MAKING TRANSFORMERS WORK FOR AUDIO CODING

JULIAN D. PARKER, STABILITY AI

WHO AM I?

- Physics -> Musical Acoustics -> DSP -> AI
- Worked as a researcher in academia and industry for 15 years.
- Most industrial work has focused on processing or generating musical sound using DSP and latterly ML/AI.

THE UNIVERSITY OF EDINBURGH

Aalto University

NATIVE INSTRUMENTS

TikTok

stability.ai

stability.ai

Prem Akkaraju
CEO

Sean Parker
Executive Chairman

James Cameron
Board Member

- We make **open-weights** generative models for **professional media production workflows**, across many modalities (image, video, 3d, audio).
- I'm part of the **Audio** team, which is primarily concentrated on **music + general audio** (not speech).
  - We released most popular open-weights model for audio generation - **Stable Audio Open**

WHAT IS THE TOPIC FOR TODAY?

"Scaling Transformers for Low-Bitrate High-Quality Speech Coding"

Julian D. Parker, *Anton Smirnov*, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, *Xubo Liu*

Accepted at ICLR 2025 in Singapore - come meet us if you're there!

SETTING THE SCENE

STATUS QUO IN MARCH 2024

- Most successful codecs for generative use (especially music) are **Encodec** and **DAC**, both of which use broadly the same arch.
- **Convolutional** arch built on fairly **old** (circa 2016 or earlier) structures (ResNet, dilated convs etc).
- **Relatively small model size**, with no clear path to scaling.
- Improvements mainly coming from adding more **complicated training objectives** and **discriminators**.

