Generative Machine Listener

—

GUANXIN JIANG, LARS VILLEMOES, AND ARIJIT BISWAS

155TH AES CONVENTION, NEW YORK, OCTOBER 26, 2023

Outline

Generative machine listener (GML)

Aims at simulating the MUSHRA scores s of an arbitrary number of listeners

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

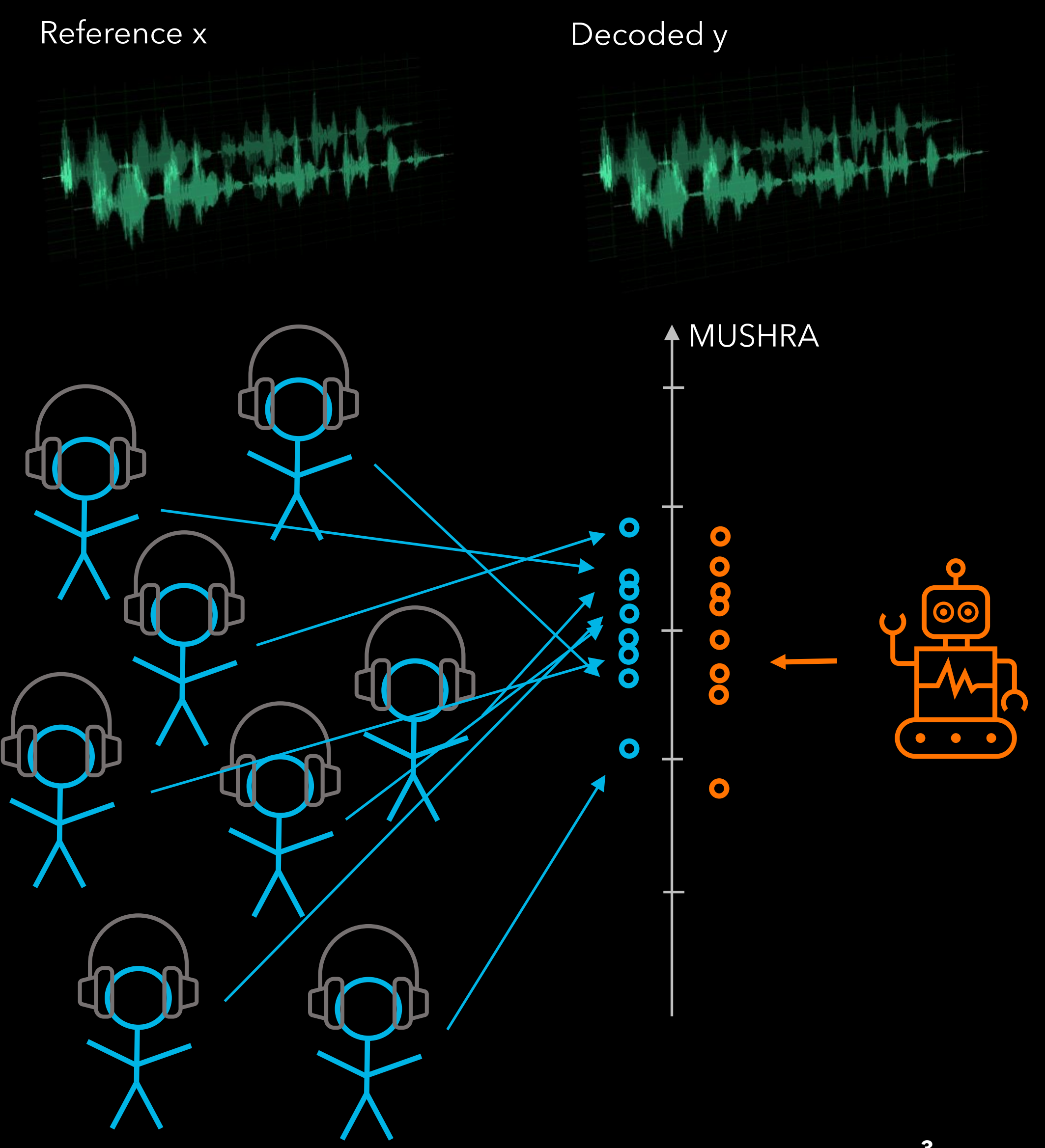
Pros

- can capture confidence intervals
