reference signal channels.
- ❖ Transfer weights from GML (first In-A block).

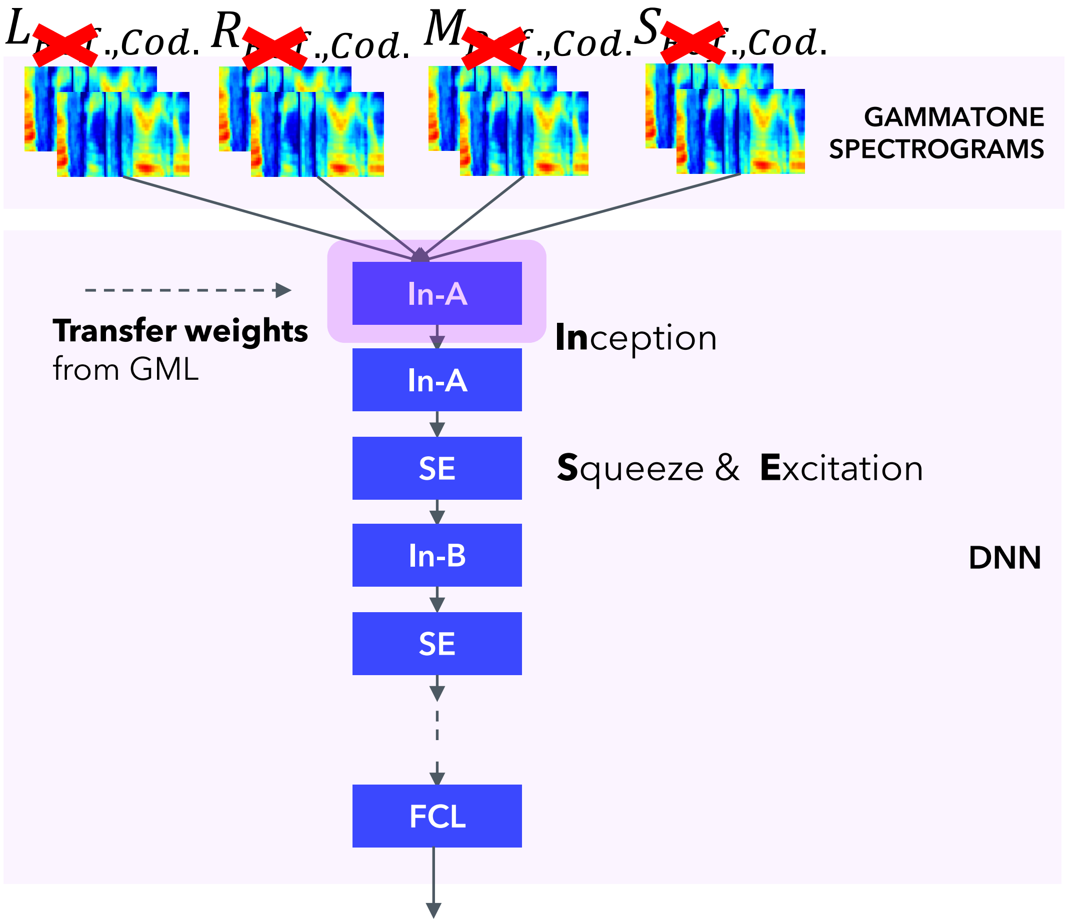
Loss (uses individual quality scores): Negative log likelihood (NLL)

- ❖ $-\log p(s|y)$

	pdf	loss
Logistic $(\mu, \log a)$	$\frac{1}{4a}\operatorname{sech}^2\left(\frac{s-\mu}{2a}\right)$	$\log 4a + 2\log \operatorname{sech}\left(\frac{s-\mu}{2a}\right)$

