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

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

[1] Princeton University

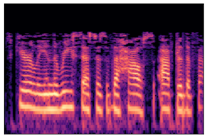
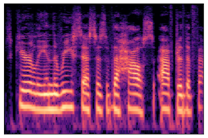
[2] Facebook Reality Labs Research

Neural Information Processing Systems (NeurIPS) 2021

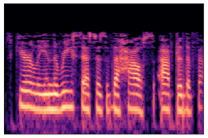
Motivation

Rate the audio quality of a recording:

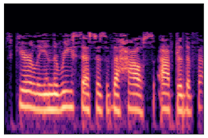
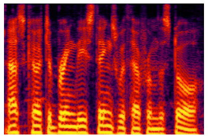
1. With its clean-`matching' reference

2. Without a reference



3. Non-matching reference!

Related works

Full reference metrics:

Traditional Metrics: PESQ *[Flessner '17]*, *VISQOL* *[Hines '15]*

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

Learned Metrics: DPAM and CDPAM *[Manocha '20 and '21]*

Learned from human annotated data; differentiable; need to exact same matching reference

Related works

No-reference metrics:

Traditional Metrics: ITU and SRMR *[Flessner '17]*, *VISQOL* *[Hines'15]*

Complex hand-crafted metric; Non-differentiable

Learned Metrics Quality-Net [Fu '18], DNSMOS *[Reddy '20]*

Learned from objective or MOS ratings; generalization to unseen perturbations; large variance (noisy labels) in MOS ratings -> challenge in training robust model

Related works

No-reference metrics:

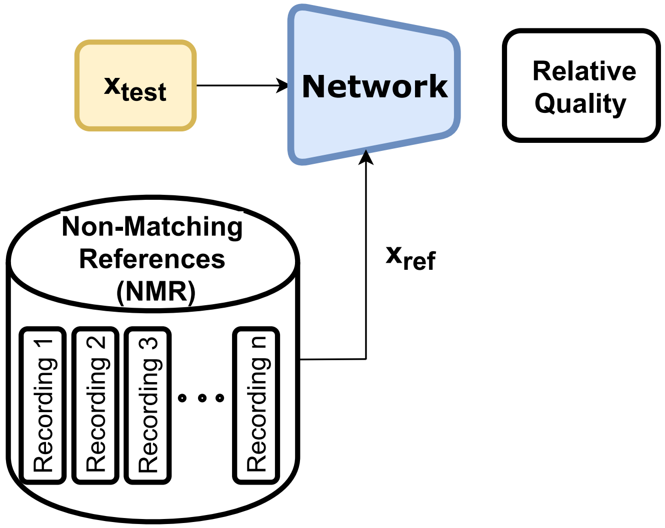
- **Challenge due to lack of a reference**

Learn the distribution of clean references that are used by human listeners.

1. Varied, d/o past experiences, mood …..
2. Difficult, especially when large label noise in ratings.

Our idea

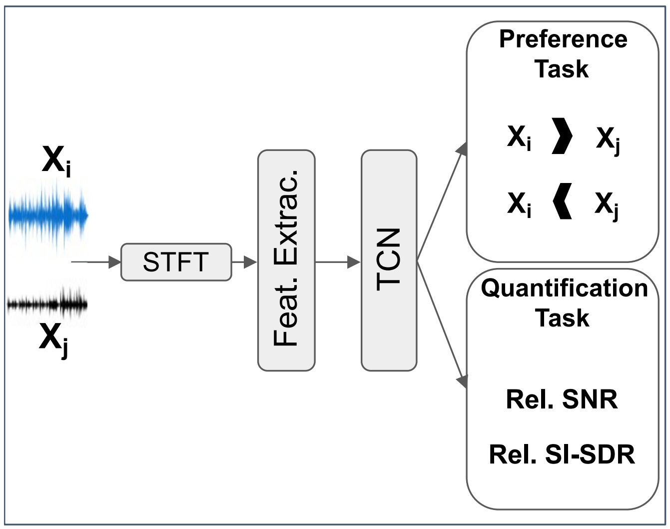
- SQA using non-matching references (NMRs) (of known quality)
- Relative assessments are easier than absolute ratings
- Inspired by human behavior: can also compare quality when diff. speakers, languages etc.



