NORESQA: A framework for Speech Quality Assessment using Non-Matching References

Pranay Manocha[1], Buye Xu[2], Anurag Kumar[2]

[1] Princeton University, [2] Facebook Reality Labs Research

Neural Information Processing Systems (NeurIPS) 2021

Speech Quality Assessment (SQA)

Task

- Accurate and reliable assessment of speech quality
- Useful for telephony, VoIP, Hearing Aids etc.

Gold Standard

Not scalable;
Costly and Time consuming
(repeated many times per recording

Speech Quality Assessment (SQA)

Objective Metrics

Reference Signal

Models

PESQ [Rix ‘01],
VISQOL [Hines ‘15],
HASQI [Kates ‘14]

Complex hand-crafted;
Sensitive to perceptual transformations;
Need a matching clean reference;
Non-differentiable

Speech Quality Assessment (SQA)

ML based Objective Metrics

Dataset → Deep NN   Metrics

Objective scores
(e.g., *PESQ*)

Subjective ratings
(e.g., *MOS*, *JND*)

Full-reference metrics

DPAM and CDPAM *[Manocha '20 and '21]*

Reference
Signal

Deep NN

Ratings
(e.g., PESQ)

Correlate well with perception; differentiable *but:*
Always require a paired clean signal for reference

Speech Quality Assessment (SQA)

ML based Objective Metrics

No-reference metrics

Quality-Net [Fu ‘18], DNSMOS *[Reddy ’20]*

Reference-free but:

Generalize poorly to unseen perturbations

Collecting MOS dataset is difficult

- Consistency in listening environments, equipment etc.

- Large variance (noisy labels) in MOS ratings

[Formuation]

Generalization problems due to lack of a reference

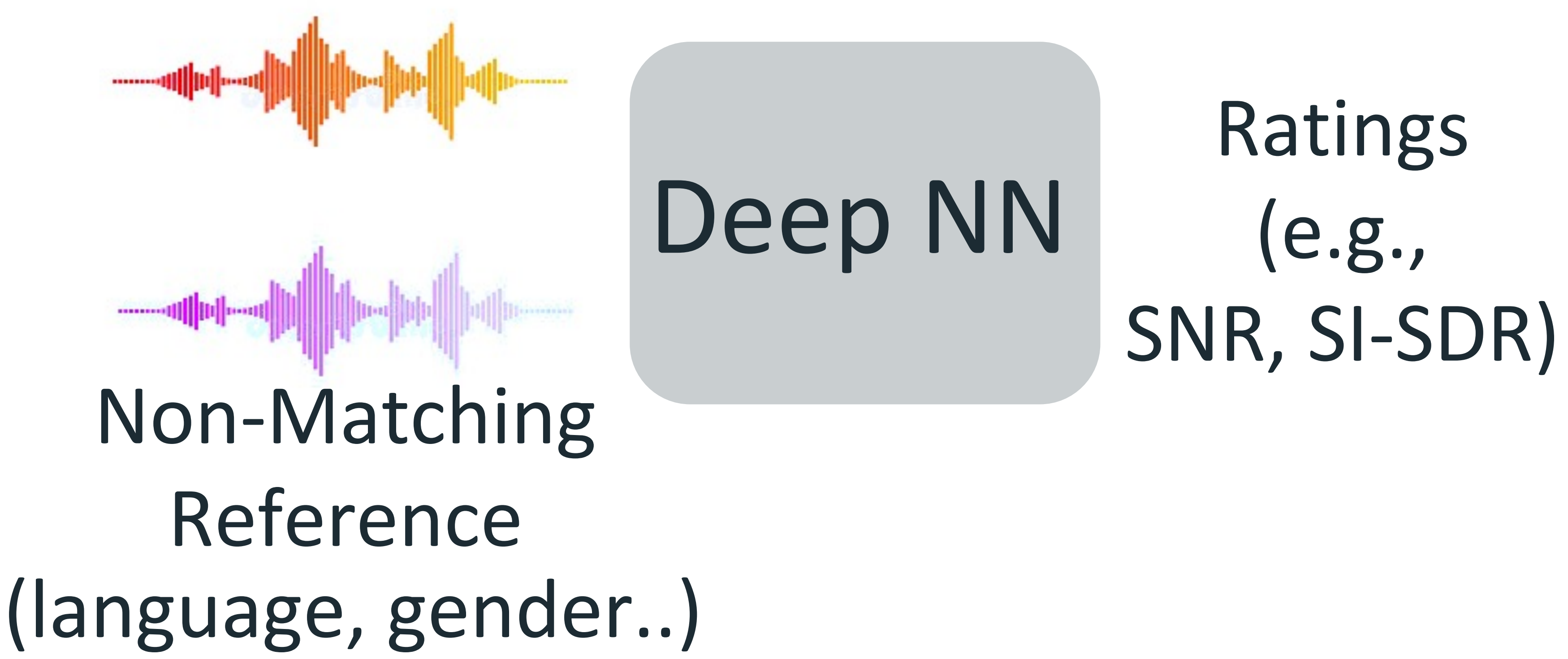
- Varied, experience /- mood dependent

Speech Quality Assessment (SQA)

