Design Choices and Effective Evaluation of Modern Speech and Audio Codecs Based on Neural Networks

Nicola Pia

Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

ICASSP2026 Barcelona

LRAC 2026 WORKSHOP

© Fraunhofer IIS/Rida El Ali

Fraunhofer IIS

History of excellence in audio

- Codecs
  - EVS
  - MPEG-H
  - MPEG-I
- And more
  - Allinga (TTS)
  - Symphoria
    (immersive rendering in cars)
  - JPEG XS
  - …

© Fraunhofer IIS

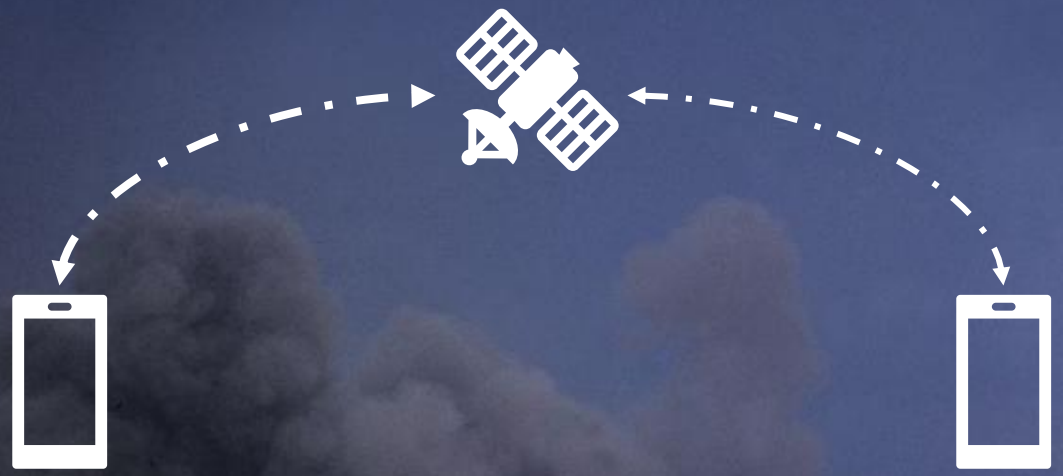
We live in highly connected world

Coding technology has huge impact

© Fraunhofer IIS

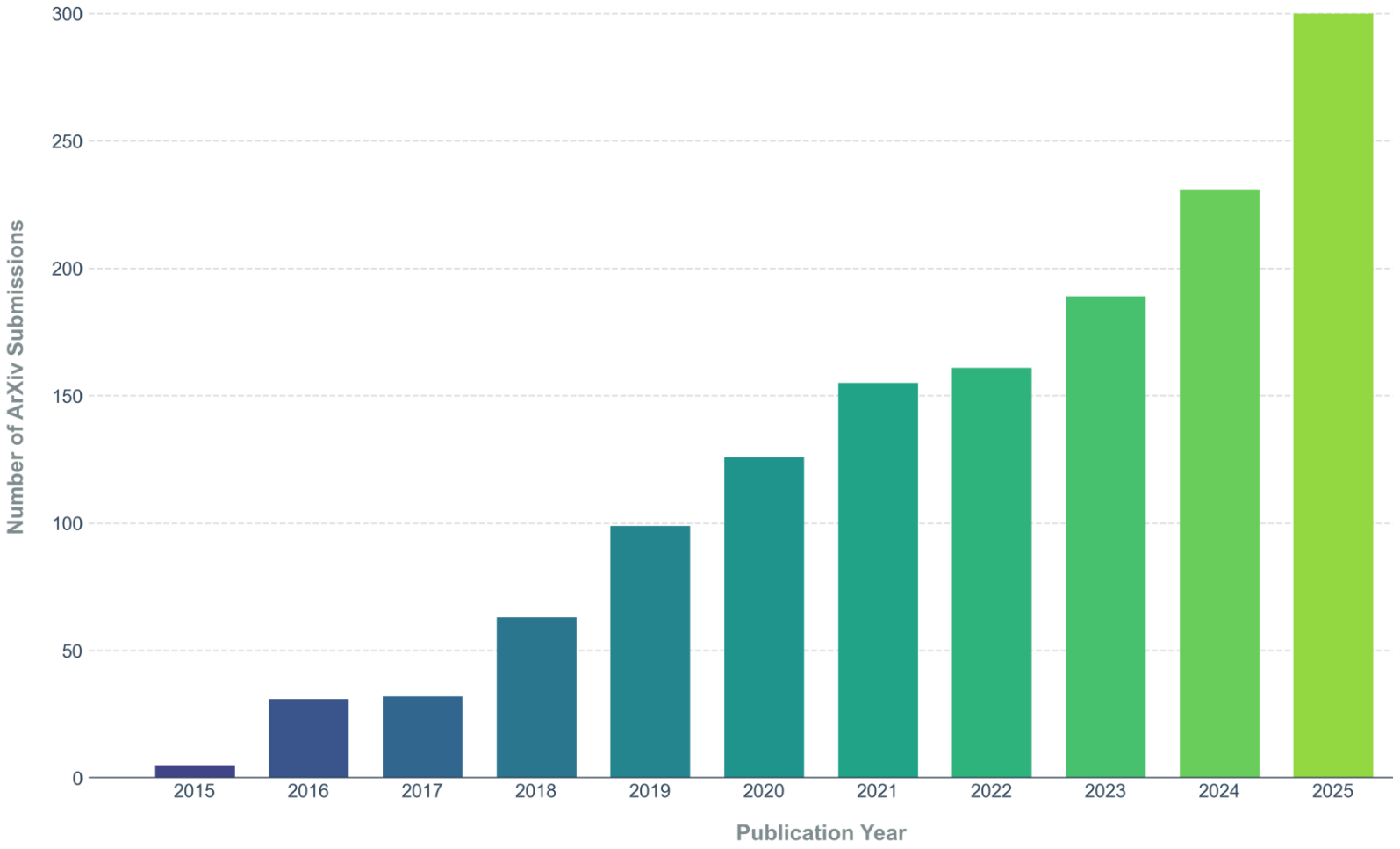
There is a need for low bit rates

Satellite communication



The research is exploding

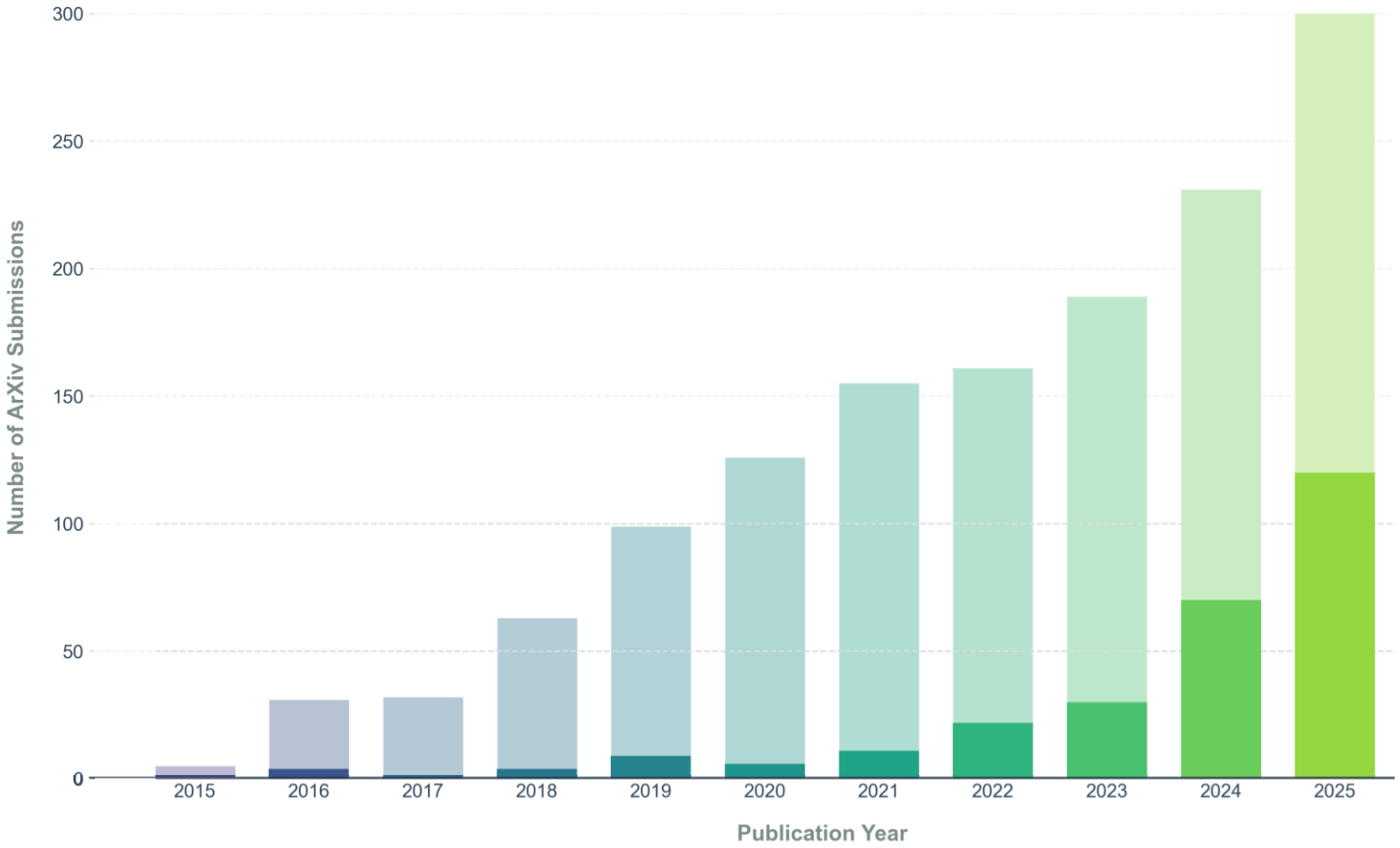
Number of papers related to coding, compression and neural networks