Features

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

Broad Framework Overview



2 inputs

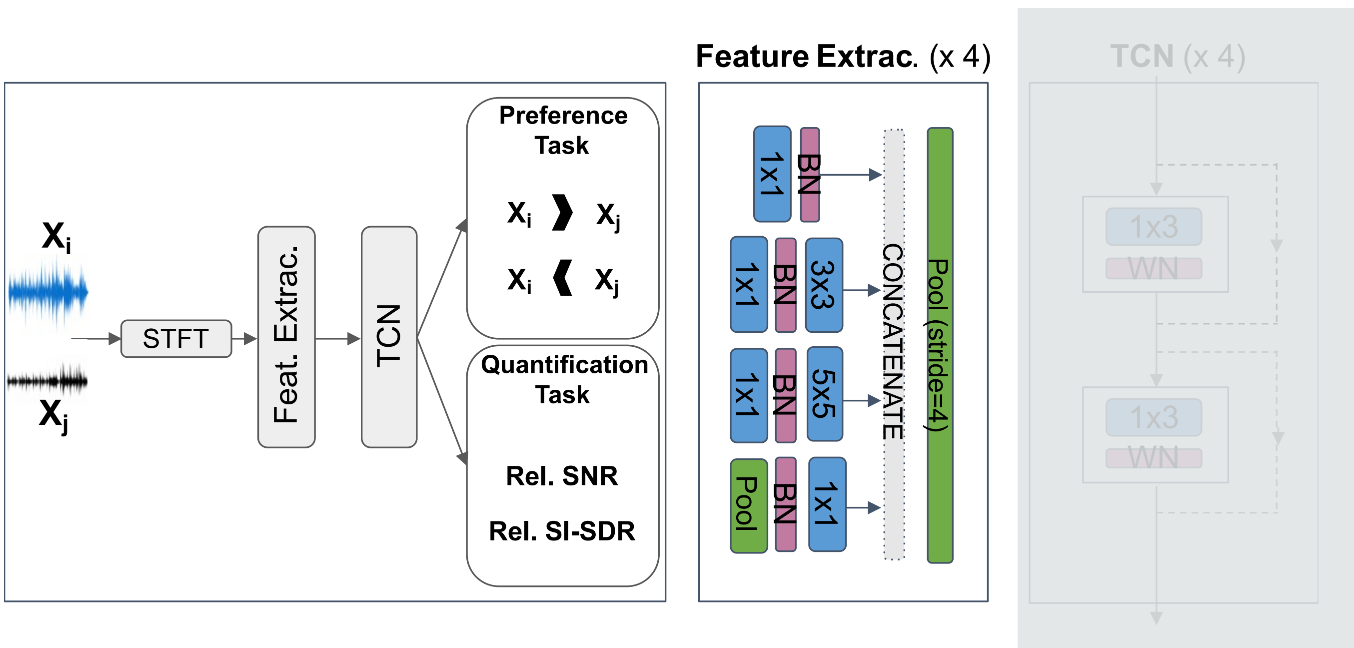
Processing pipeline

- Feature Extraction
- Temporal Aggregation
- Multi-task and multi-head learning head:
  - Preference and quantification task
  - Relative SNR and SI-SDR prediction