Features

- Usable in real world where no references exist.
- Addresses the problem of lack of a reference
- Does not require any labeled dataset (low variance)

- SQA using non-matching references (NMRs)
- Inspired by human behavior: can compare quality across diff. speakers, languages etc.
- Relative assessments are easier than absolute ratings



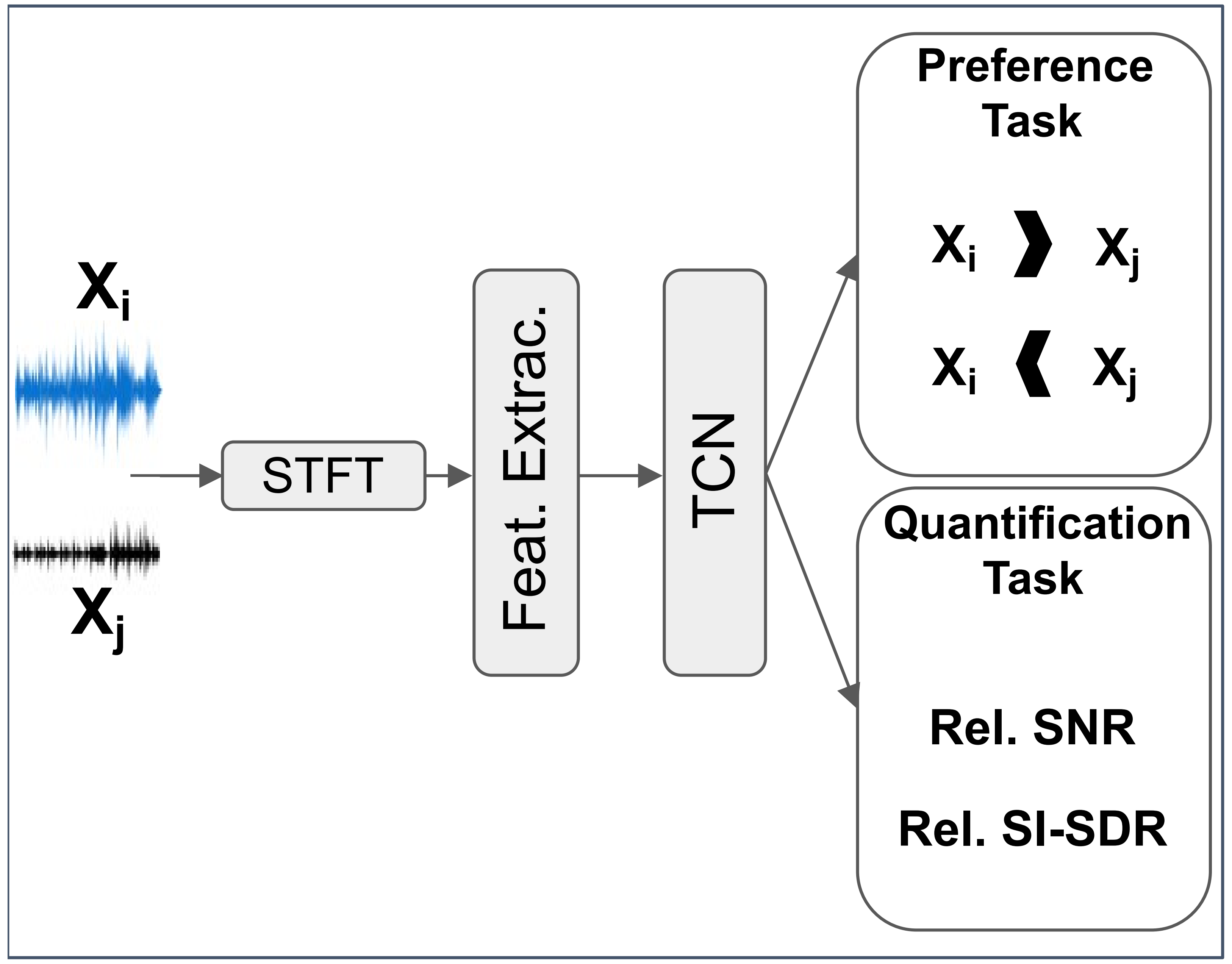
NORESQA

Broad Framework Overview

2 (non-matching) inputs

NORESQA processing pipeline

- Feature Extraction
- Temporal Aggregation
- Multi-task and multi-head learning head:

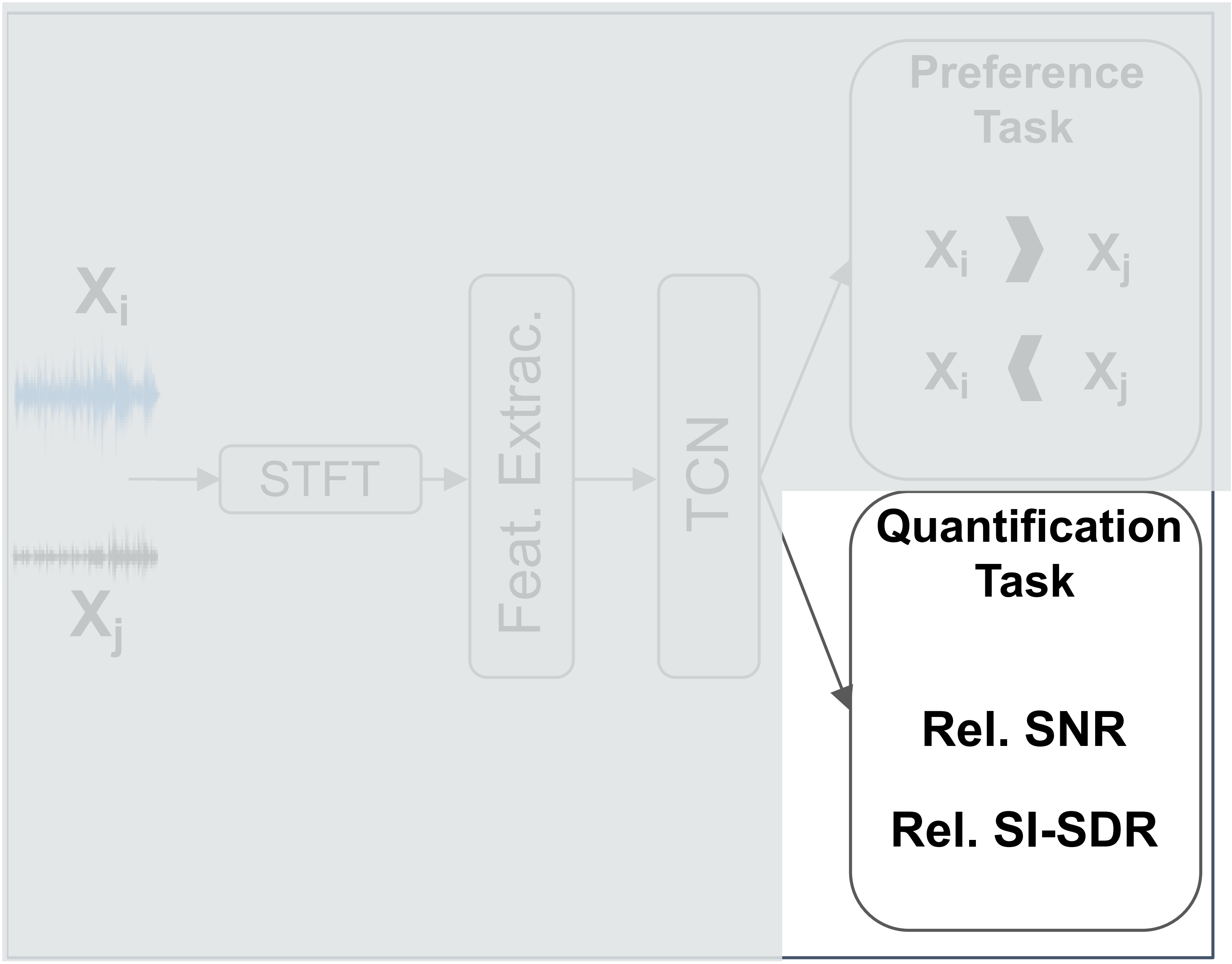


NORESQA Framework

Multi-objective learning

Relative SNR and SI-SDR prediction:

- No labeled data; Most fundamental measures
- Desirable Properties (distn. metric; scale invariance)
- Works across realistic tasks