- uses individual scores of dataset

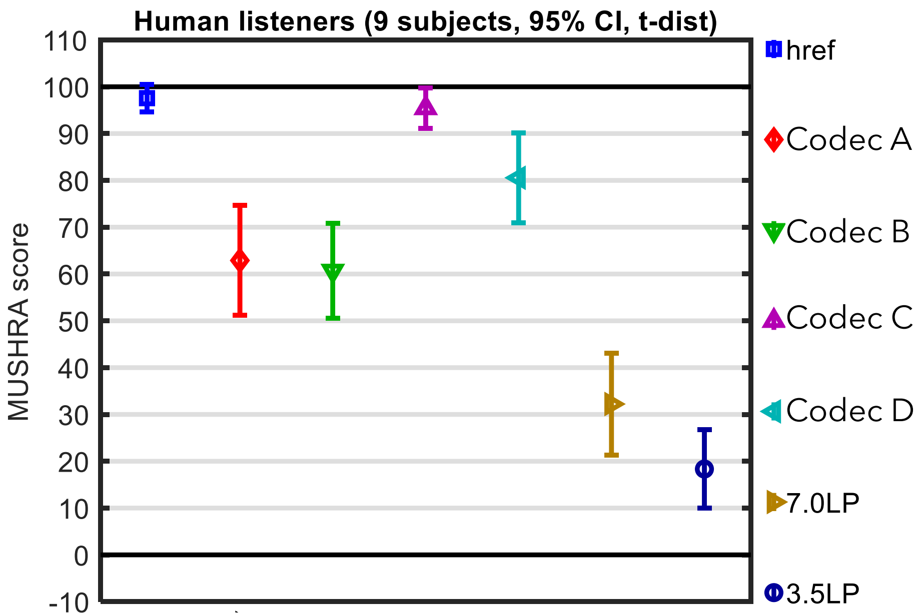
Con

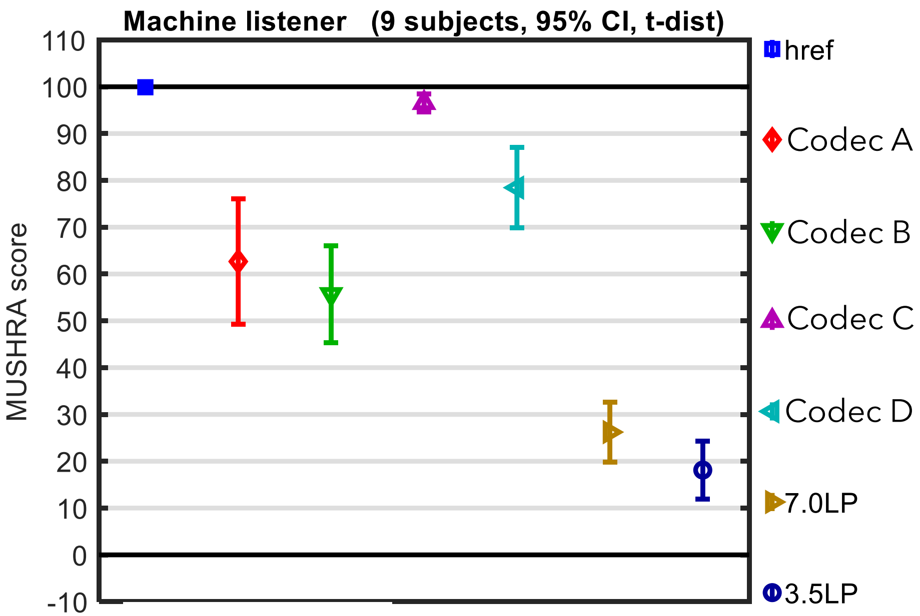
- more demanding to train than mean score regression

How good is the match of y to x?



Human vs machine listening





In this example, neither the excerpt nor the human listeners were seen during training.

DATASETS

Training (80%) and validation (20%)