2 tasks; 2 objectives

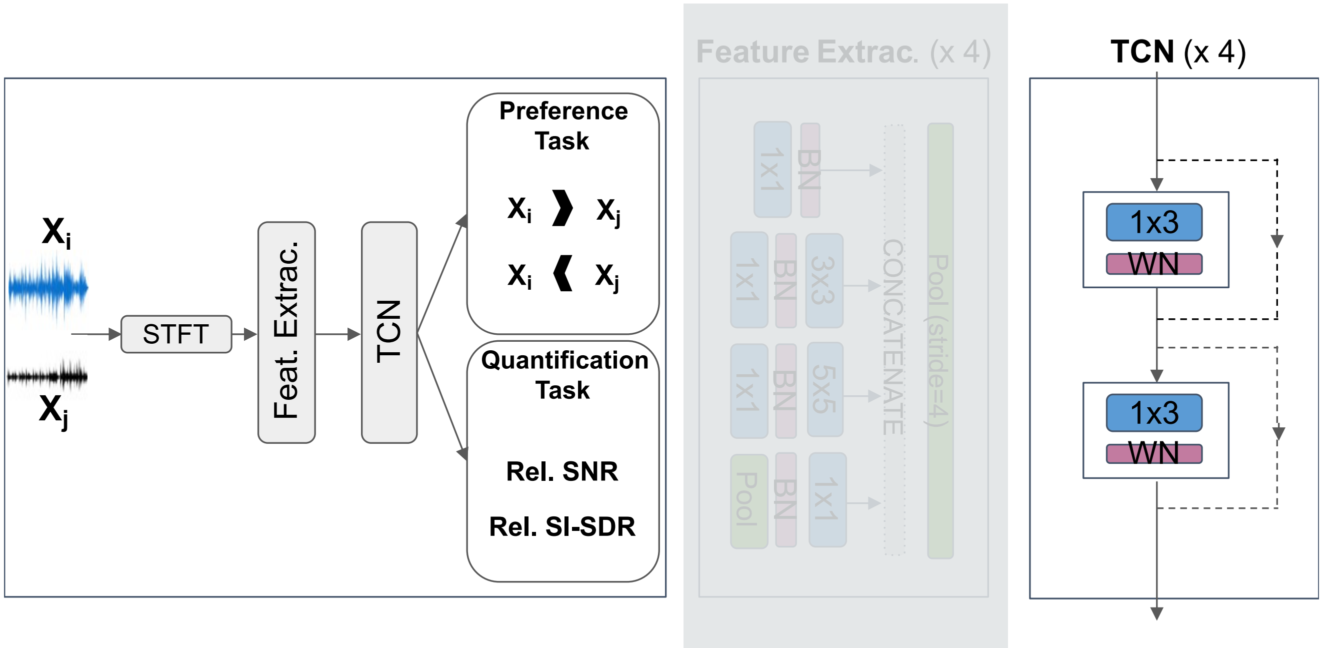
Architecture

Feature Extraction



Architecture

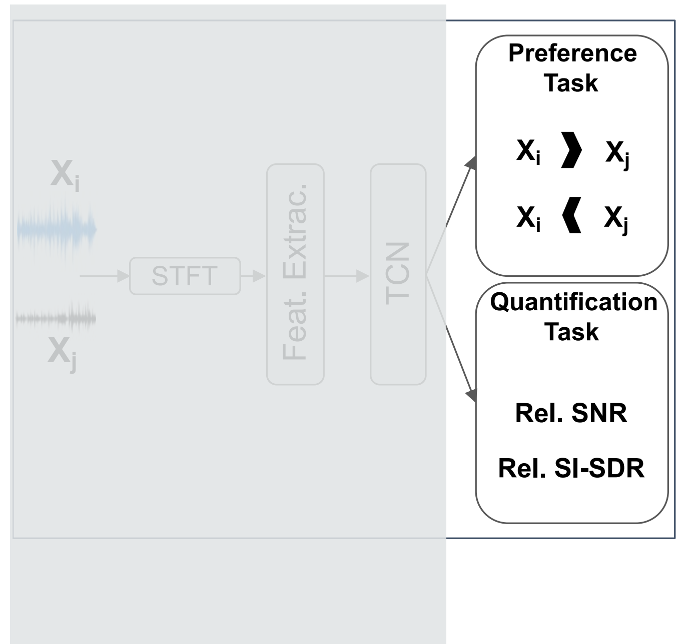
Temporal Aggregation



Architecture

Multi-task learning