NORESQA Framework

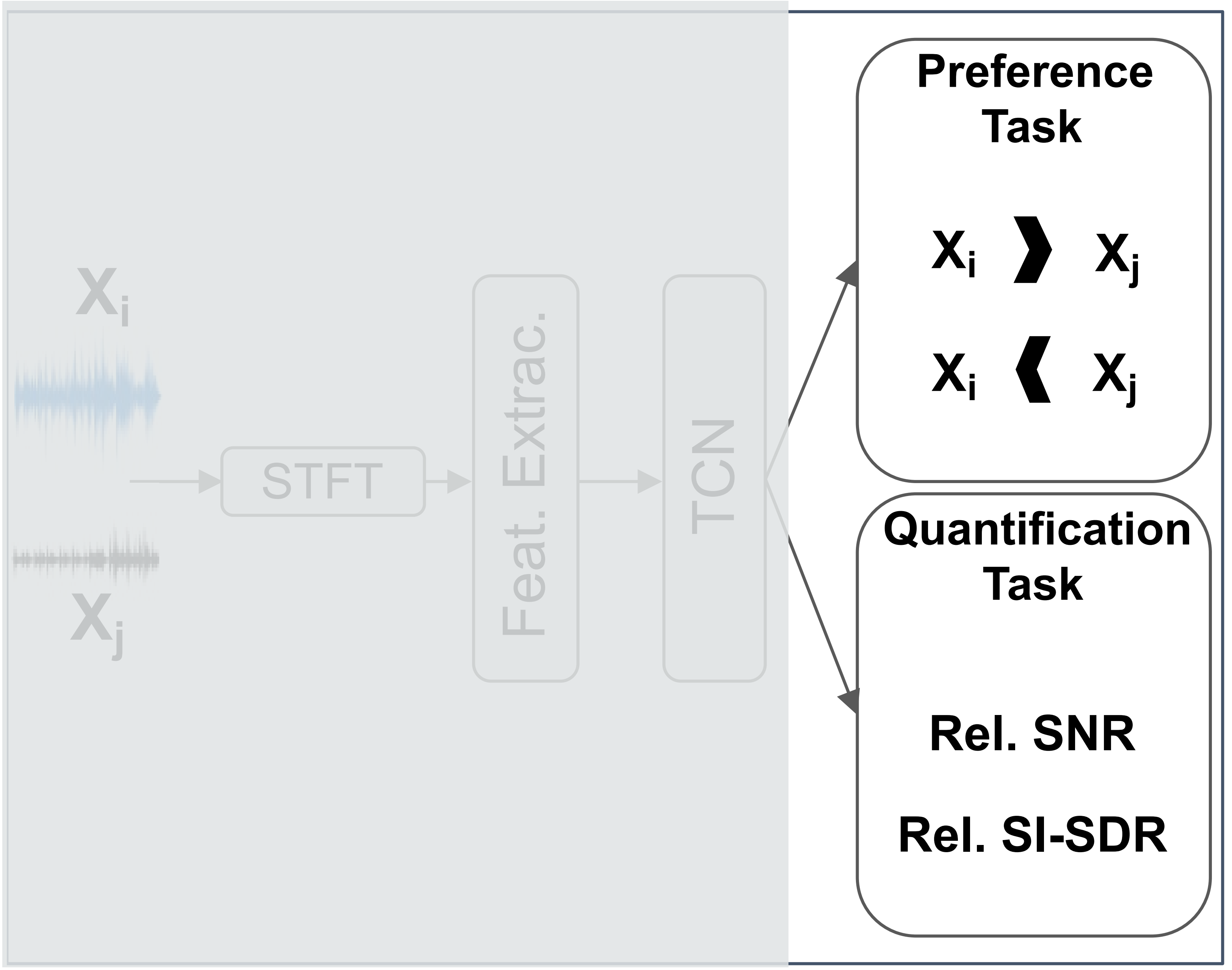
Multi-task learning

Preference task - which input audio is of better quality

Quantification task - quality difference between the two audio inputs

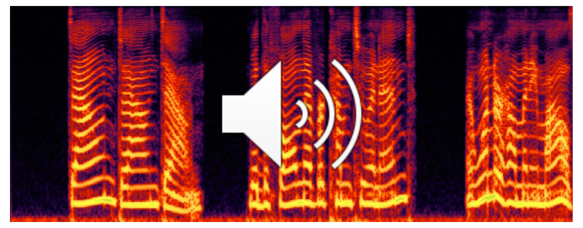
Two tasks important because:

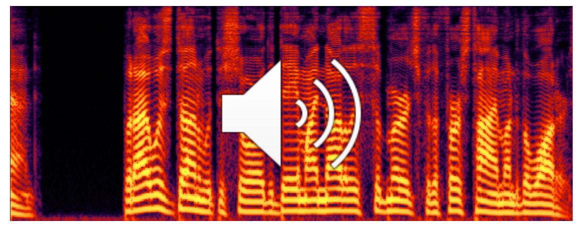
- Focus on quality attributes
- Easier to use - adjust individual model
- Easy extension to > 2 inputs



