PLCMOS - a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms

Lorenz Diener, Marju Purin, Sten Sootla, Ando Saabas,
Robert Aichner, Ross Cutler

https://aka.ms/plcmos
lorenzdiener@microsoft.com

Speech Quality Assessment

- Problem for every researcher working in speech – *how do we know our methods actually work?*
- **Gold standard: Human ratings**
  - Slow and expensive
  - Large number of raters required for reliable results
- **Classical objective metrics**
  - Correlation with actual human ratings not great
  - Sometimes require reference audio
  - Sometimes even cost money
- **Task-specific neural metrics**
  - Learn to replicate human ratings given audio
  - Data-driven – fewer implicit assumptions
  - Open models – broadly applicable
  - Cheap and fast to use – enables fast iteration
  - Validated for a specific task

➢ **PLCMOS** – metric for evaluating audio files after **packet loss concealment (PLC)**

PLC / PLCMOS task

- PLC: Given…
  - Audio file with cutouts
  - Location of the cutouts
- … restore the original audio data
- Different from audio inpainting task:
  - Real-time and low latency
  - Causal processing
  - Small models!
- PLCMOS: ***Given the output of models that solve the PLCMOS task, estimate human mean opinion score (MOS), and rank models based on this***

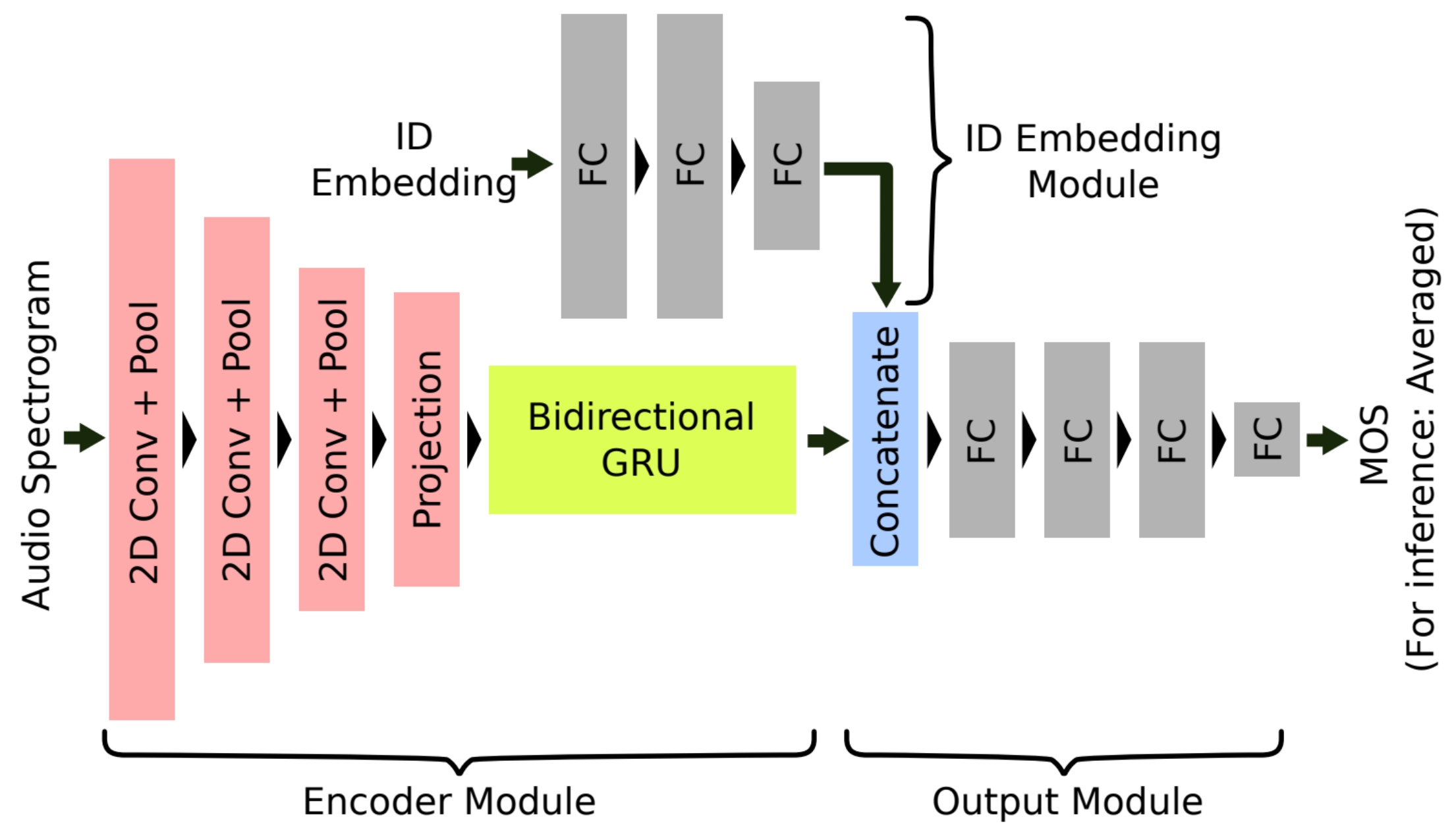
Dataset

1. Base audio data…
   - LibriSpeech (read speech)
   - LibriVox Podcast (conversational)
2. …combined with loss data…
   - Basic (realistic distribution)
   - Heavy packet loss
   - Long bursts (median >80ms)
3. …processed with models…
   - No-Op, Oracle
   - Codec PLC (Silk/Satin/Lyra)
   - Neural PLC models (Internal, INTERSPEECH 2022 PLC Challenge)
4. …labeled using P.808

		#Models		#Votes	
Audio data	Trace set	Train	Eval	Train	Eval
LibriSpeech	Basic	78	21	333740	22165
LibriSpeech	Long bursts	10	2	15550	990
Podcasts	Heavy loss	17		82110	
DNSMOS				16800	

➢ Realistic public domain based dataset with no audio from real calls: No privacy or copyright issues

PLCMOS Model Structure



Results

- Versus classical metrics (restricted to aligned reference available data)