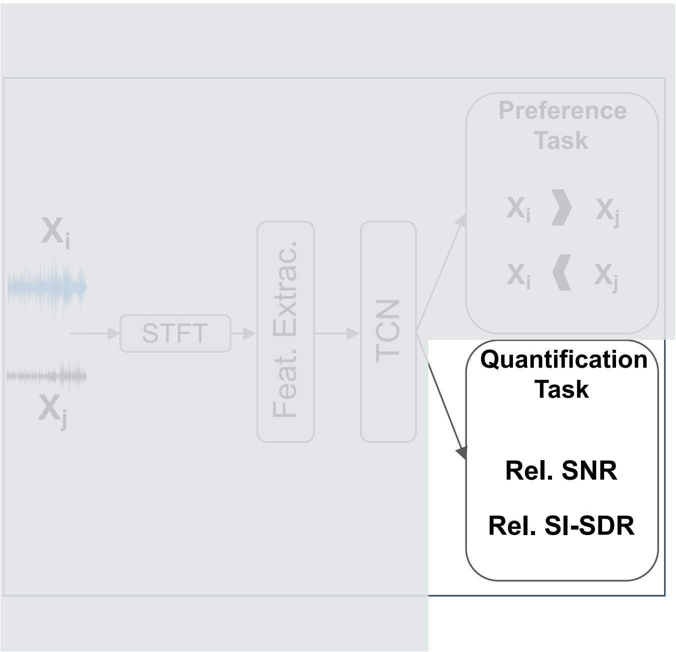
Preference and Quantification task



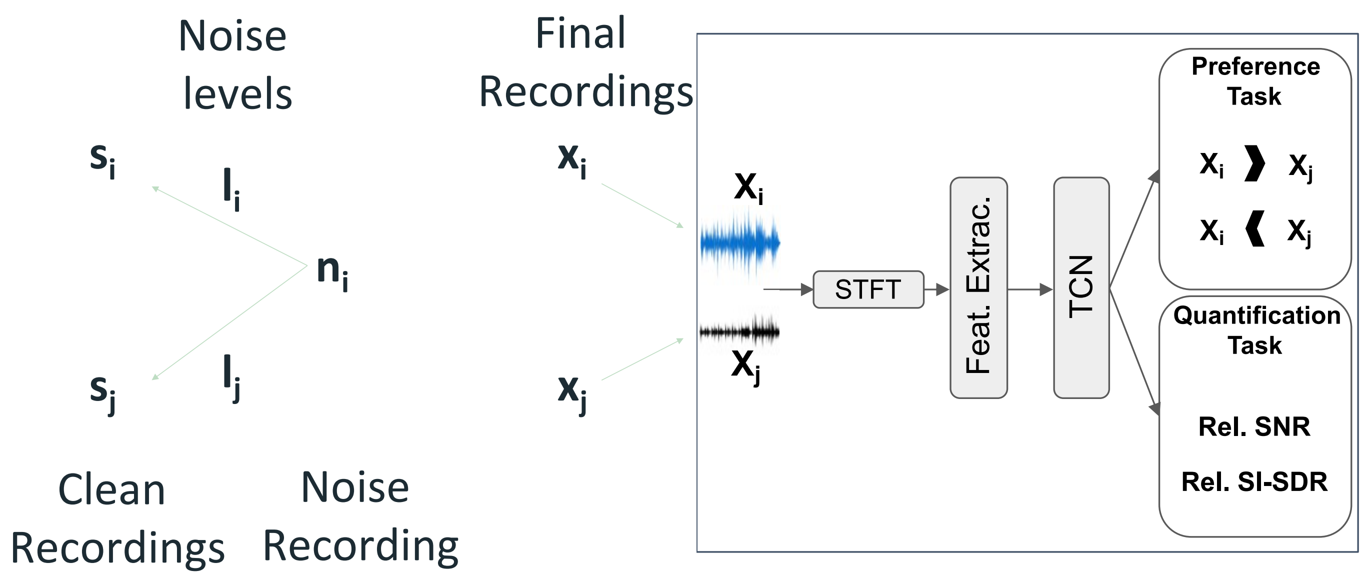
Architecture

Multi-objective learning

SNR and SI-SDR prediction



Training Procedure



Loss

Preference Task (L_P)