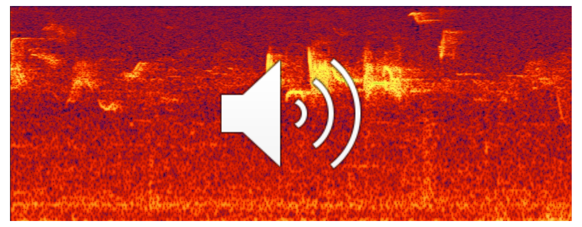
Training Procedure

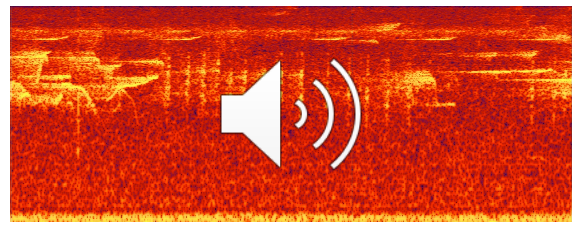
1. Clean Recordings





2. Noise Recording



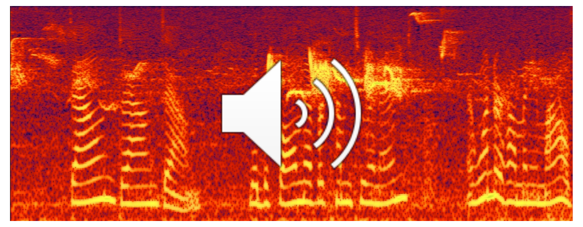


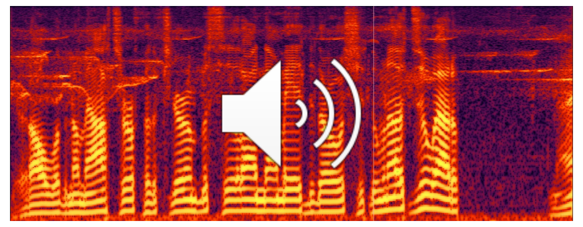
3. Noise levels

5dB

40dB

4. Final Recordings





Perturbations: Noise, EQ, Reverb…

NORESQA

5. Loss

Preference Task

[0,1]

- Binary Cross-Entropy loss (**L_P**)

Quantification Task

- Pose as classification
- Inter-class relationships
- *Gaussian* smoothed-labels

$L_Q = L_{SNR} + L_{SDR}$

Final loss ($L_P + L_Q$)

Usage

NORESQA Score:

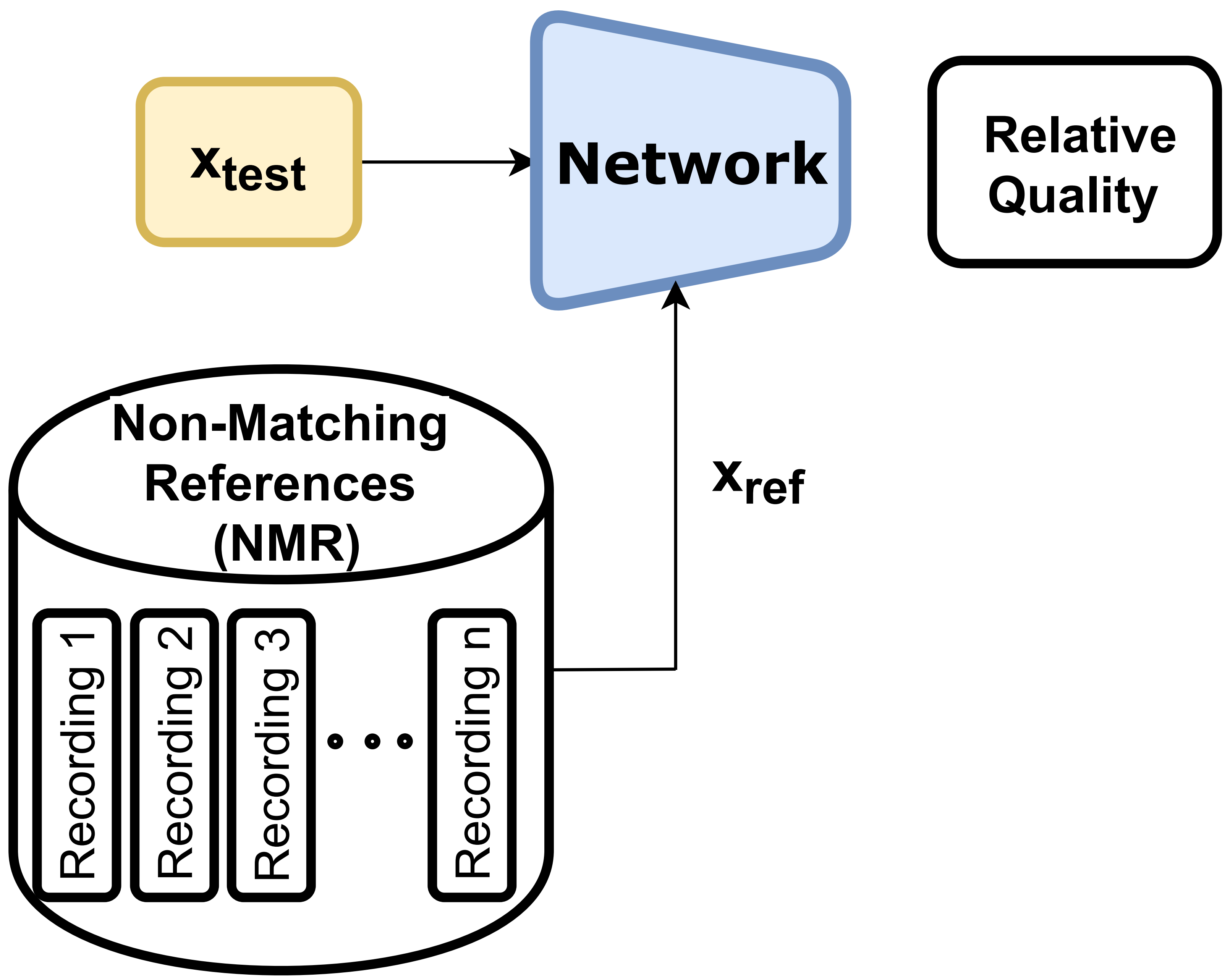
- *Preference* task shows '*sign*'
- *Quantification* task shows magnitude
- Aggregated over all *k* classes

$$\text{NORESQA}_{x_{test},x_{ref}} = \sum_{k=1}^{K} d^k_{x_{test},x_{ref}} \mu^k$$