© Fraunhofer IIS

The research is exploding

Number of papers about neural coding



© Fraunhofer IIS

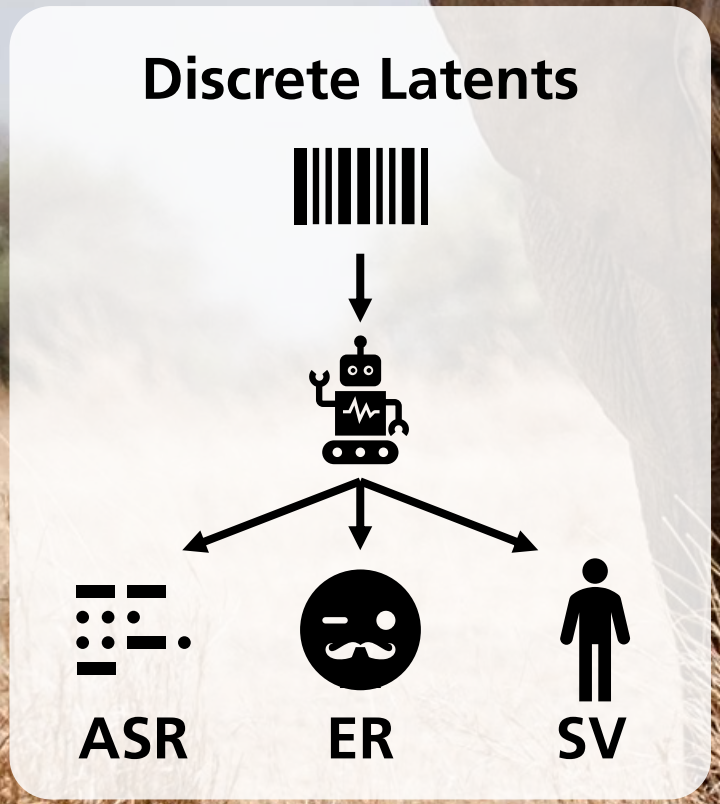
Navigating the jungle

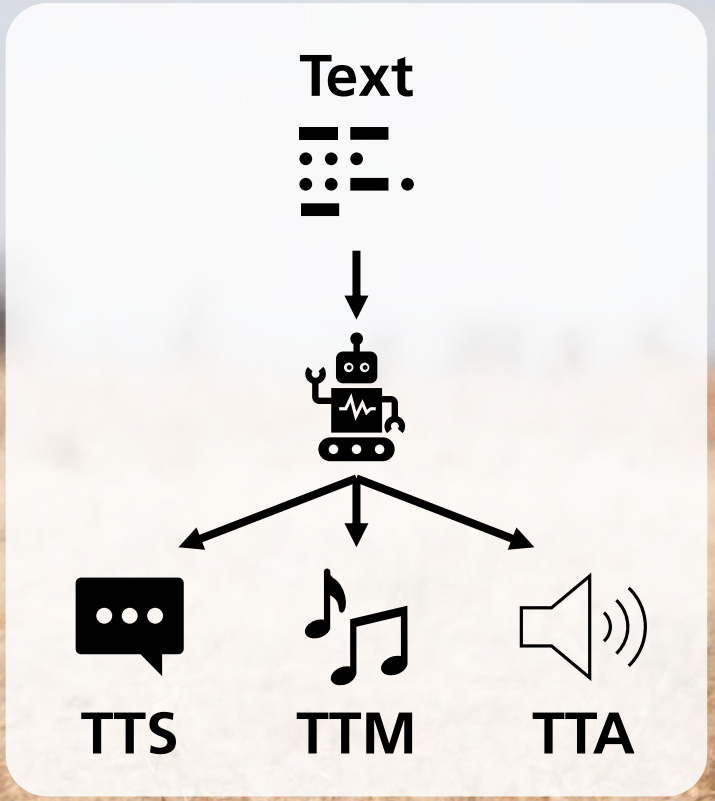
What makes it difficult

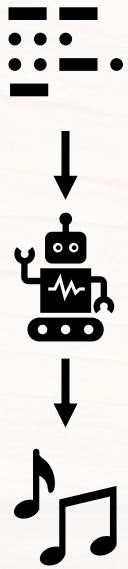
Navigating the jungle
What makes it difficult

Many papers tackle other applications

The elephant in the room





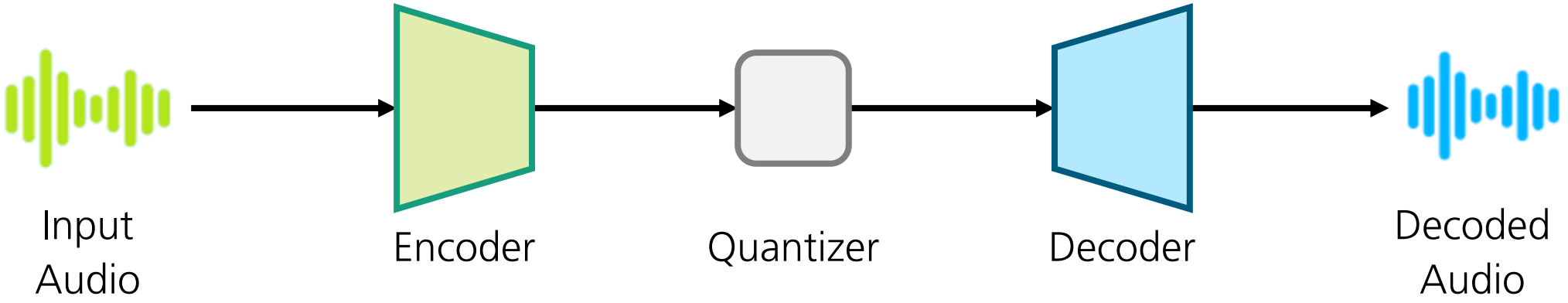


- Could sound bad on other signal types
- Can be high-complex
- Can have high delay

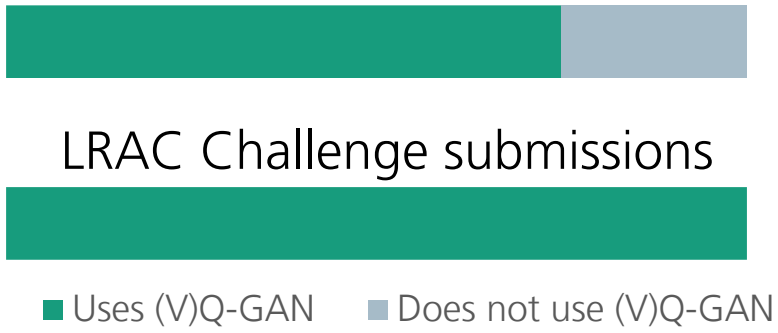
Navigating the jungle
What makes it difficult

Neural speech and audio coding

The most common framework: End-to-end (V)Q-GAN



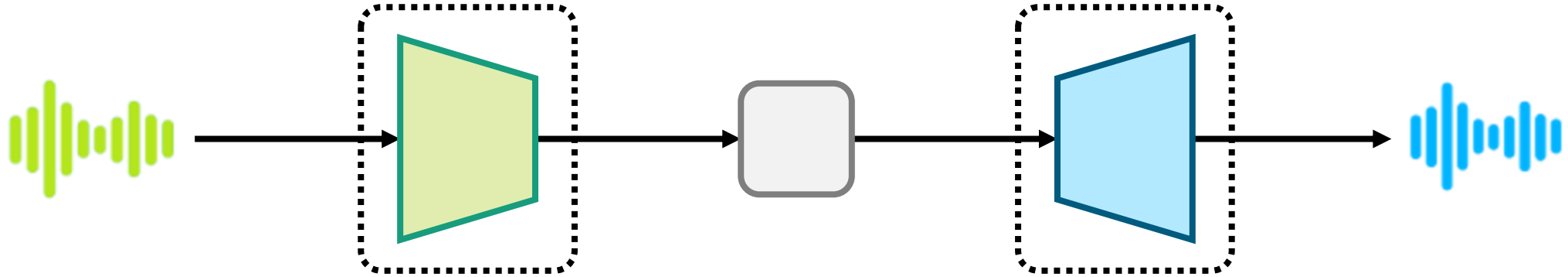
Discrete audio tokens: More than a survey!