- Cross Entropy loss

Quantification Task (L_Q)

- Pose as classification, but take into consideration inter-class relationships
  - *Gaussian* smoothed labels (for both SNR and SI-SDR)
  - k-equal intervals,
  - $L_Q = L_{SNR} + L_{SDR}$

Final Loss

- $L_P + L_Q$

Usage

***NORESQA Score*:**

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

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

Absolute Quality:

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

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

Datasets

DNS Challenge

FSDK50

ESC-50

TIMIT

- Clipping
- Frequency Masking
- Reverberation
- Gaussian Noise
- Mu-law and MP3 compression

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

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

1. Performance on preference and quantification tasks
2. Invariance to language
3. Commutativity and indiscernibility of identicals
4. Quality based retrieval

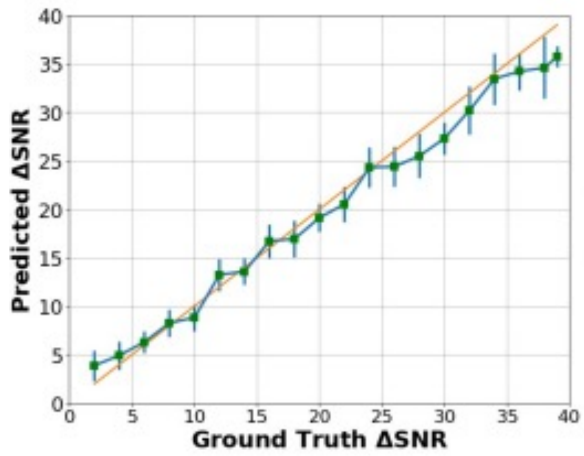
Results: Objective evaluation

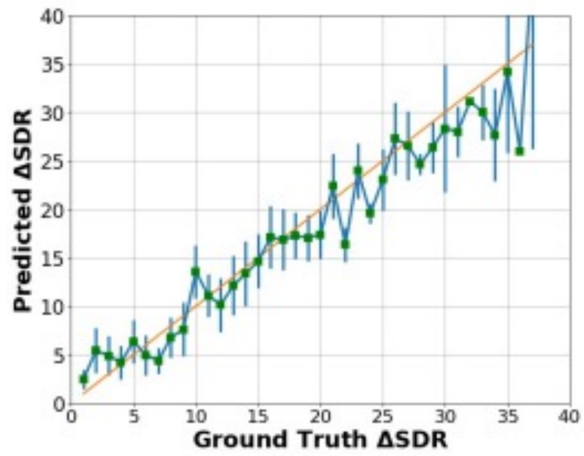
Performance on preference and quantification tasks