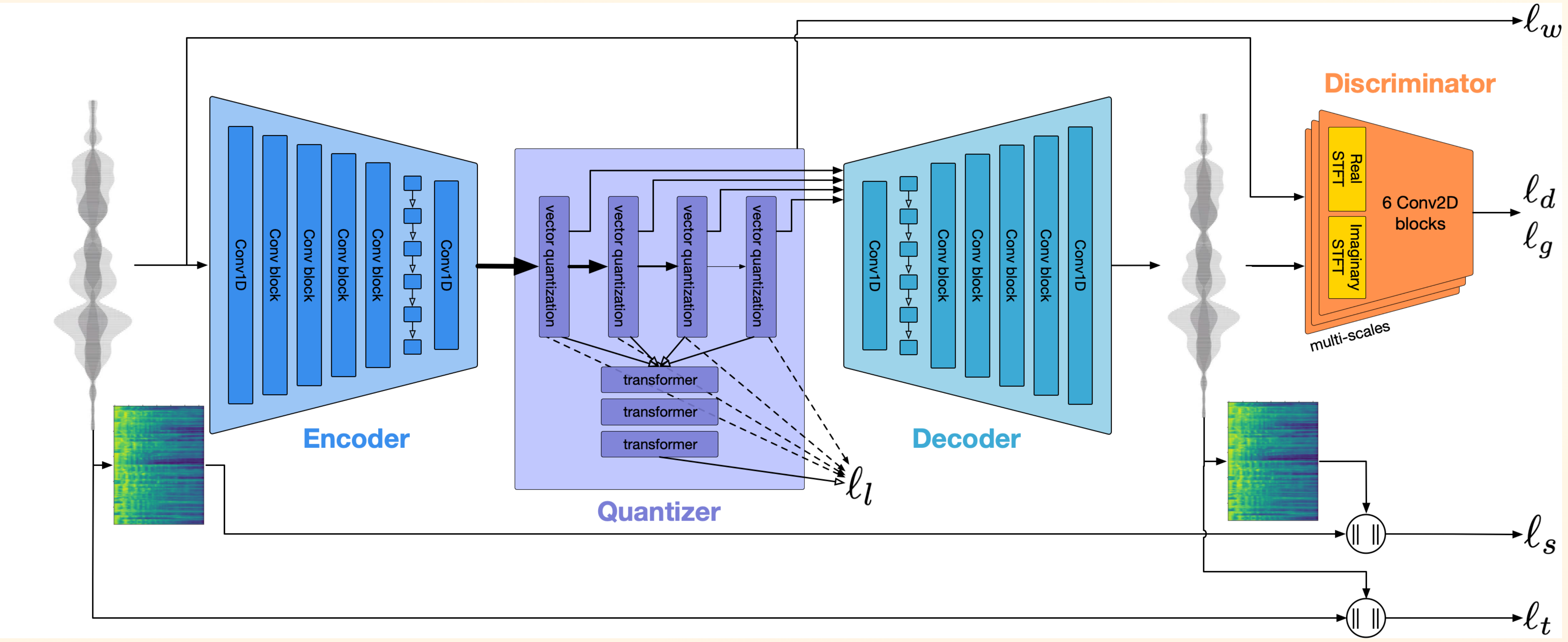


Figure adopted from A. Défossez

MOTIVATION - WHY TRANSFORMERS?

- *Obvious reason*: **transformers** have become **default architecture** for most problems
  - Great **scaling** properties
  - Mature + **optimized** implementations
- *Personal reason*: been hurt too many times by the `**bitter-lesson**`, why not try a very generic approach?
- *Principled reason*: most **traditional compression** algorithms **heavily leverage non-uniform compression** across a sequence, convolutional neural codecs do not.
  - Attention is very effective at moving and rearranging information across a sequence - maybe there's potential to exploit this.

MOTIVATION - WHAT DID WE WANT TO ACHIEVE?

- **Viable architecture** where the **majority of parameters** are in **transformer** blocks.
- **High quality** reconstruction at very **low bitrate**.
- Evidence that **scaling parameter** count **improves reconstruction** quality.
- Try out some interesting new techniques.
- **N.B.** We don’t have production speech models at Stability, so this was intended primarily as a research work. Many decisions reflect this. Novelty > performance.

DESIGN DECISIONS

BROAD DESIGN PRINCIPLES

- Stick with overall architecture from Encodec + DAC
  - **Encoder** > **Bottleneck** > **Decoder** + **Discriminator**
  - No generative decoder or post-filter
- Try to **eliminate** the majority of **convolutional elements**.
- Utilize **standard transformer blocks**.
- Aim for bottleneck to use **low number** of **tokens** per **timestep** (no 8-level RVQ).
- Prioritise **natural** audio **quality**.

TRANSFORMERS NEED EMBEDDINGS

OPTIONS

- Spectrograms
  - Mel spectrograms limited by inversion techniques ❌
  - Linear complex spectrograms not *critically sampled* (apart from special cases) ❌
- Existing convolutional up/downsampling networks add extra complexity ❌
- MDCT/wavelets work well, but errors more audible, + technically not perfect reconstruction
- Patching is better! Critically sampled, perfect reconstruction, very simple + error more noise-like ✅

BOTTLENECK OPTIONS

- **VQ / RVQ**
  - Requires **auxillary losses** and s**traight-through gradient estimation**
  - Generally used successfully with many residual tokens.
- **FSQ**
  - appealingly simple (**no auxillary losses**)
  - can get around **straight-through gradient estimation** using noise
  - downside - very few configurations that lead to sensible codebook sizes
    - Residual version?