Data augmentation

- ➢ Swap L and R; but keep same quality
- ➢ CutMix

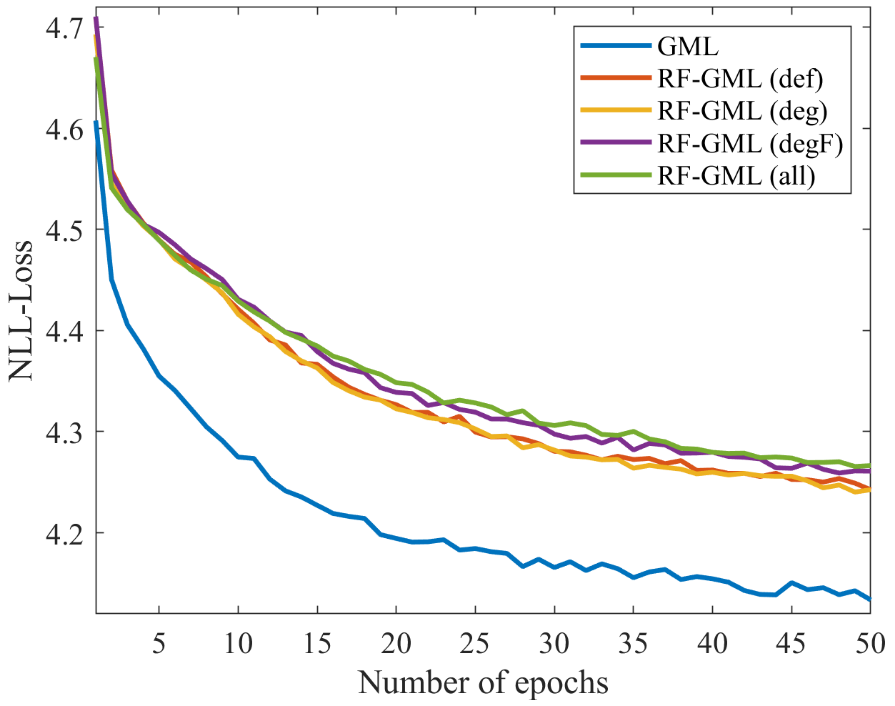


Mid: $\boldsymbol{M} = \frac{\boldsymbol{L+R}}{\boldsymbol{2}}$

Side: $\boldsymbol{S} = \frac{\boldsymbol{L-R}}{\boldsymbol{2}}$

Mean quality (μ)
additional parameter ($\log a$)
enables prediction of confidence interval (CI)

NLL training loss



RF-GML (def) Trained from scratch.

RF-GML (deg) First inception block (In-A) initialized from GML.

RF-GML (degF) Same as (deg), but weights frozen during training.

RF-GML (all) All inception blocks initialized from GML.

DATASETS

Training (80%) and validation (20%)

67,505 internal subjective scores

Codecs:

- ➢ AAC (stereo)
- ➢ HE-AAC v1/v2 (stereo)
- ➢ AC-4 (stereo)
- ➢ AC-4 A-JOC (binaural)
- ➢ DD+JOC (binaural)
- ➢ 3GPP IVAS (binaural)

Test

Mean scores from Unified Speech and Audio Coding (USAC) verification tests and two internal binaural tests.

Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Codecs	USAC, HE-AAC v1/v2, AMR-WB+			DD+JOC, AC-4 IMS	
Bitrates [kb/s]	8–24	16-24	32-96	256-448	64-256
#Conditions	12	10	11	5	5
#Excerpts	24	24	24	11	12
#Subjects	66	44	28	9	11

No-reference subjective tests were not used due to lack of reliable datasets for audio codecs.

