Dolby

Stereo InSE-NET

Stereo audio quality predictor transfer learned from mono InSE-NET

ARIJIT BISWAS

GUANXIN JIANG

153RD AES CONVENTION, NEW YORK, ONLINE EVENT, 27 OCTOBER 2022

Vision & Scope

Vision

- Develop a consistent and reliable objective quality metric for audio.
- Provide a valuable tool, e.g., for audio codec-related R&D.

Scope

- Demonstrate consistent and reliable stereo audio quality prediction for coded audio, including non-waveform coding tools, e.g., spectral band replication, parametric stereo.

We aim to make use of existing data and work with pre-trained models!

PRE-INVESTIGATION*

Mimicked the quality score predicted by an objective quality metric (ViSQOL-v3) with a deep neural network (DNN), followed by improving over it
- completely utilizing programmatically generated data!

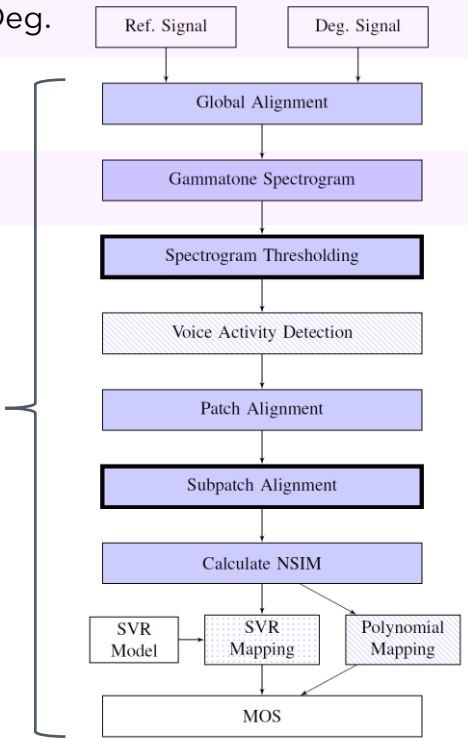
*G. Jiang, A. Biswas, C. Bergler, and A. Maier, "InSE-NET: A Perceptually Coded Audio Quality Model based on CNN," in *151st AES Convention, 2021* (Best Student Technical Paper Award).

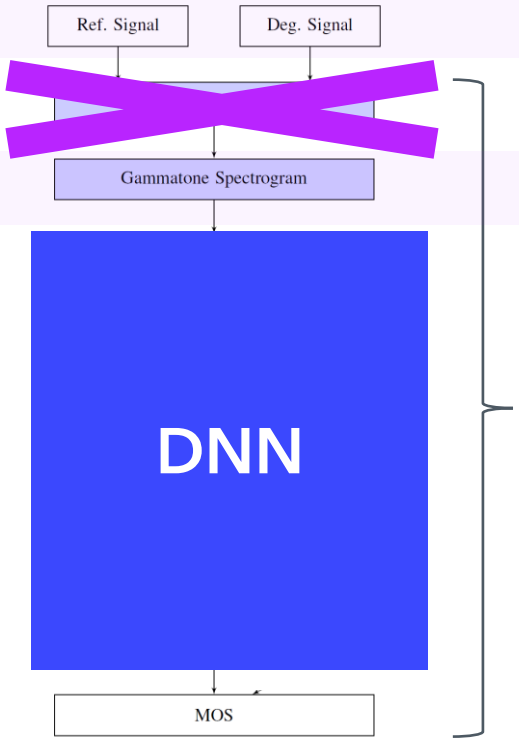
ViSQOL-v3 to InSE-NET

Training set: (HE)-AAC as Deg.