© Fraunhofer IIS
Mousavi, Pooneh, et al. "Discrete audio tokens: More than a survey!." *arXiv preprint arXiv:2506.10274* (2025).

A plethora of encoder-decoder architectures …

There is no universal backbone (yet)



Transform domain (e.g., STFT) vs StridedConv + ConvTranspose

Symmetrical encoder-decoder?

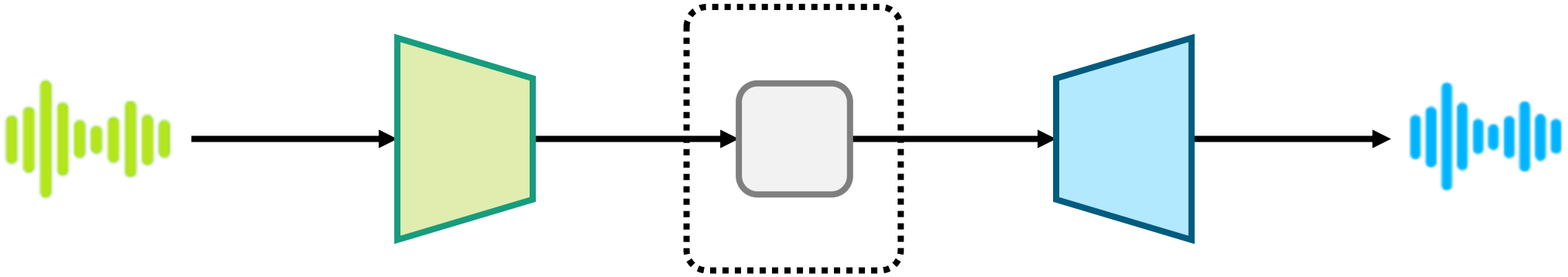
Use RNN/Transformer for long term dependencies?

Block:
Conv1d vs Conv2d
LKCA
ConvNeXt
…

© Fraunhofer IIS

… and of quantizers

There is no universal backbone (yet)



RVQ design:
Number of stages
Unused codebooks
…

Grouped RVQ
Multi-Scale RVQ
Product RVQ
…

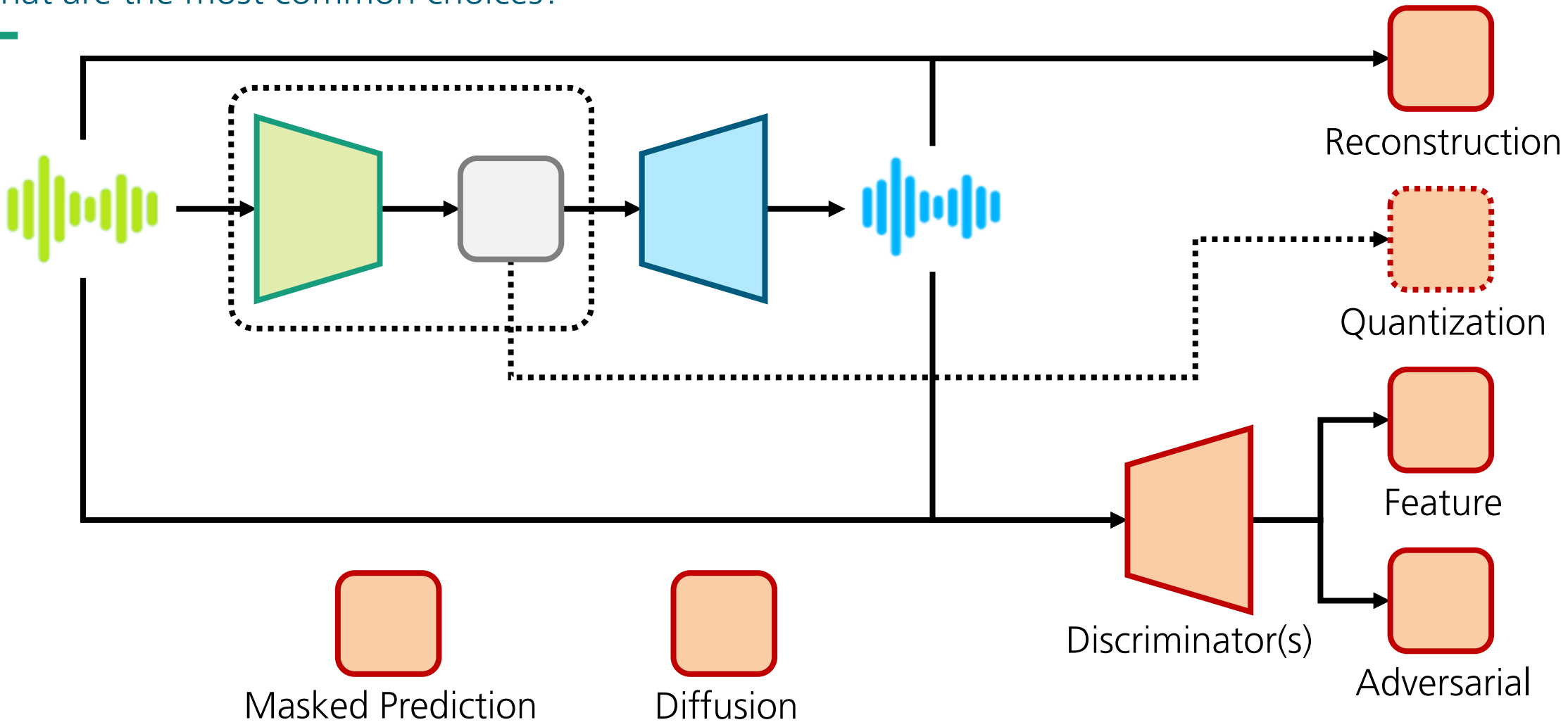
Finite Scalar Quantization?

Training tricks:
Entropy Constraints
Distillation
…

© Fraunhofer IIS

The training losses

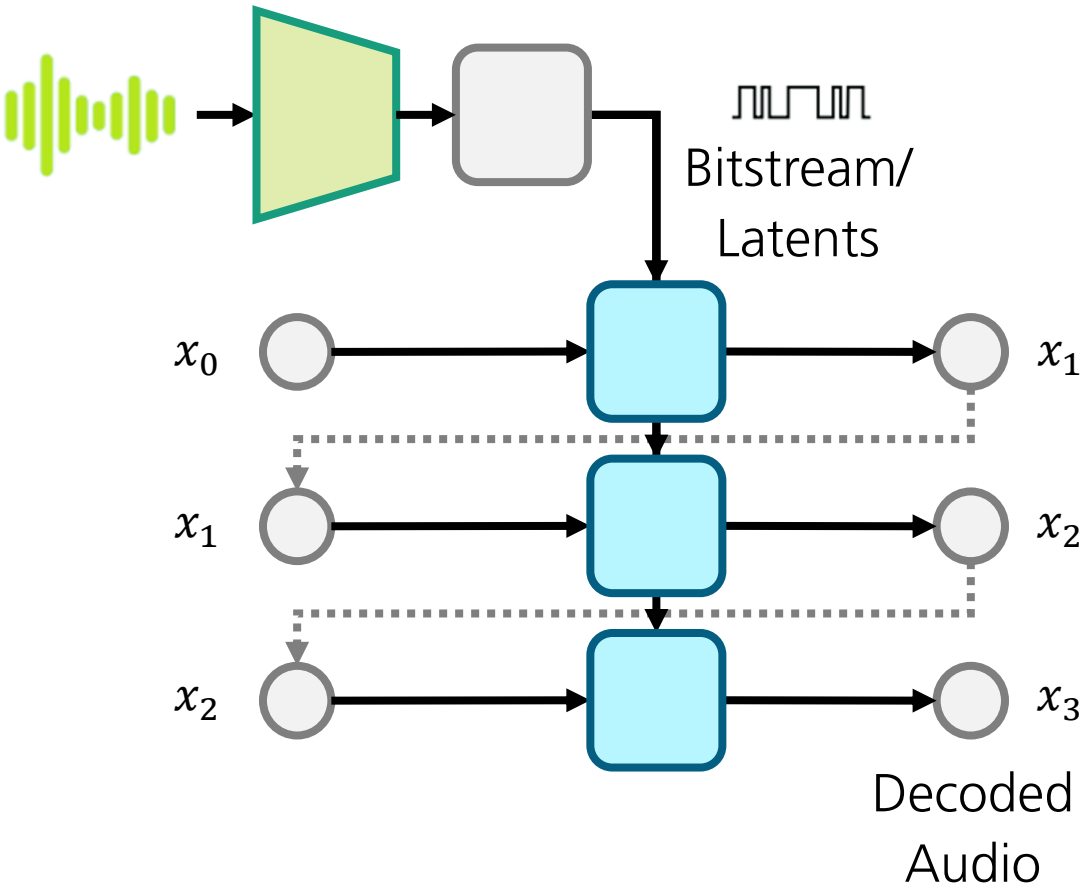
What are the most common choices?