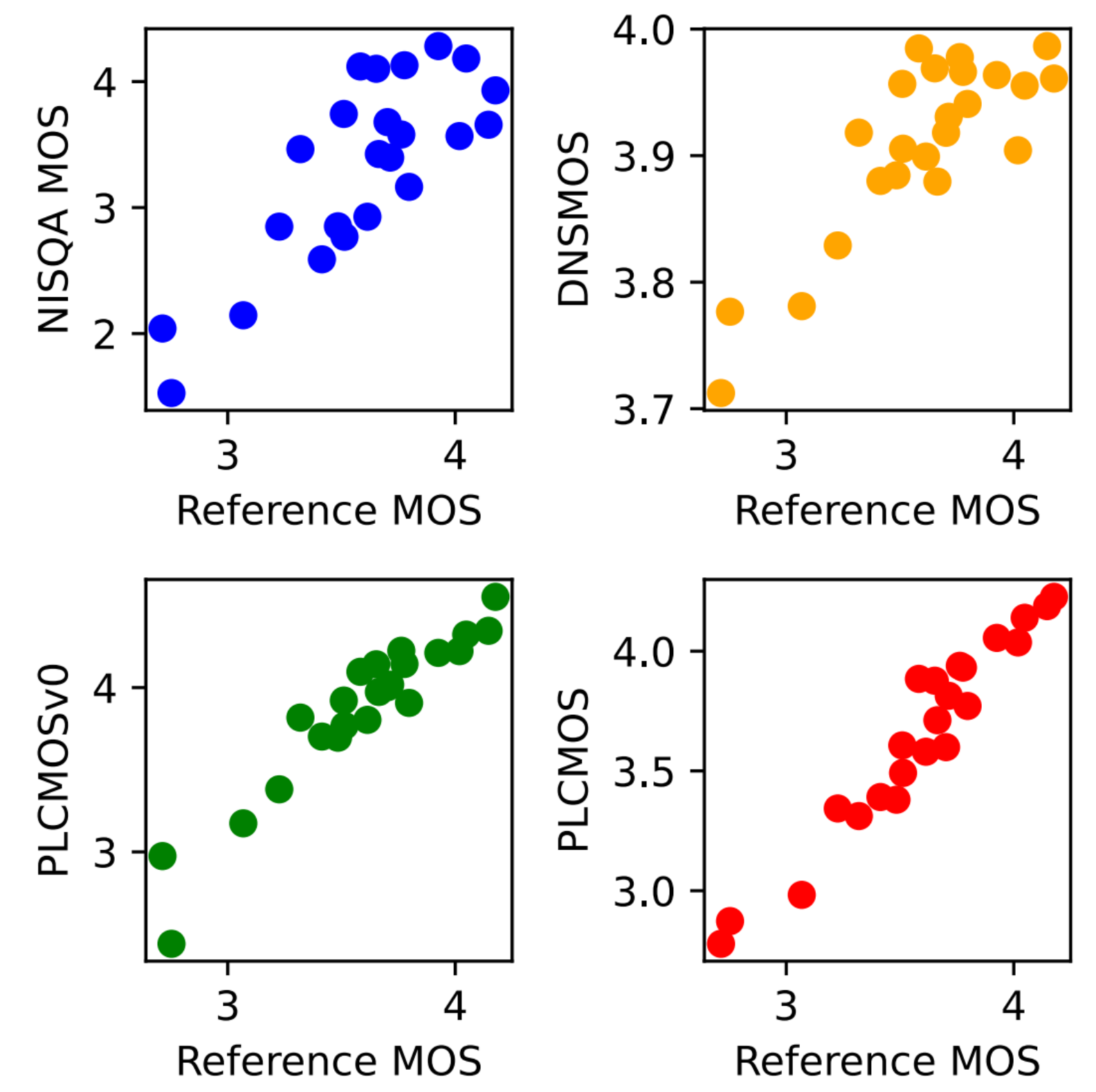
	Filewise		Modelwise	
Metric	PCC	SRCC	PCC	SRCC
MCD	0.14	0.21	0.23	0.06
PESQ	0.70	0.76	0.52	0.54
STOI	0.03	0.17	0.21	0.26
PLCMOS (ours)	**0.87**	**0.85**	**0.98**	**0.97**

- Versus other neural metrics

	Filewise			Modelwise		
Metric	PCC	SRCC	MAE	PCC	SRCC	MAE
DNSMOS	0.52	0.45	0.71	0.85	0.68	0.37
NISQA (MOS)	0.69	0.66	0.67	0.81	0.71	0.47
NISQA (DIS)	0.63	0.63	0.72	0.66	0.66	0.51
PLCMOSv0	0.81	0.79	0.48	0.94	0.92	0.29
PLCMOS (no ID)	0.83	0.80	0.45	0.95	0.95	0.20
PLCMOS (ours)	**0.85**	**0.83**	**0.40**	**0.97**	**0.95**	**0.09**

Discussion & Limitations

- PLCMOS beats classical metrics on PLC task evaluation, by a large margin
  - Dramatically better at ranking models
  - **More suitable for use during research and development than any other metric**
  - However: **Final evaluation should still be a human listening test!**
- PLCMOS beats other neural metrics on PLC task evaluation (including NISQA DIS)
  - Does not mean it is better for every task, just for the PLC task
- Potential limitations:
  - Sample rate (16kHz only)
  - Language (trained & validated on, mostly, modal English speech)



Links

Paper preprint (arXiv):
https://arxiv.org/abs/2305.15127

Speechmos package on PyPi (includes PLCMOS):
https://pypi.org/project/speechmos/

2022 INTERSPEECH PLC Challenge
https://aka.ms/plc_challenge

Conclusions and Future Work

- **PLCMOS provides good estimate of human MOS ratings for the PLC task (SRCC ~0.95 for models) without requiring a reference!**
- **Model available freely to anyone (ONNX format + PyPi package for ease of use) – easy to integrate into your evaluation pipeline!**
- **Future work: Extend to more diverse data (models, languages, sample rates) and validate more use cases**

Microsoft Teams
IC3
INTERSPEECH 2023