GAMMATONE SPECTROGRAMS

ViSQOL-v3*

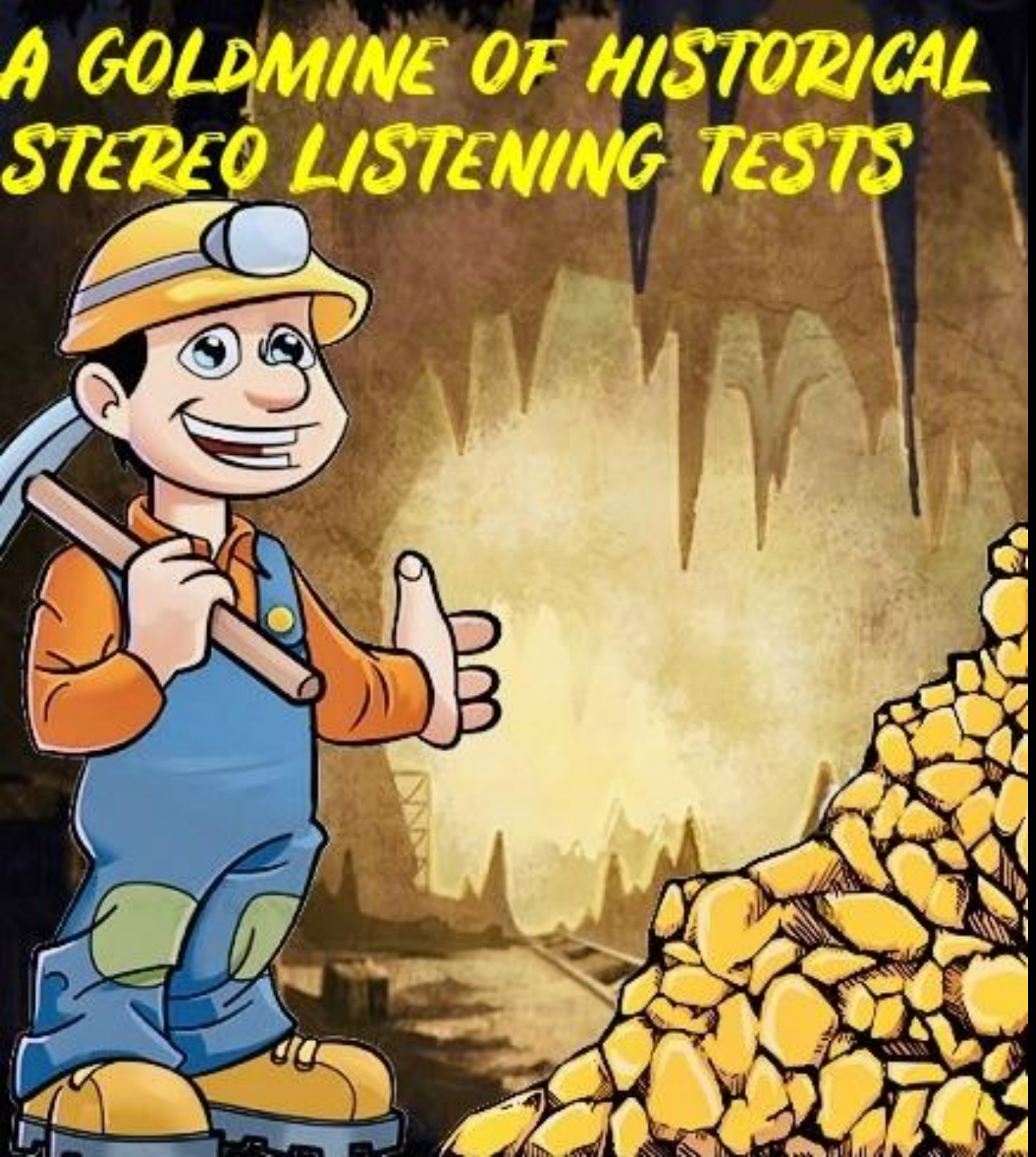




Mono InSE-NET

Regress with loss function

*https://github.com/google/visqol



Towards Stereo InSE-NET

PRIOR RESEARCH – AUDIO QUALITY MODELS BEYOND MONO

All the relevant papers are discussed in our paper.