© Fraunhofer IIS

AutoRegressive decoders

This is where the field started



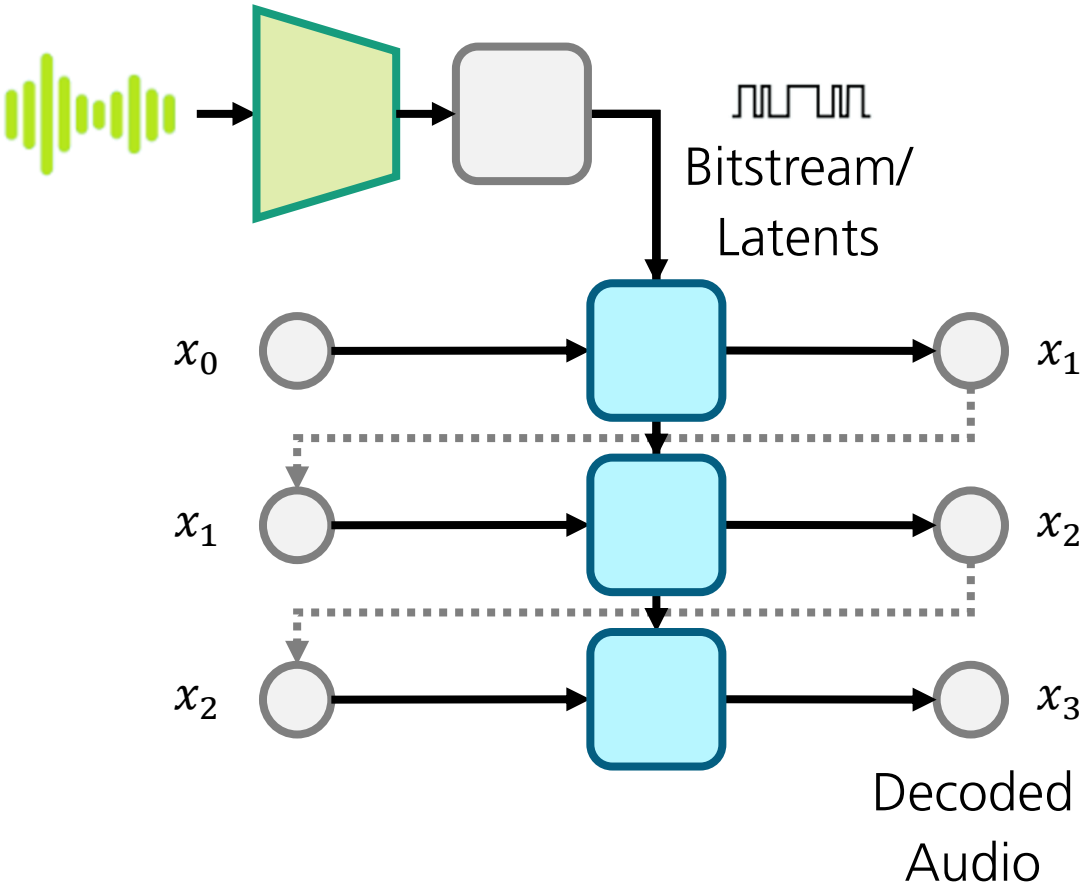
$$NN(\boldsymbol{x};\theta) = p_\theta(\boldsymbol{x}) = \prod_{t=1}^{n} p_\theta(x_t|x_1,\ldots,x_{t-1})$$

$$\mathcal{L}_{AR}(\theta) = -\mathbb{E}_{\boldsymbol{x}\sim p_{data}} \log\big(NN(\boldsymbol{x};\theta)\big)$$

© Fraunhofer IIS

AutoRegressive decoders complexity

Need clever design for low complexity



- For naïve RNN in waveform domain:
  - each parameter used once per output sample, i.e.

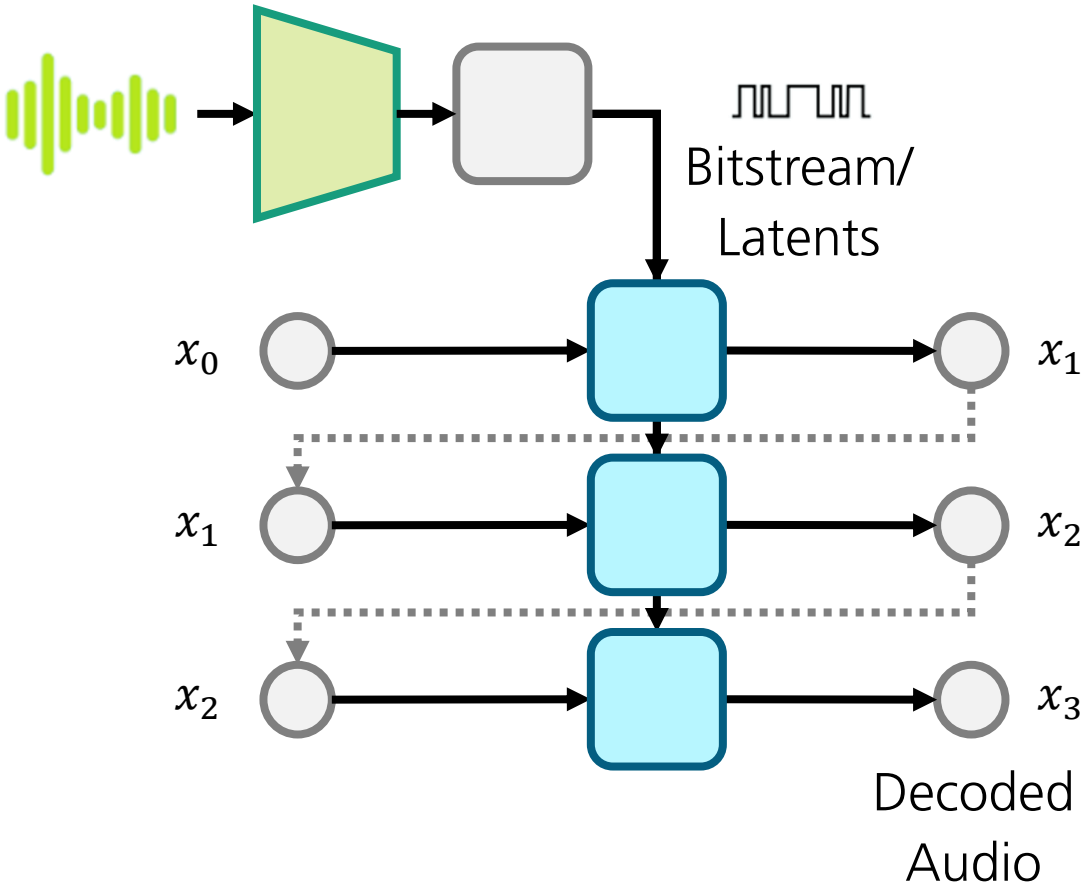
```
tot_macs = num_params * sampling_rate

# LRAC assumptions
tot_macs = 300_000_000
sampling_rate = 24_000
# ->
num_params = tot_macs / sampling_rate = 12_500
```


© Fraunhofer IIS

AutoRegressive decoders design

Clever design is possible

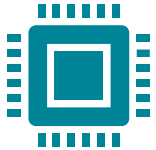
1. **Integration with classical DSP**
   LPC, MDCT, Subband decoding, …
2. **Use only for submodules**
   PLC, Entropy Coding, …
3. **Student-teacher pipeline**



© Fraunhofer IIS

AutoRegressive decoders in summary

Challenging design choices that might pay-off?



Easy to train and flexible
WaveNet Codec2 Decoder

Great compatibility with classical DSP
LPCNet

With correct design (and hardware) good alternative to VQ-GAN!