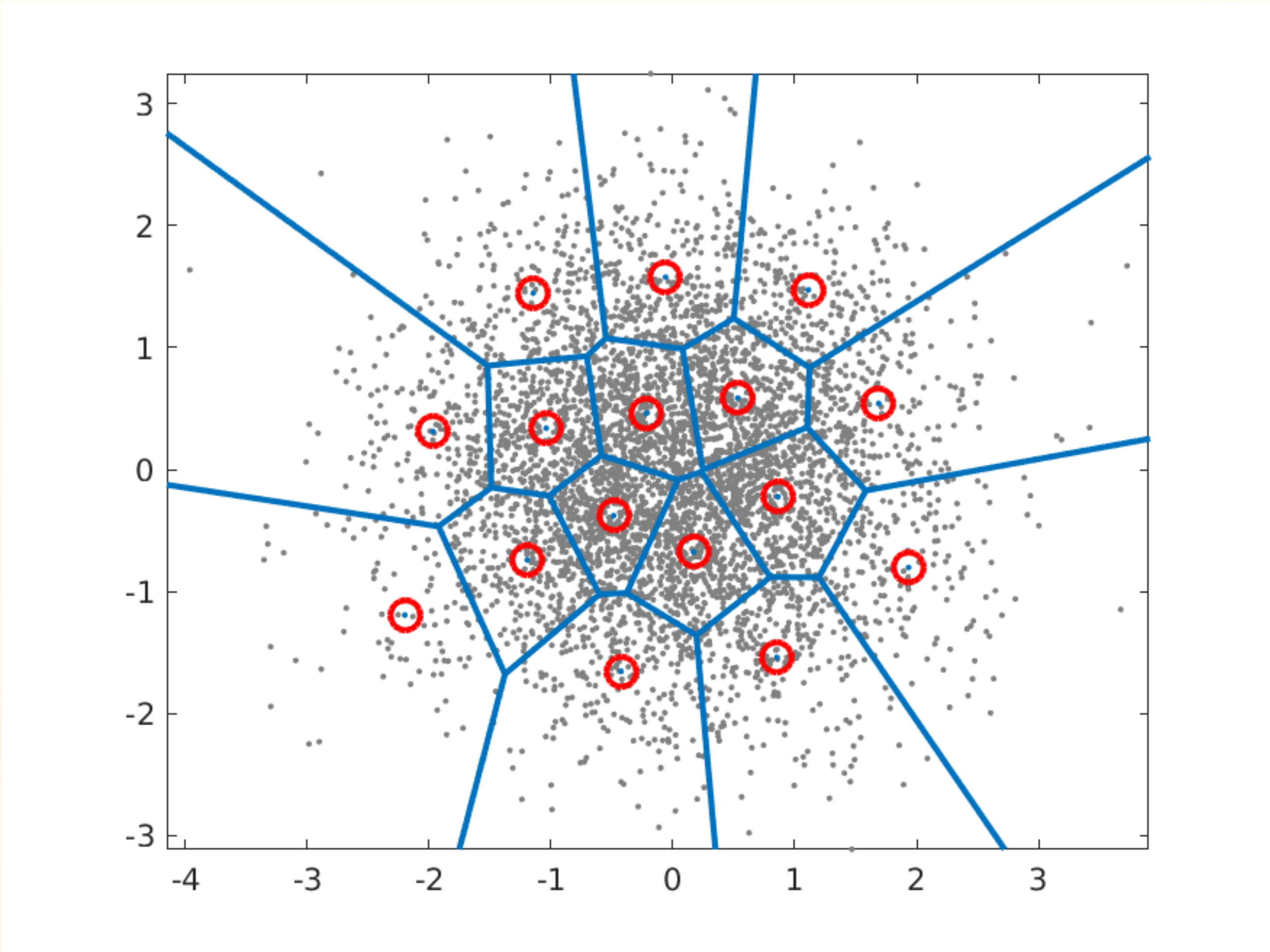


Figure adopted from *"Introduction to Speech Processing"*, Aalto University

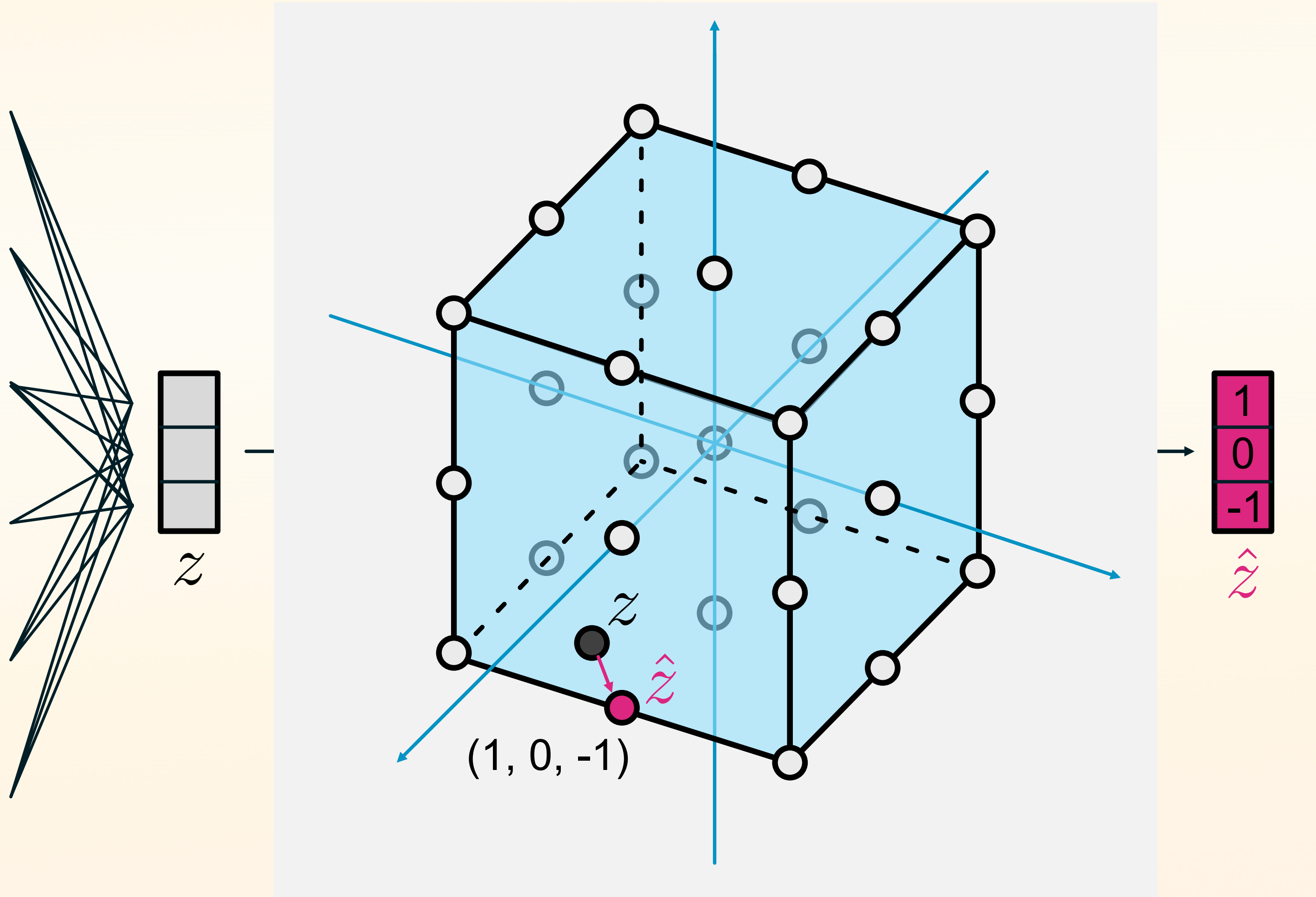


Figure adopted from *"Finite Scalar Quantization: VQ-VAE made simple"* by F. Mentzer et al

RESIDUAL FSQ

- We noticed a couple of interesting properties of FSQ.
  - Certain sets of levels are **purely supersets** of other sets of levels.
  - These sets can be combined with scaling to produce each other (with some caveats).
  - This property can be used to **decompose single FSQ** bottleneck into **residual version**, after training.