Deep Learning-based models

Spatial audio quality metric (SAQAM)*: evaluates similarity between any pair of binaural signals in terms of localization accuracy and sound fidelity degradation, however:

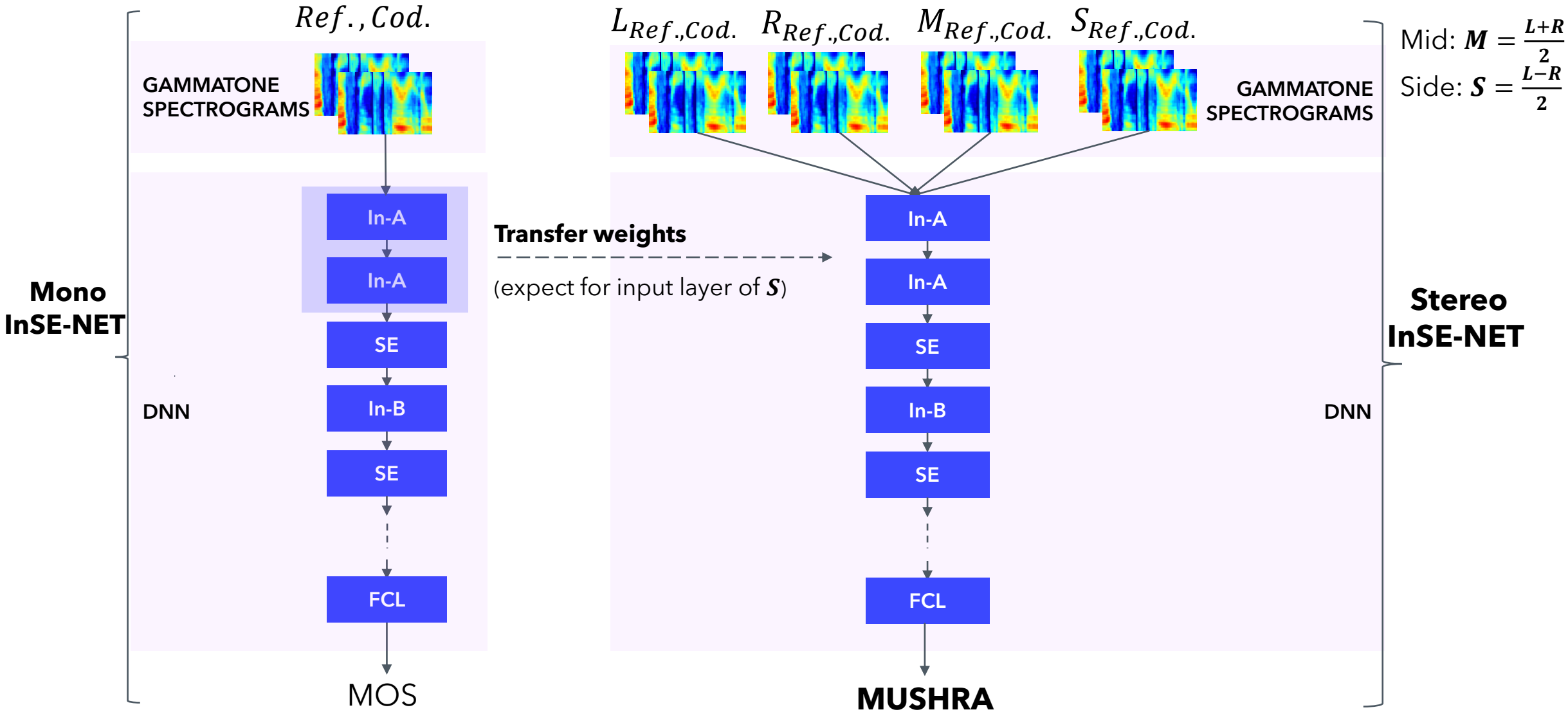
- ❑ **Does not predict an easy-to-interpret (e.g., MUSHRA) quality score.**

  Predict distances from the features delivered by the DNN

- ❑ **Only trained for speech signals at 16 kHz.**

Unaware of any deep learning-based (coded) stereo audio quality prediction model for general audio signals at 48 kHz!