Might be prone to error accumulation?
LPCNet vs StreamwiseStyleMelGAN

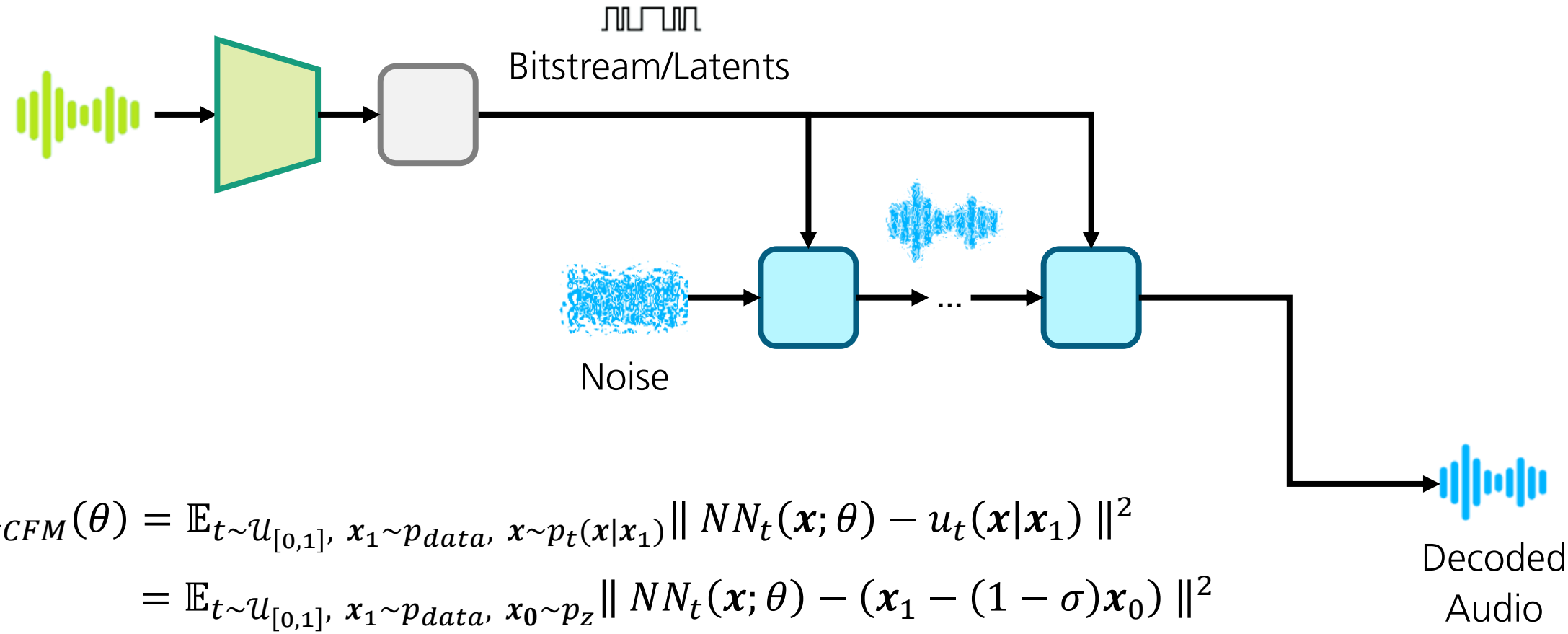
Can achieve high-quality (at high computational cost?)
MDCTNet

© Fraunhofer IIS

Kleijn, W. B., et al. "Wavenet based low rate speech coding." ICASSP 2018.
Valin, J.-M. and Skoglund J. "A Real-Time Wideband Neural Vocoder at 1.6 kb/s Using LPCNet." Interspeech 2019.
Mustafa, A., et al. "A streamwise gan vocoder for wideband speech coding at very low bit rate." WASPA 2021.
Davidson, G., et al. "High quality audio coding with MDCTNet." ICASSP 2023.

Diffusion models

Multiple design choices for the architecture



$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim\mathcal{U}_{[0,1]},\ \boldsymbol{x}_1\sim p_{data},\ \boldsymbol{x}\sim p_t(\boldsymbol{x}|\boldsymbol{x}_1)} \| NN_t(\boldsymbol{x};\theta) - u_t(\boldsymbol{x}|\boldsymbol{x}_1) \|^2$$

$$= \mathbb{E}_{t\sim\mathcal{U}_{[0,1]},\ \boldsymbol{x}_1\sim p_{data},\ \boldsymbol{x_0}\sim p_z} \| NN_t(\boldsymbol{x};\theta) - (\boldsymbol{x}_1 - (1-\sigma)\boldsymbol{x}_0) \|^2$$

© Fraunhofer IIS

Elucidating the design space

Multiple design choices for training and inference

Loss

- Predict noise or signal
- Stochastic vs deterministic
- Select vector field and coupling

Hyperparameters

- Noise schedule
- Timestep sampling

Inference

- Select solver and parameters
- Stochastic vs deterministic
- Distill?

Question: Should we study this more systematically for coding?

© Fraunhofer IIS

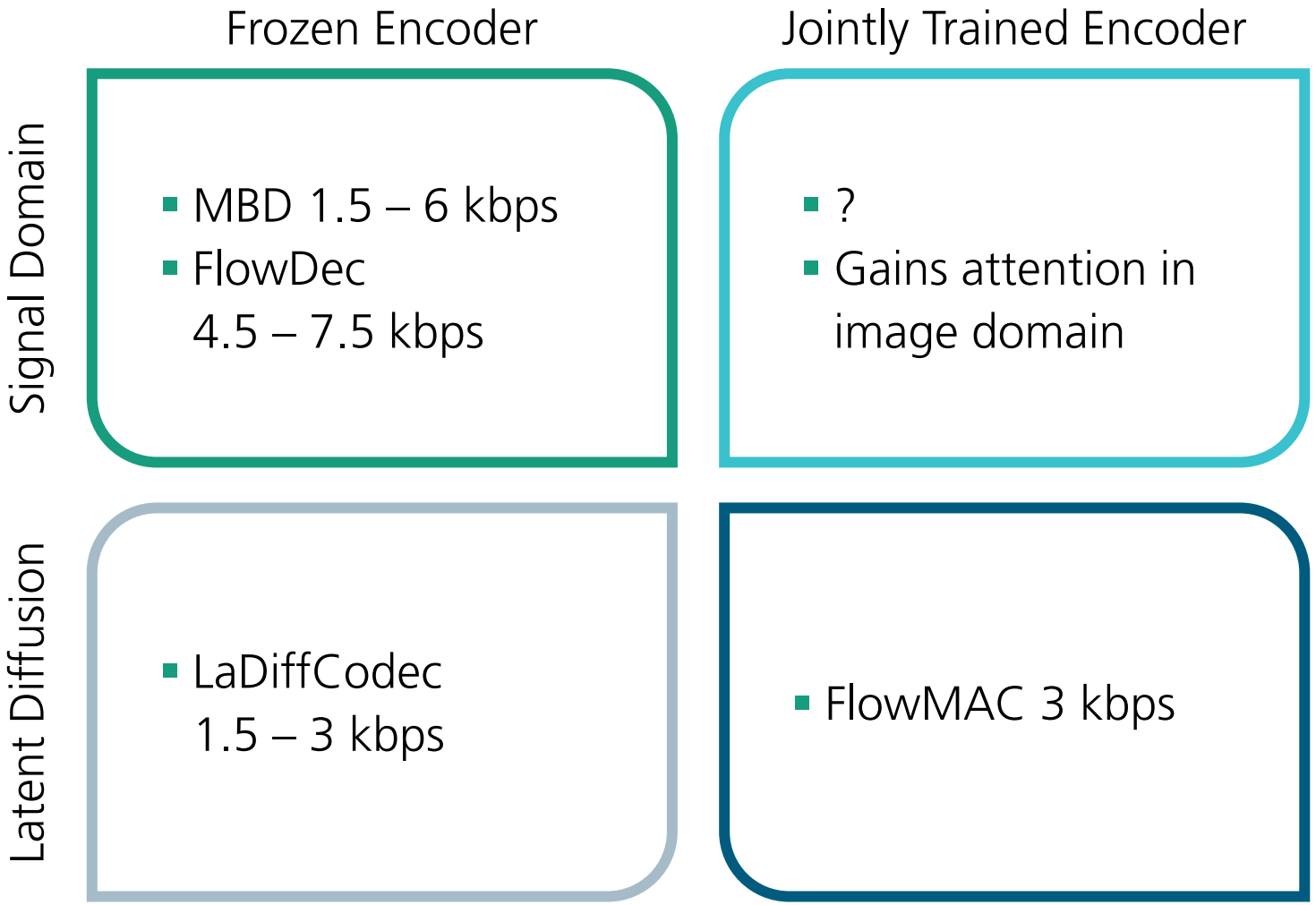
Karras, T., et al. "Elucidating the design space of diffusion-based generative models." Advances in neural information processing systems, 2022

Diffusion models in the literature

Many approaches have been tried

✅ GAN-free
❌ higher complex

✅ lower complex
❌ multi-phase decoder training

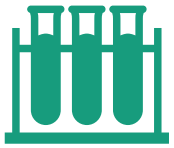


San Roman, R., et al. "From discrete tokens to high-fidelity audio using multi-band diffusion." NeurIPS 2023
Yang, H., et al. "Generative de-quantization for neural speech codec via latent diffusion." ICASSP 2024
Welker, S., et al. "FlowDec: A flow-based full-band general audio codec with high perceptual quality." ICLR 2025
Pia, Nicola, et al. "FlowMAC: Conditional flow matching for audio coding at low bit rates." ICASSP 2025

© Fraunhofer IIS

Diffusion models in summary

Computational cost is still too high?

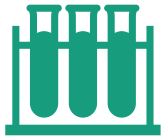


Easy to scale, needs from careful tuning
Loss, inference steps, solver, …

Computational cost is high
For now …

Quality comparable to GANs
GANs still more efficient

Gains momentum in many applications
Mel-to-audio, SE, …



© Fraunhofer IIS

1. **GANs dominate the literature**
   Harder to train but best trade-offs
2. **AR and Diffusion show promising quality**
   Computational cost still too high?
3. **A fair comparison is hard**
   Control variables: model type, hyperparameters, …

Is this all there is?

Encoder-Quantizer-Decoder Architecture + Training?

Navigating the jungle

What makes it difficult

There is no widely adopted method!

When is the evaluation believable?