$$\ell_3 + \frac{\ell_3}{2} + \frac{\ell_3}{4} \supset \ell_9$$

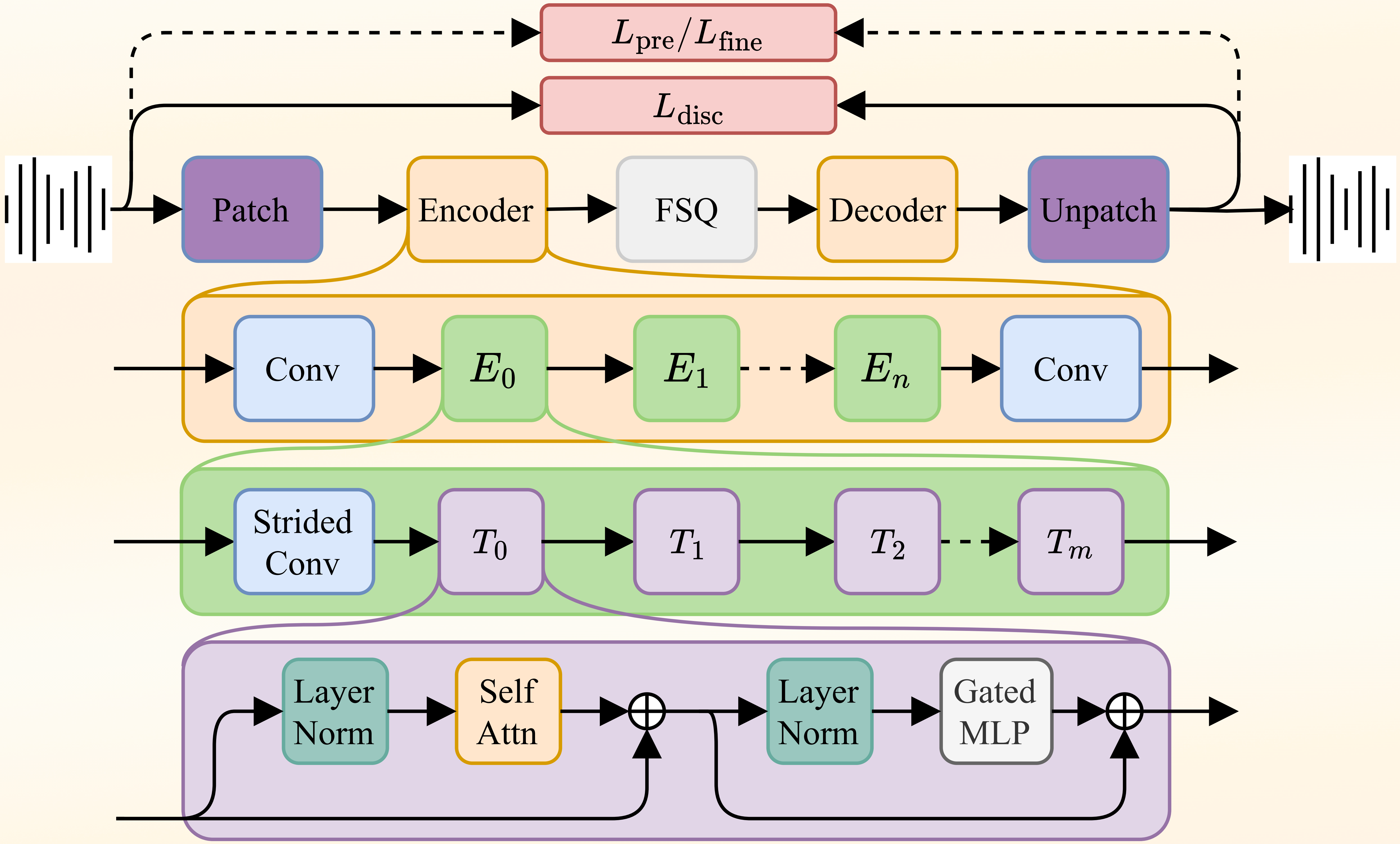
	Quantized Positions
ℓ_3	$\{-1, 0, 1\}$
ℓ_5	$\{-1, -0.5, 0, 0.5, 1\}$
ℓ_9	$\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$

Table 1: FSQ quantization points for level numbers conforming to $L = 2^n + 1$, $n \in \mathbb{Z}^+$, up to $n = 3$.

PUTTING EVERYTHING TOGETHER

ARCHITECTURE

- **Minimal** amount of **convolution**.
    - Needed to mitigate upper limit on patch size.
- Standard attention blocks with **RoPE** + **non-causal sliding window** mask.