RESULTS

Evaluation metrics

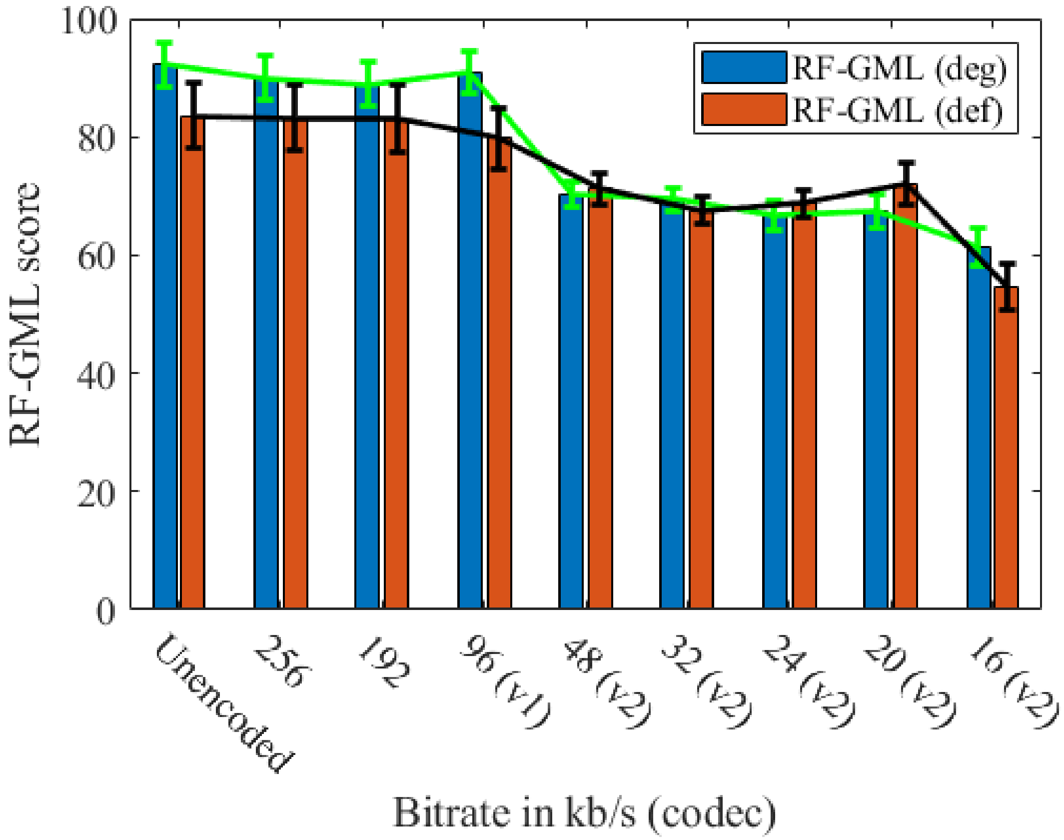
- Pearson linear correlation **R_p**
- Spearman rank correlation **R_s**
- Mean quality of unencoded audio **MU**

Benchmarking

Models \ Metric	Mono Bitrates			Stereo Low Bitrates			Stereo High Bitrates			Binaural Test-1			Binaural Test-2		
	R_p	R_s	MU	R_p	R_s	MU	R_p	R_s	MU	R_p	R_s	MU	R_p	R_s	MU
FR models															
ViSQOL-v3	0.81	0.84	94.64	0.77	0.78	94.64	0.82	0.90	94.64	0.90	0.93	94.64	0.96	0.85	94.64
GML	**0.88**	**0.88**	**100**	**0.89**	**0.86**	**100**	**0.92**	**0.94**	**100**	**0.98**	**0.95**	**100**	**0.98**	**0.92**	**100**
RF models															
SESQA	0.26	0.26	64.17	0.28	0.31	64.23	0.22	0.22	64.23	0.32	0.33	54.06	0.14	0.17	45.36
RF-GML (def)	**0.82**	**0.83**	82.14	0.78	**0.77**	84.54	0.78	0.65	84.54	**0.86**	0.68	82.43	**0.97**	**0.81**	96.25
RF-GML (deg)	0.78	0.76	**89.56**	**0.81**	0.75	**93.78**	**0.86**	**0.81**	**93.78**	0.79	0.73	**88.82**	0.90	0.76	**99.38**
RF-GML (degF)	0.76	0.75	78.79	0.79	0.75	81.57	0.85	0.79	81.57	0.84	0.71	79.67	0.95	0.79	84.35
RF-GML (all)	0.73	0.72	71.04	0.76	0.71	71.38	0.81	0.70	71.38	**0.86**	**0.79**	70.58	0.93	0.76	68.21
Performance on speech															
SESQA	**0.82**	**0.83**	82.29	**0.78**	**0.80**	81.48	0.69	0.59	81.48	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
RF-GML (def)	0.79	0.78	90.95	0.71	0.63	92.18	0.79	0.71	92.18	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
RF-GML (deg)	0.75	0.74	**93.21**	0.73	0.65	**93.02**	**0.84**	**0.79**	**93.02**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