Quality evaluation

The methods for quality evaluation

Objective metrics
(POLQA, STOI, ViSQOL, ScoreQ, NOMAD, …)

✅ Quick and reproducible
✅ Good for big test sets

❌ Hard to model all aspects of human perception

In-Lab listening tests
(MUSHRA, P.800, …)

✅ Extremely reliable (if conducted well)

❌ Costly (time and resources)

Crowd-sourced LTs
(P.808, MUSHRA-1S, …)

✅ Cost efficient and fast

❌ Not always as reliable as or consistent with in-lab

Question: can we keep the best of these two?

© Fraunhofer IIS

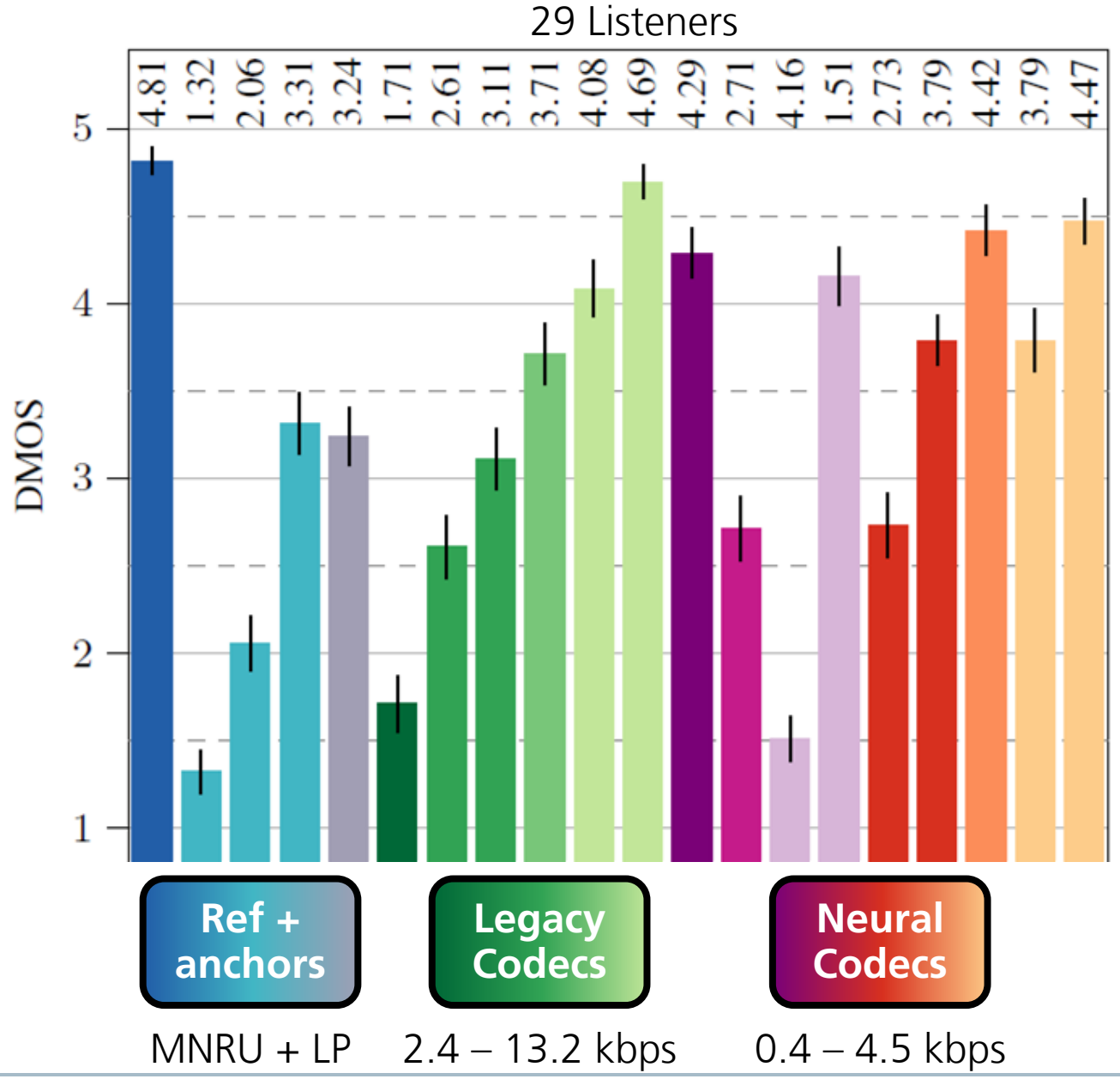
In-Lab P.800 is a gold standard

For industry and standardization bodies

In-Lab P.800

How we did it

- **Preparation of the test**
  - P.800 DCR on English clean speech
  - Choose correct anchors
  - Listener pool with balanced demographics
    - (mostly) German native speakers
- **Listening sessions**
  - Controlled environment
  - Gather statistics and audiogram
- **Data filtering and post-processing**
  - Minimal filtering following standard



DCR = Degradation Category Rating

© Fraunhofer IIS

Crowd-Sourced P.808

Cost-effective but unreliable?

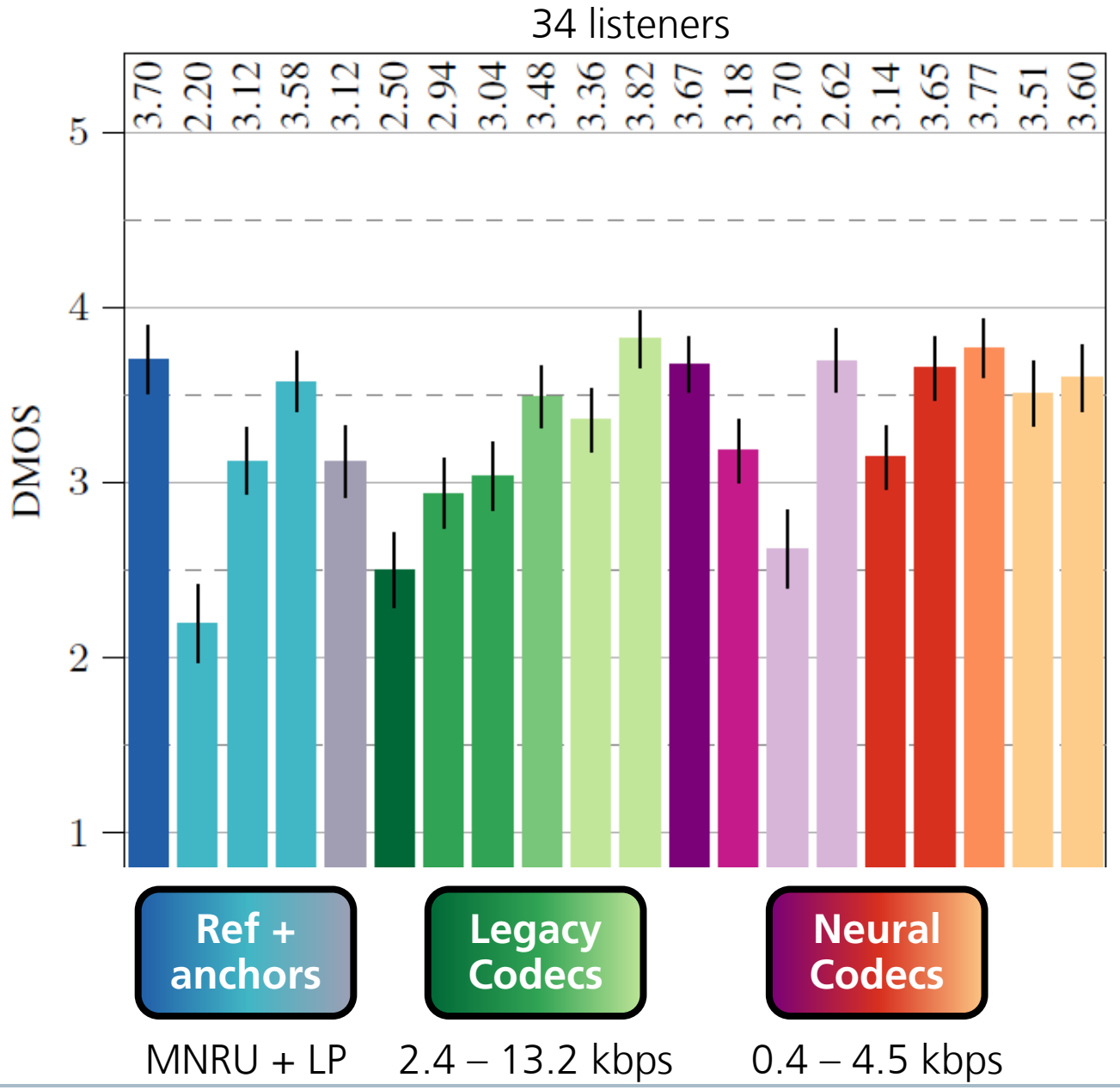
(without screening)