DATA

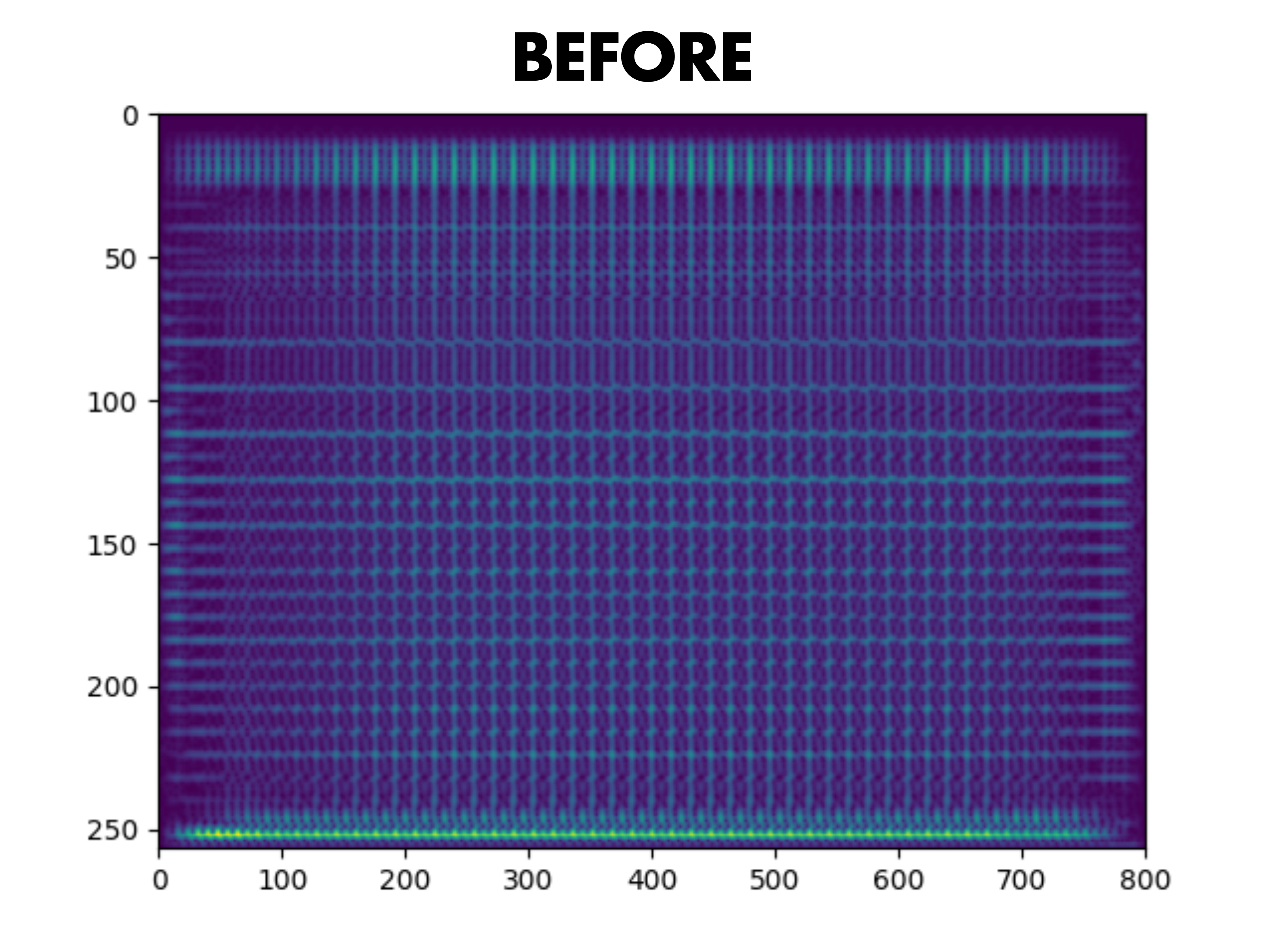
- Initially decided to train two variants of model - **speech (16kHz mono)** and **music & general audio (44.1kHz stereo)**.
- For speech, keeps things simple by focusing on **LibriLight**.
- For music + general audio, we can use the same dataset as Stable Audio Open - **Freesound + Free Music Archive**
- **Modest dataset sizes** in both cases.