67,505 internal subjective scores

Codecs: AAC, HE-AAC v1/v2, Dolby AC-4, A-JOC, DD+JOC, 3GPP IVAS

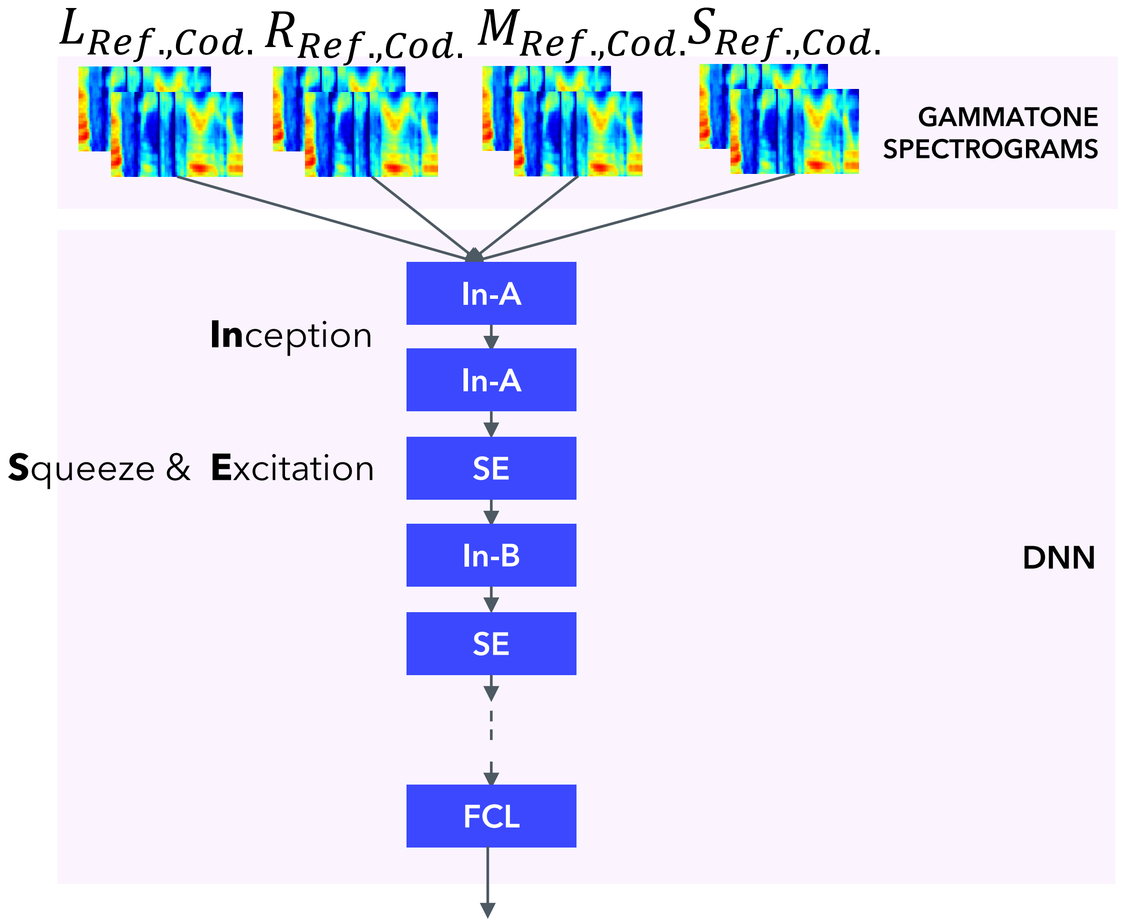
Test

Means and confidence intervals (CI) 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, 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

MODEL

Stereo InSE-NET



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

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

Only architectural change ➡ [**Mean MUSHRA** additional parameter]

"Stereo InSE-NET: Stereo Audio Quality Predictor Transfer Learned from Mono InSE-NET," A. Biswas, and G. Jiang, Paper 21, (AES October 2022)

Output stage and loss

Given signals x and y, the model outputs two parameters controlling the conditional pdf

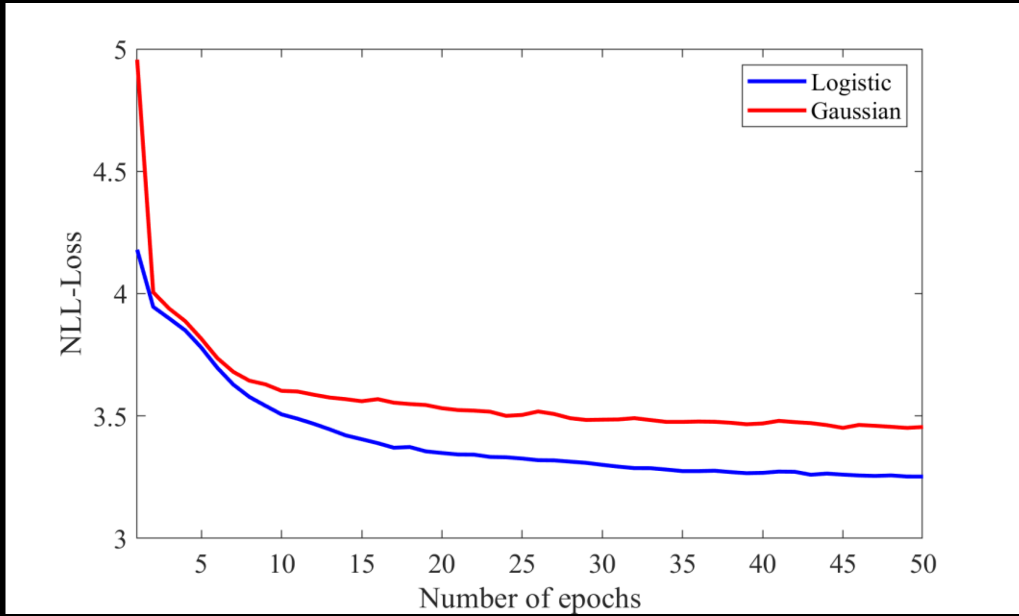
$$p(s|x,y)$$

The loss is the nonnegative log likelihood (NLL)