❌ Scale is not fully used

❌ Reference is scored below 4

❌ Some conditions are better than the reference

❌ All scores are close together



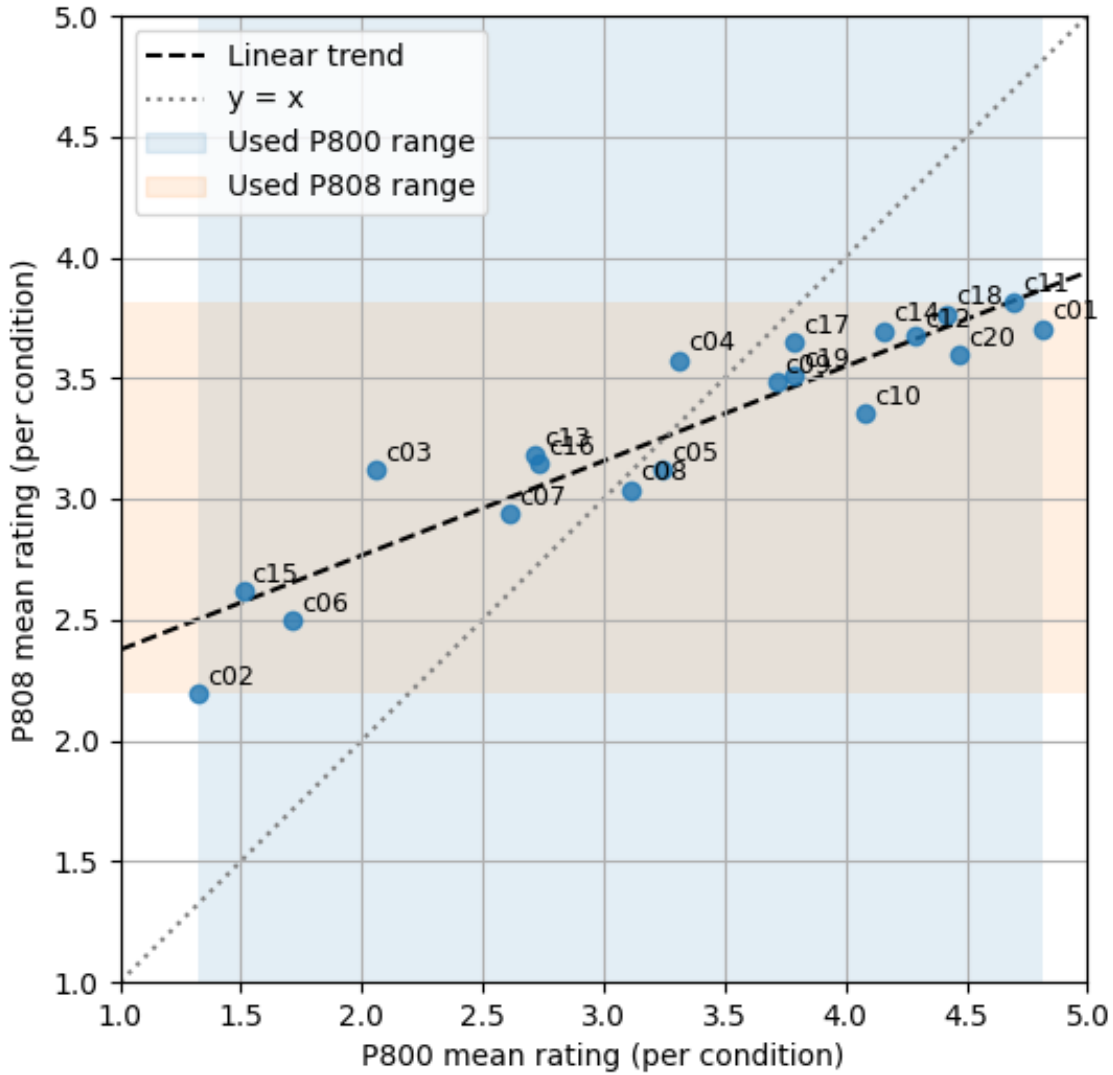
© Fraunhofer IIS

Do we have a problem?

Correlation is high but ratings are different

- Measure **Closeness** of results distribution via
  - Mean Absolute Error (MAE)
  - Root Mean Square Error (RMSE)
- Measure **Correlation** via
  - Pearson correlation r
  - Spearman's rank ρ
- **Question:** How do we make P.800 and P.808 as close as possible?

❌ $MAE = 0.573$ ❌ $RMSE = 0.659$ ✅ $r = 0.929$ ✅ $\rho = 0.929$



© Fraunhofer IIS
Fraunhofer IIS

How to solve the issue

Select only the reliable results of the crowd-sourced listening test

Before the test

- **Pretest**: Check if listeners have suitable hardware and environment
- **Questionnaire**: Gather relevant statistics

During the test

- **Traps questions**: Randomly require to select a specific rating
- **Gold standards**: Monitor the ratings of the reference (or anchor)

After the test

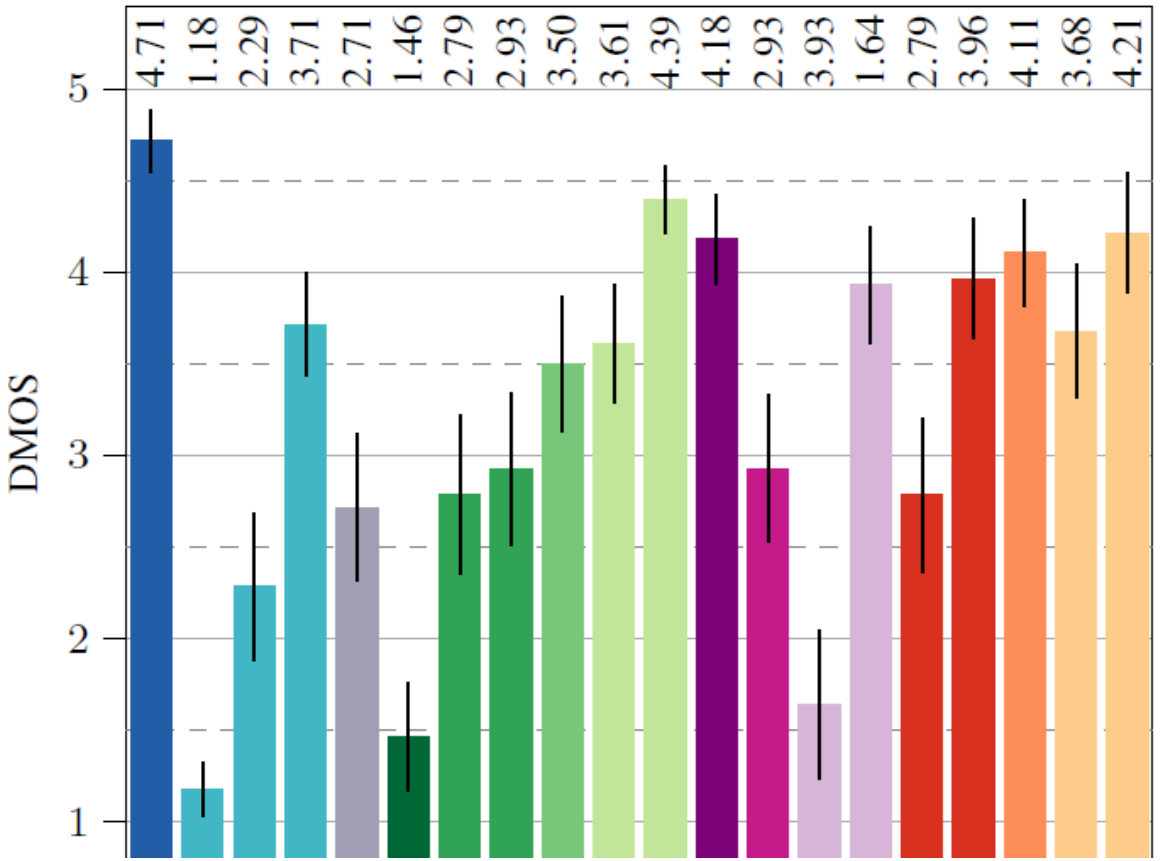
- **Rating span**: Filter out based on difference between reference and lowest anchor
- **Anchor ordering**: Filter out if wrong ordering of the MNRUs (10, 17, 24)