Absolute Quality:

- *Averaging over a set of n non-matching references*

$$\text{NORESQA}^{avg}_{x_{test},x_{ref}} = \frac{1}{n} \sum_{i=1}^{n} \text{NORESQA}_{x_{test},x^i_{ref}}$$



Baselines

Full reference metrics:

- *PESQ*: hand-crafted, complex
- *CDPAM*: learned metric on *JND* ratings

No-reference metric:

- DNSMOS: learned metric on *MOS* ratings

Our proposed *NORESQA*:

- Entirely trained using simulated data

Results

1. Objective evaluation
2. Subjective Evaluation
3. Use as a ‘*differentiable*’ loss

Results: Objective evaluation

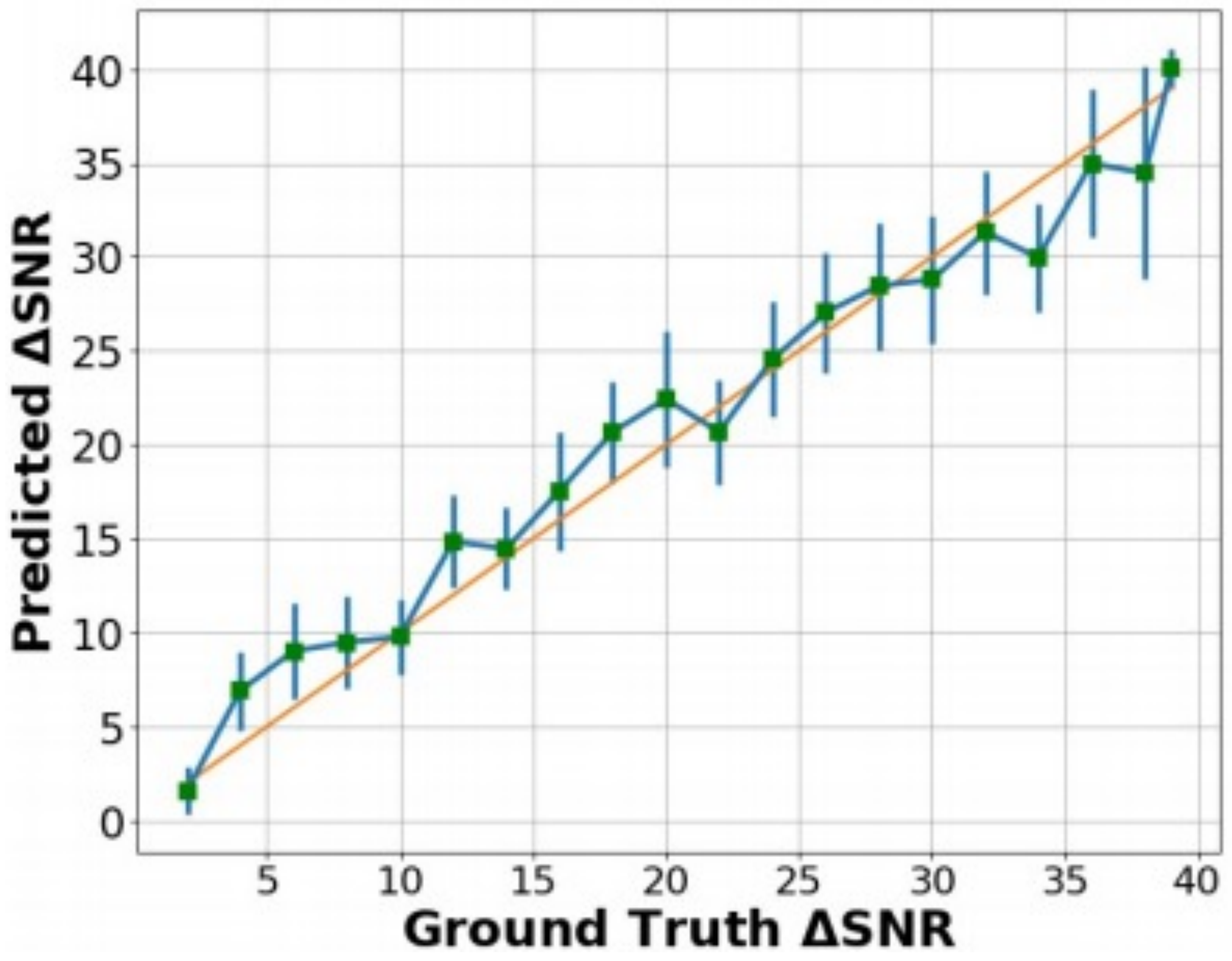
Invariance to language and gender;