$$-\log p(s|x,y)$$

Given validation loss performance, we use the **logistic** model

	pdf	loss
Gaussian $(\mu, \log\sigma)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}}$	$\log\sqrt{2\pi}\sigma + \frac{(s-\mu)^2}{2\sigma^2}$
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

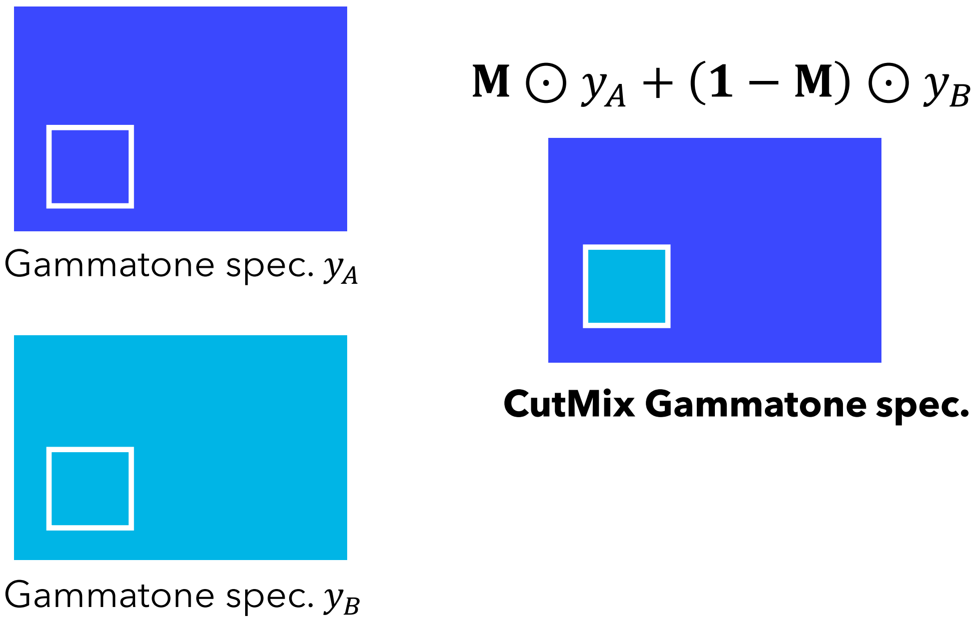
CutMix

1. Sample $\lambda \sim \boldsymbol{B}(\alpha, \alpha)$

   pdf: $c \cdot (t(1-t))^{\alpha-1}, \; 0 < t < 1$

2. Draw a randomly positioned gammatone spectrogram patch of normalized area λ
3. Cut out the patch from one spectrogram and insert it in the other
4. Interpolate the two subjective scores

(We use $\alpha = 0.7$)



CutMix **score** → $\lambda s_A + (1-\lambda) s_B$

RESULTS

Evaluation metrics

For mean MUSHRA scores

- Pearson linear correlation **R_p**
- Spearman rank correlation **R_s**
- Outlier ratio **OR**: proportion of scores outside of subjective confidence interval

For confidence intervals

- Pearson linear correlation **R_p**
- Spearman rank correlation **R_s**
- Root mean squared error **RMSE**

Mean MUSHRA scores

Pearson linear correlation **R_p**↑

Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Model					
VisQOL-v3	0.81	0.77	0.82	0.90	0.96
Non-GML	0.87	0.87	**0.93**	**0.98**	0.98
GML	0.84	0.82	0.90	0.96	**0.99**
GML + CutMix	**0.88**	**0.89**	0.92	**0.98**	0.98
Non-GML+CutMix	0.87	0.87	0.90	**0.98**	**0.99**

Mean MUSHRA scores

Spearman rank correlation $\mathbf{R_s}\uparrow$

Test	**Mono**	**Stereo low bitrates**	**Stereo high bitrates**	**Binaural 1**	**Binaural 2**
Model					
VisQOL-v3	0.84	0.78	0.82	0.93	0.85
Non-GML	0.82	0.83	0.93	**0.96**	0.89
GML	0.80	0.75	0.90	0.94	**0.95**
GML + CutMix	**0.88**	**0.86**	**0.94**	0.95	0.92
Non-GML+CutMix	0.83	0.80	0.89	0.95	**0.95**