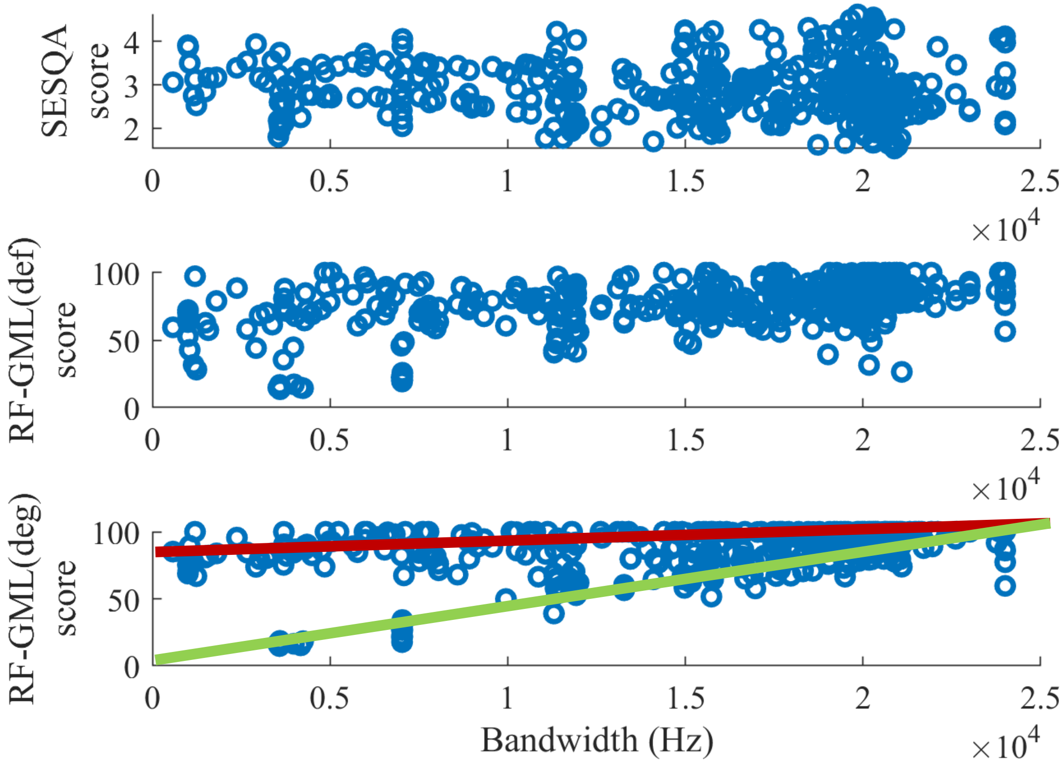
J. Serrà, J. Pons and S. Pascual, "**SESQA:** Semi-Supervised Learning for Speech Quality Assessment," *ICASSP,* 2021.

Rate-quality scaling



- RF-GML scores scale well with HE-AAC v1/v2 and AAC bitrates.
- RF-GML (deg) has better separation between bitrates than RF-GML (def).

Unencoded audio & bandwidth



Superior ability to rate unencoded audio closer to 100

CONCLUSION

Conclusion

Reference-Free Generative Machine Listener (RF-GML)

Transfer learning

Future research

Evaluating diverse audio content types and coding scenarios without an unencoded reference.

Transfer learning from a full-reference GML is highly effective.

Evaluate the CI prediction accuracy with RF subjective tests.

Explore new training methods.

THANK YOU

Arijit Biswas
arijit.biswas@dolby.com

Linkedin