Preference Task

97.3%

Quantification Task





Results: Objective evaluation

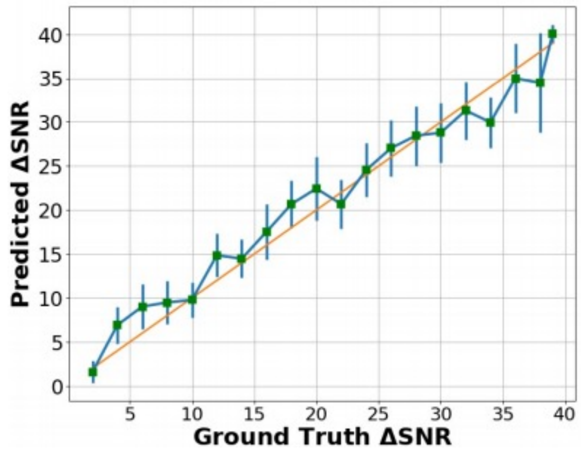
Invariance to language and gender

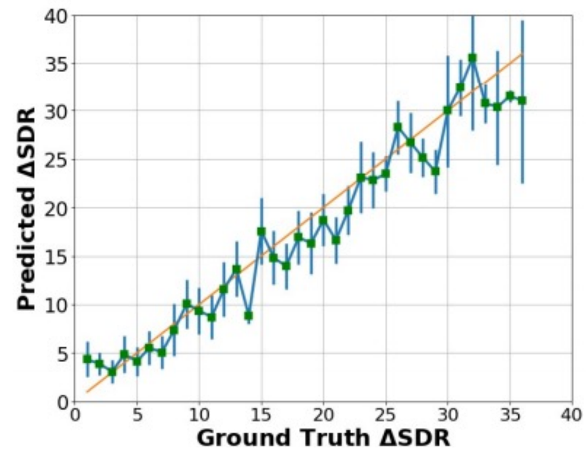
- Given $\mathbf{x}_{test}$, doesn't matter the language or gender of NMRs