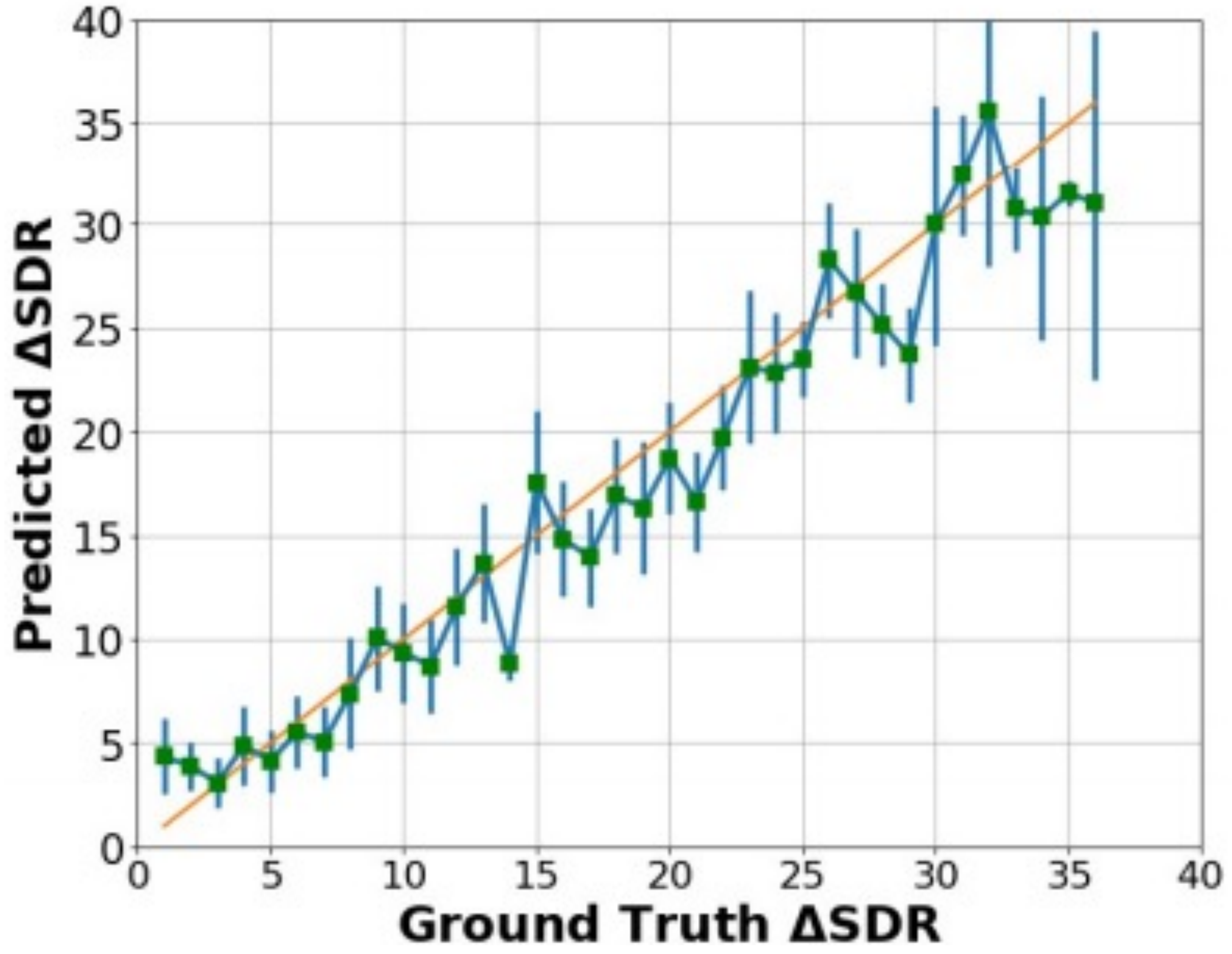
- Given $\mathbf{x}_{\mathbf{test}}$, doesn't matter the language or gender of NMRs.

Preference Task

97.3%

Quantification Task





Results: Subjective evaluation

Evaluation Datasets

- Synthesis tasks (*VoCo, FFTnet*)
- Speech Enhancement (*Dereverberation, Noizeus, HiFi-GAN*)
- Voice Conversion (*VCC-2018*)
- Speech Source Separation (*PEASS*)
- Telephony Degradations (*TCD-VoIP*)
- Bandwidth Expansion (*BWE*)
- General Degradations

Metrics

Correlate with *MOS* ratings using:

- Pearson correlation (PC)
- Spearman rank order correlation (SC)

Check 2AFC accuracy (*Triplet*) using:

- % accuracy

NORESQA

- Paired (*n=1*)
- Unpaired (*n=100*)
- Unpaired-Local-Fixed (*n=100*)
- Unpaired-Global-Fixed (*n=100*)

Results: Subjective evaluation

MOS correlations (n=100)

- *NORESQA:* competitive to full-reference methods and DNSMOS in all cases.