GML + CutMix is advantageous
(Separate usage of GML or CutMix is not)

Mean MUSHRA scores

Outlier ratio **OR** ↓

Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Model					
VisQOL-v3	N/A	N/A	N/A	N/A	N/A
Non-GML	0.92	0.82	0.78	0.27	0.77
GML	**0.75**	**0.63**	0.62	0.34	**0.42**
GML + CutMix	0.80	0.70	**0.56**	**0.19**	0.56
Non-GML+CutMix	0.87	0.80	0.78	0.23	0.51

GML is advantageous

Confidence intervals

Pearson linear correlation **R_p** ↑

Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Model					
GML	0.36	0.31	0.38	0.37	0.21
GML + CutMix	**0.79**	**0.80**	**0.78**	**0.70**	**0.76**

Confidence intervals

Spearman rank correlation $\mathbf{R_s}\uparrow$

Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Model					
GML	0.28	0.23	0.34	0.38	0.28
GML + CutMix	**0.44**	**0.43**	**0.67**	**0.65**	**0.60**

Confidence intervals

Root mean squared error **RMSE**↓

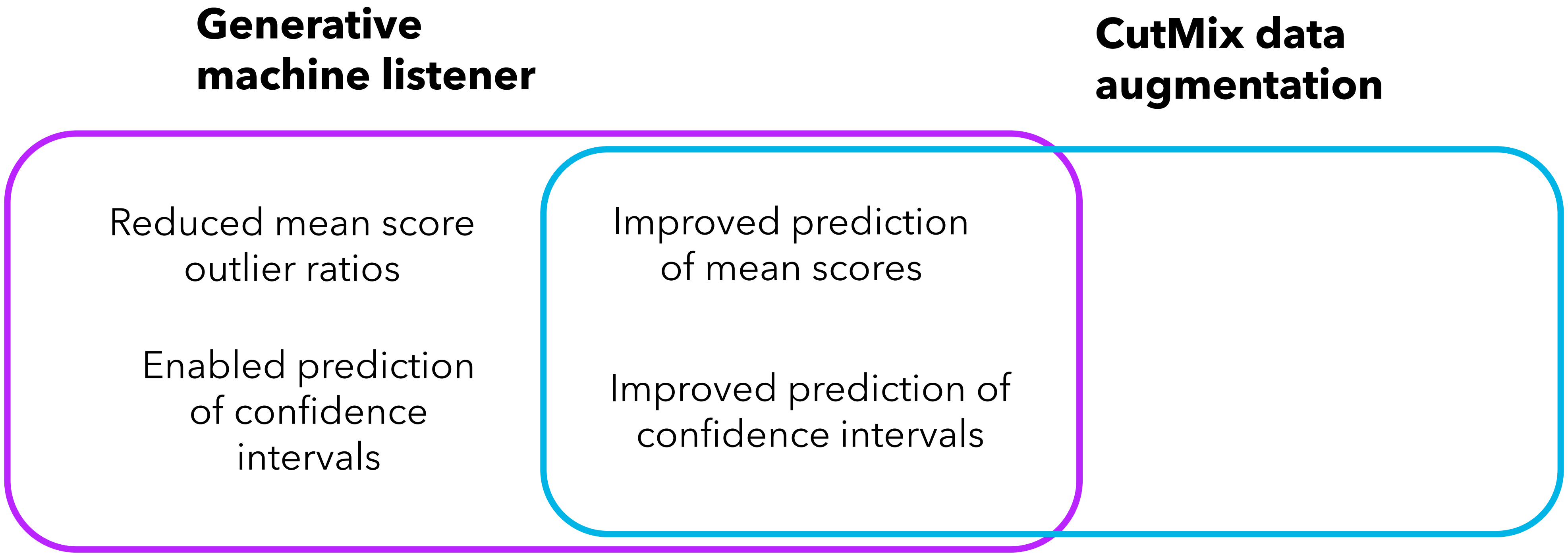
Test	Mono	Stereo low bitrates	Stereo high bitrates	Binaural 1	Binaural 2
Model					
GML	2.80	3.82	4.44	7.61	4.46
GML + CutMix	**0.87**	**1.13**	**1.50**	**3.20**	**2.25**

CutMix is advantageous in all three metrics

CONCLUSION

Conclusion

Relative to the mean score regression model (Non-GML), we observe the benefits



—

THANK YOU