Preference Task

97.3%

Quantification Task





Results: Objective evaluation

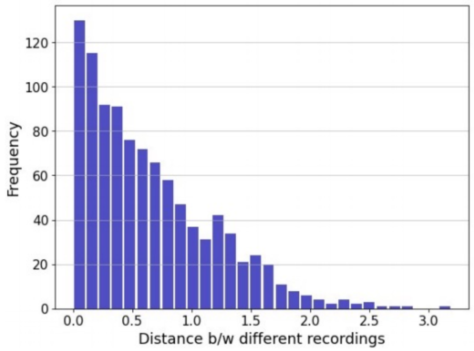
Commutativity

$$\mathcal{N}(x_{test}, x_{ref}) = \mathcal{N}(x_{ref}, x_{test})$$

Preference Task

97%

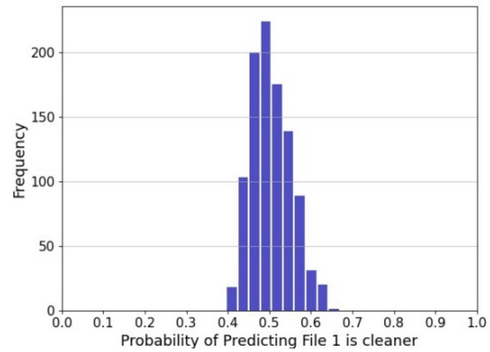
Quantification Task



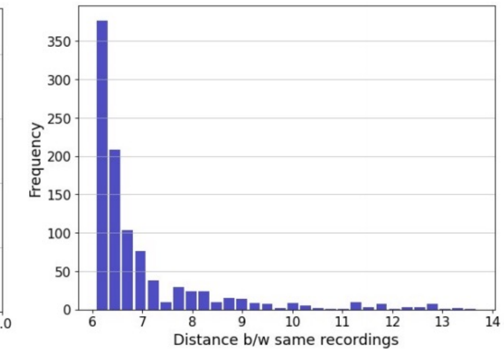
Indiscernibility of identicals

$$\mathcal{N}(x_{test}, x_{test})$$

Preference Task

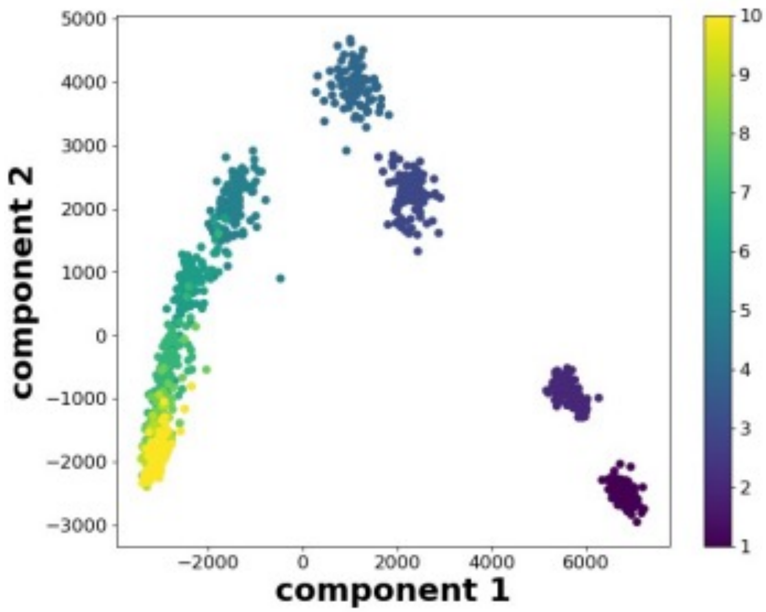


Quantification Task



Results: Objective evaluation

Quality based retrieval



PCA visualization of embeddings capturing audio quality information

Results: Subjective evaluation

MOS correlations (n=100)

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

Results: Subjective evaluation

2AFC accuracy

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

Results: Ablations

Relative VS Absolute predictions

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

Name	Type	VoCo		Dereverb		HiFi-GAN		FFTnet	
		PC	SC	PC	SC	PC	SC	PC	SC
Absolute	Sing. Inp.	0.32	0.31	0.19	0.17	0.19	0.30	0.16	0.15
	Two Inp.	0.41±0.15	0.35±0.03	0.26±0.08	0.27±0.01	0.42±0.07	0.45±0.06	0.17±0.01	0.09±0.01
NORESQA		**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**

Results: Ablations

Multi-objective learning (SNR and SI-SDR)

- Using either head performs worse than using both together

Type	Name	VoCo		Dereverb		HiFi-GAN		FFTnet	
		PC	SC	PC	SC	PC	SC	PC	SC
NORESQA	SNR only	0.43	0.39	0.39	0.38	0.49	0.42	0.2	0.1
	SI-SDR only	0.6	0.48	0.48	0.49	0.54	0.65	0.25	0.28
	SNR and SI-SDR	**0.85**	**0.68**	**0.66**	**0.67**	**0.68**	**0.78**	**0.33**	**0.44**

Results: Ablations

Number of NMRs (*n*):

- Increasing *n*->1 to 100 improves results by 15%.
- Averaging reduces the std in scores.
- No significant diff. in unpaired local and global -> equally well for any random set of references.

Type	Category	VoCo		Dereverb		HiFi-GAN		FFTnet	
		PC	SC	PC	SC	PC	SC	PC	SC
Unpaired	NMR_1	0.76±0.1	0.27±0.2	0.57±0.03	0.62±0.04	0.63±0.01	0.70±0.02	0.43±0.10	0.45±0.11
	NMR_{10}	0.87±0.01	0.43±0.07	0.64±0.01	0.73±0.03	0.63±0.01	0.70±0.01	0.45±0.03	0.48±0.06
	NMR_{100}	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	NMR_1	0.65±0.23	0.40±0.23	0.53±0.10	0.57±0.15	0.56±0.08	0.64±0.08	0.38±0.10	0.31±0.13
	NMR_{10}	0.79±0.1	0.44±0.2	0.61±0.05	0.69±0.05	0.61±0.02	0.67±0.03	**0.48±0.03**	0.50±0.04
	NMR_{100}	**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	NMR_1	0.79±0.20	0.54±0.20	0.44±0.16	0.41±0.19	0.56±0.08	0.63±0.10	0.29±0.10	0.36±0.12
	NMR_{10}	0.84±0.05	0.63±0.08	0.62±0.08	0.62±0.09	0.63±0.01	0.71±0.02	0.33±0.03	0.41±0.07
	NMR_{100}	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±0.02

Results: Speech Enhancement

As a Pretraining strategy: consistently improves scores

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**

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