*P. Manocha, et al., "SAQAM: Spatial Audio Quality Assessment Metric," in *Interspeech 2022*.



Stereo MPEG USAC Verification Listening Tests

	Stereo low-rates		Stereo high-rates		
	R_p	R_s	R_p	R_s	
ViSQOL-v3*	0.777	0.782	0.825	0.906	
Mono InSE-NET**	0.806	0.788	0.847	0.895	Trained to mimic mono ViSQOL-v3 (w/ noise & silence)
Stereo InSE-NET	0.888	0.838	0.892	0.874	Trained w/ stereo listening tests
Stereo InSE-NET	0.897	0.861	0.907	0.899	+ stereo listening tests with swapped LR
Stereo InSE-NET	0.915	0.88	**0.912**	**0.911**	+ hybrid stereo coding listening tests (& swapped LR)
Stereo InSE-NET (w/o M)	**0.922**	**0.900**	0.910	0.910	**Equivalent performance w/o mid-channel**

*ViSQOL-v3 compares the mid-signal: $M = \frac{1}{2}(L + R)$

**Signals fed to the model for comparison are the mid-signal.

Codecs included in the MUSHRA tests were AMR-WB+, HE-AAC, and USAC.
AMR-WB+ and USAC codecs were not seen during training. The lowest stereo HE-AAC bitrate was also not seen.

Mono MPEG USAC Verification Listening Tests

Mono low-rates

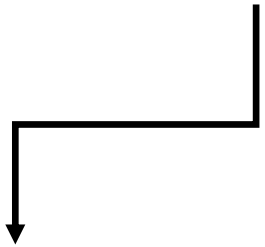
	R_p	R_s
PEAQ Advanced	0.650	0.700
ViSQOL-v3	0.810	0.840
Mono InSE-NET	0.830	0.835
Stereo InSE-NET*	**0.905**	**0.903**

Training with listening tests improves the correlation coefficients!

*The stereo signal fed to the model for comparison is dual-mono ($L = R$).

Mono listening tests were not used in training.
Codecs included in the MUSHRA tests were AMR-WB+, HE-AAC, and USAC.
AMR-WB+ and USAC codecs were not seen during training. The lowest stereo HE-AAC bitrate was also not seen.

Summary

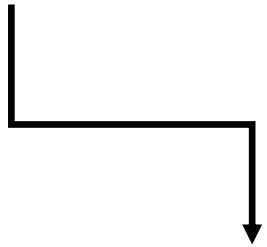


Stereo InSE-NET: deep learning-based coded stereo audio quality predictor at 48 kHz.

Programmatically generated data for training is powerful, but …

Listening tests and data derived from listening tests also provides a benefit.

Next steps



Data augmentation: continue to *engineer training data* utilizing audio (coding) domain expertise.

Beyond stereo: binaural, multi-channel, …

—

THANK YOU

RESULTS PER CODEC

Stereo MPEG USAC Verification Listening Tests

Stereo low-rates

	Mono InSE-NET		Stereo InSE-NET	
Codecs	R_p	R_s	R_p	R_s
AMR-WB+	0.868	0.842	**0.960**	**0.904**
HE-AAC	0.830	0.790	**0.945**	**0.877**
USAC	0.891	0.860	**0.976**	**0.943**

Stereo MPEG USAC Verification Listening Tests

Stereo high-rates

	Mono InSE-NET		Stereo InSE-NET	
Codecs	R_p	R_s	R_p	R_s
AMR-WB+	0.864	0.852	**0.955**	**0.925**
HE-AAC	0.871	0.925	**0.946**	**0.949**
USAC	0.909	0.920	**0.964**	**0.942**

Mono MPEG USAC Verification Listening Tests

Mono low-rates

	Mono InSE-NET		Stereo InSE-NET	
Codecs	R_p	R_s	R_p	R_s
AMR-WB+	0.889	0.856	**0.948**	**0.922**
HE-AAC	0.853	0.791	**0.945**	**0.887**
USAC	0.873	0.881	**0.950**	**0.939**