© Fraunhofer IIS
required/recommended in the standard

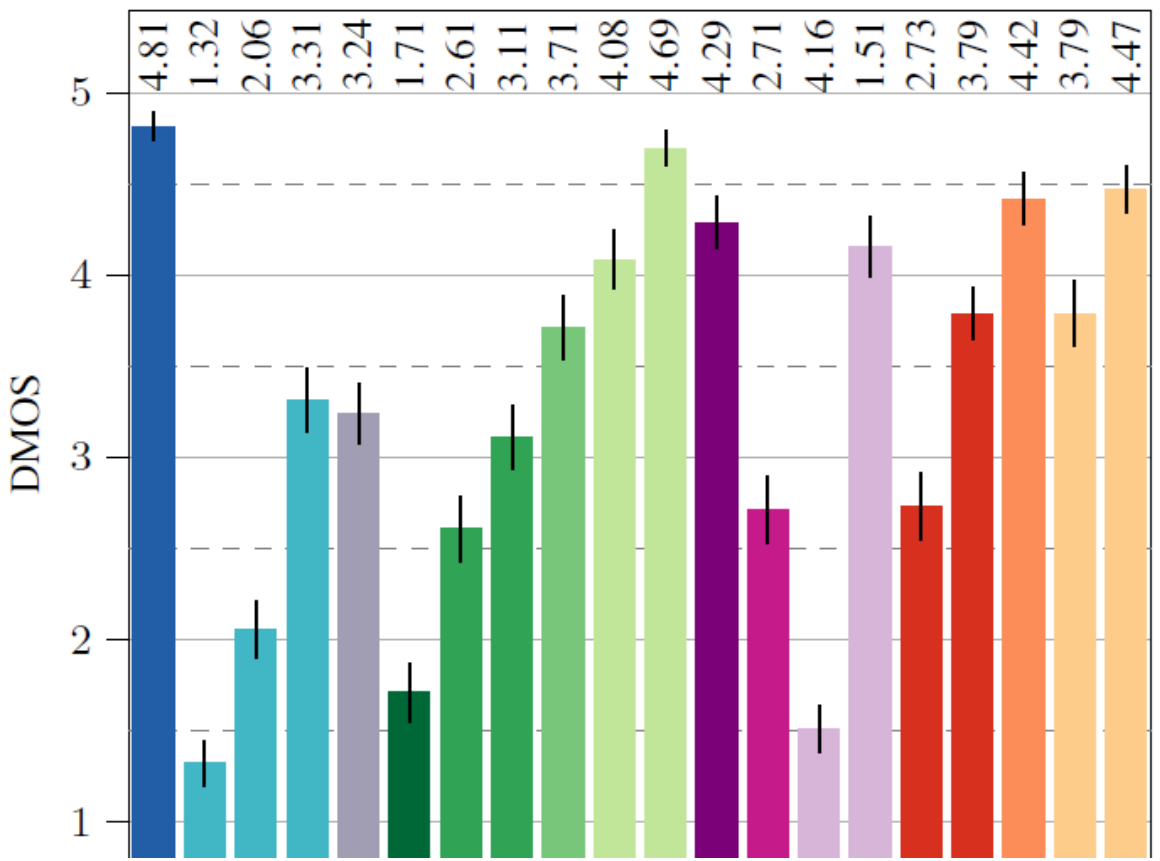
Screening is important

Gold standard + scale span + anchor ordering

P.808 filtered results (14 listeners)



P.800 (29 listeners)

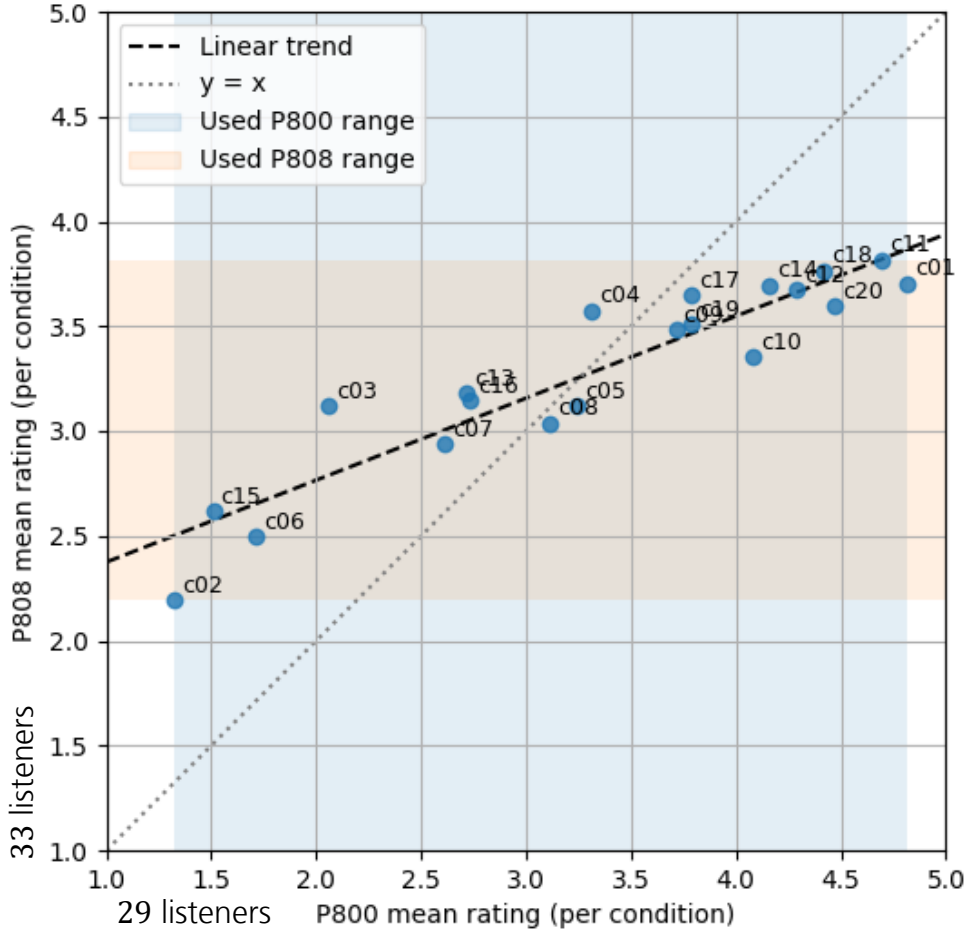


© Fraunhofer IIS

We can solve the problem!

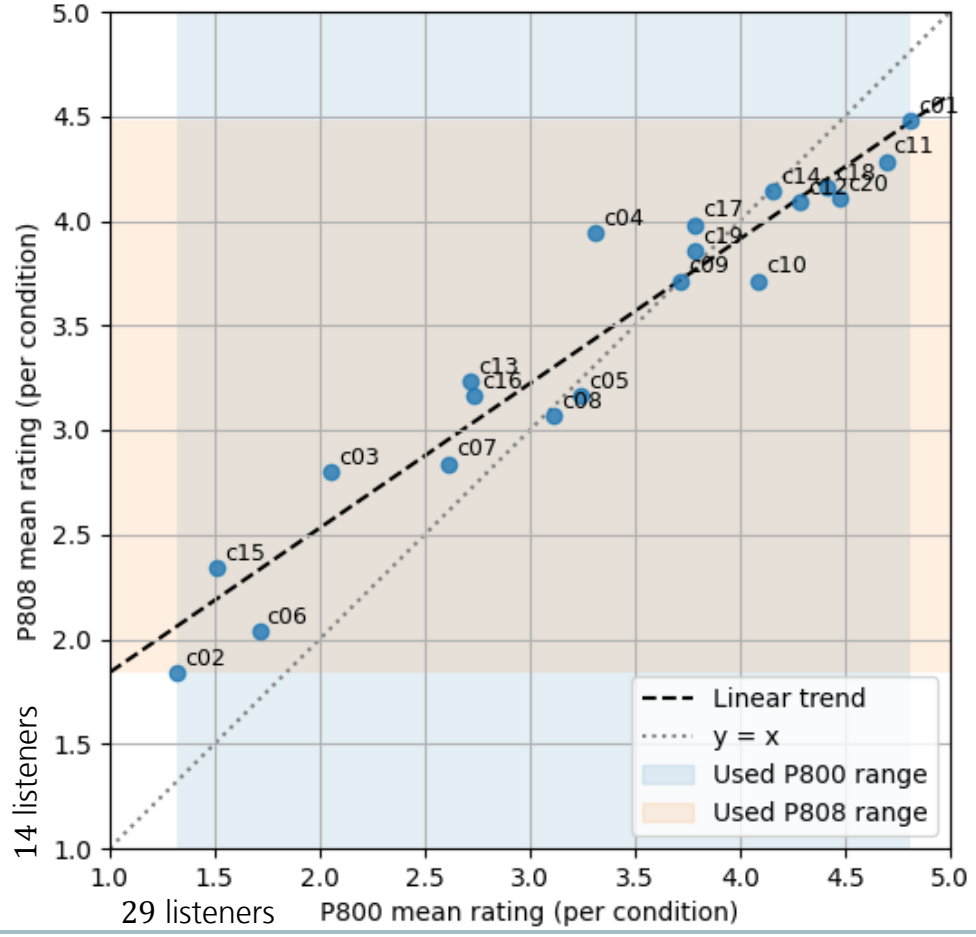
Make P.800 and P.808 as close as possible improving correlation

❌ $MAE = 0.573$ ❌ $RMSE = 0.659$ ✅ $r = 0.929$ ✅ $\rho = 0.929$



Filter results

✅ $MAE = 0.230$ ✅ $RMSE = 0.259$ ✅ $r = 0.974$ ✅ $\rho = 0.958$



© Fraunhofer IIS

Can we fix crowd-sourced P.808?

Yes, with correct screening!

1. Standard screening is ineffective
2. Gold standard + scale rating + anchor ordering works
3. P.808 with ~3x listeners still cheaper than P.800 but as effective


Goal
Evaluation
Techniques

Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

Thank you for your time!

nicola.pia@iis.fraunhofer.de