Type	Name	VoCo [65]		Dereverb [66]		HiFi-GAN [67]		FFTnet [68]	
		PC	SC	PC	SC	PC	SC	PC	SC
Full-ref.	PESQ	0.68	0.43	**0.86**	0.85	0.72	0.7	0.51	0.49
	CDPAM	-	**0.73**	-	**0.93**	-	0.68	-	**0.68**
Non-Int.	DNSMOS	0.6	0.48	0.7	0.73	**0.93**	**0.88**	**0.59**	0.53
NORESQA	Paired	0.64	0.6	0.46	0.65	0.59	0.81	0.46	0.47
	Unpaired	0.88±0.01	0.41±0.06	0.63±0.01	0.75 ±0.02	0.63±0.01	0.71±0.01	0.46±0.01	0.51±0.02
	+Local-Fixed	**0.89±0.01**	0.44±0.06	0.63±0.01	0.75±0.01	0.61±0.01	0.73±0.01	0.46±0.01	0.51±0.02
	+Global-Fixed	0.85±0.01	0.68±0.03	0.66±0.02	0.67±0.02	0.68±0.01	0.78±0.01	0.33±0.01	0.44±0.01

Type	Name	PEASS [69]		VCC-2018 [70]		Noizeus [71]		TCD-VoIP [72]	
		PC	SC	PC	SC	PC	SC	PC	SC
Full-ref.	PESQ	**0.86**	0.71	**0.51**	0.56	0.43	0.42	**0.89**	**0.90**
	CDPAM	-	**0.74**	-	**0.61**	-	**0.71**	-	0.88
Non-Int.	DNSMOS	0.39	0.21	0.37	0.42	0.41	0.59	0.71	0.72
NORESQA	Paired	0.26	0.43	0.48	0.39	0.47	0.46	0.38	0.44
	Unpaired	0.38±0.01	0.40±0.01	0.61±0.01	0.41±0.02	**0.50±0.02**	0.39±0.05	0.43±0.01	0.46±0.02
	+Local-Fixed	0.40±0.04	0.52±0.06	0.65±0.04	0.39±0.02	0.45±0.01	0.44±0.02	0.43±0.02	0.41±0.04
	+Global-Fixed	0.41±0.05	0.57±0.05	0.47±0.01	0.41±0.01	0.48±0.02	0.51±0.01	0.56±0.01	0.52±0.03

MOS Correlations; higher is better

Results: Subjective evaluation

2AFC accuracy

- *NORESQA* generalizes to other perceptual tests (like MOS and 2AFC) whereas DNSMOS works best only on MOS tasks.

Name	Simulated [6]	FFTnet [68]	BWE [73]	HiFi-GAN [67]
PESQ	86.0	67.0	38.0	88.5
CDPAM	**87.7**	**88.5**	**75.9**	**96.5**
DNSMOS	49.2	58.8	45.0	62.3
NORESQA	68.7	73.3	53.3	81.6

2AFC Accuracy; higher is better

Results: Ablations

Relative VS Absolute predictions:

- Predicting relative quality performs better than absolute rating
- Utility of providing a reference (even NMR) helps

Multi-objective learning (SNR and SI-SDR):

- Using either head performs worse than using both together

Number of NMRs (*n*):

- Increasing *n:*1 to 100 improves correlations by 15%.
- No significant diff. in unpaired local and global -> works for any random set of references.

Results: Speech Enhancement

- As a Pre-training strategy (large un-labeled corpus) + small labeled fine-tuning
- Consistently improves scores (esp. STOI)

Type	Data%	PESQ	STOI	SNRseg	CSIG	CBAK	COVL
Noisy		1.97	91.50	1.72	3.35	2.44	2.63
	33%	2.22	91.7	8.18	3.26	2.98	2.72
Baseline	66%	2.30	92.23	8.54	3.45	3.04	2.85
	100%	2.39	91.89	8.71	3.55	3.10	2.95
	33%	2.28	92.30	8.33	3.43	3.03	2.83
Pre-trained	66%	2.35	92.90	8.77	3.53	3.1	2.92
	100%	**2.46**	**93.53**	**8.81**	**3.59**	**3.17**	**2.99**

Speech enhancement; higher is better

Summary

1. Speech Quality assessments using non-matching references (NMRs)
2. Addresses a key limitation of no-reference metrics
3. Competitive against existing metrics, w/o any training on subjective ratings
4. *Differentiable* metric; good *pretraining* strategy for Speech Enhancement

Future Work

1. All new models under NORESQA framework that correlate better with human perception