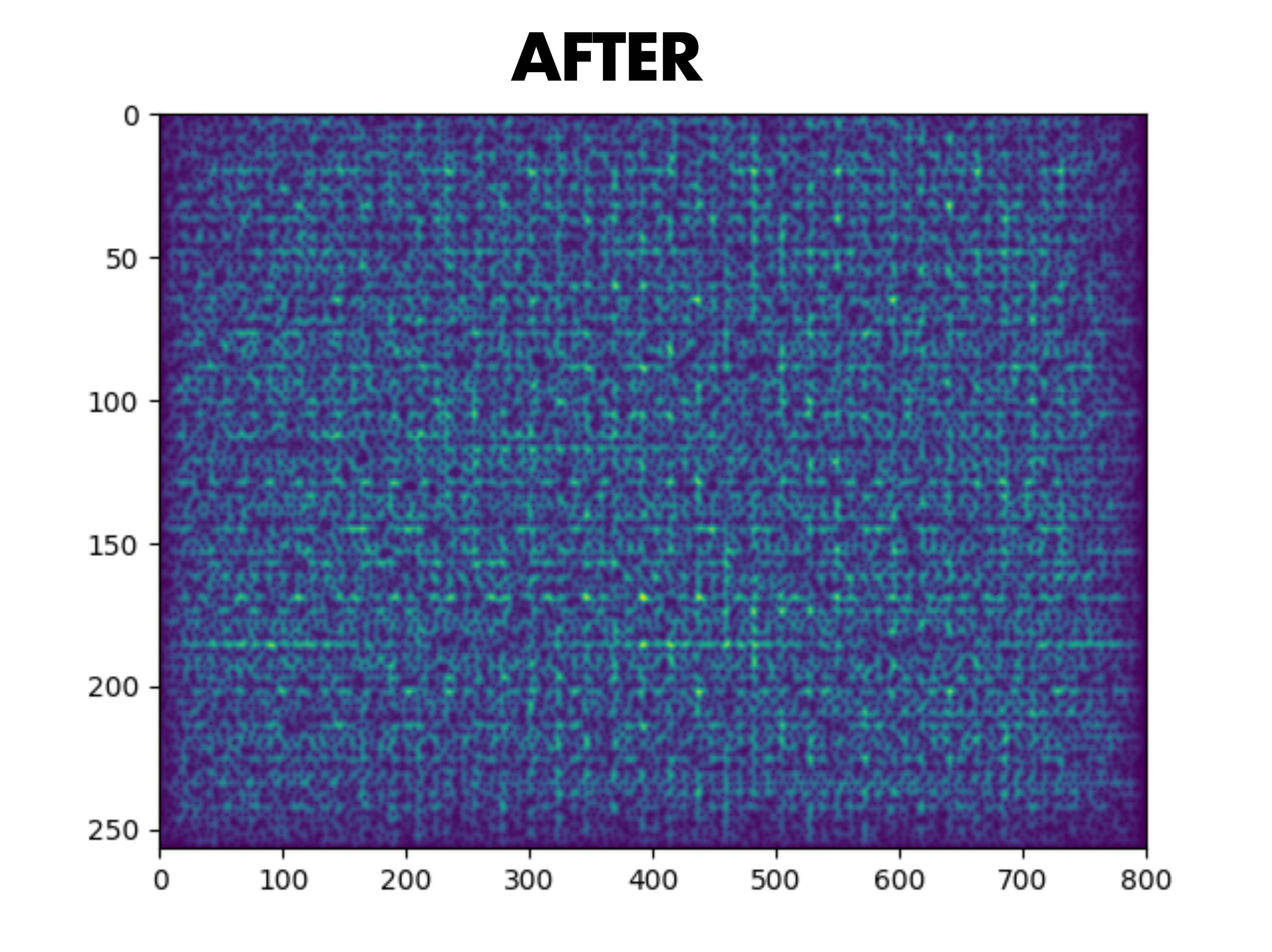
PROBLEMS

- Zero embedding issue
  - First trainings marred with **instability** unless we **aggressively stripped silence** from training data - **not practical**!
  - Traced issue to **LayerNorm** in transformer - can relax the epsilon to mitigate.
- Powerful transformer decoder likes to **over-fit on biases** introduced by loss functions.
  - **STFT loss produces periodic artefacts** - de-emphasise it
  - **Discriminator causes spotty artefacts** along predictable grid - examine discriminator for bias and de-emphasise adversarial component in favour of feature matching.

DISCRIMINATOR BIAS

- Seems to be present in basically all current discriminator archs (those with MPD are the worst)
- Can be partially mitigated by inharmonically spaced FFT sizes.





RESULTS OF INITIAL LARGE RUNS

- Speech **intelligibility not perfect**
  - **Audio quality very good**, but **rare phonemes dropped** or slurred
  - *Solution*: **Finetune** model with perceptual loss on decoder output using internal embeddings of **WavLM**

Model	SI-SDR ↑	Mel ↓	STFT ↓	PESQ ↑	STOI ↑
TAAE	4.73	0.86	1.26	3.09	0.92
w.o. perceptual loss	4.80	1.18	1.59	2.82	0.88

- Music version **too generative**
  - Musical version of intelligibility problem?
  - Audio quality is good, but not possible to evaluate in MUSHRA due to large differences (dropped instruments, changed timbre etc)
  - Drop for future work as we have no strong equivalent of WavLM for music.

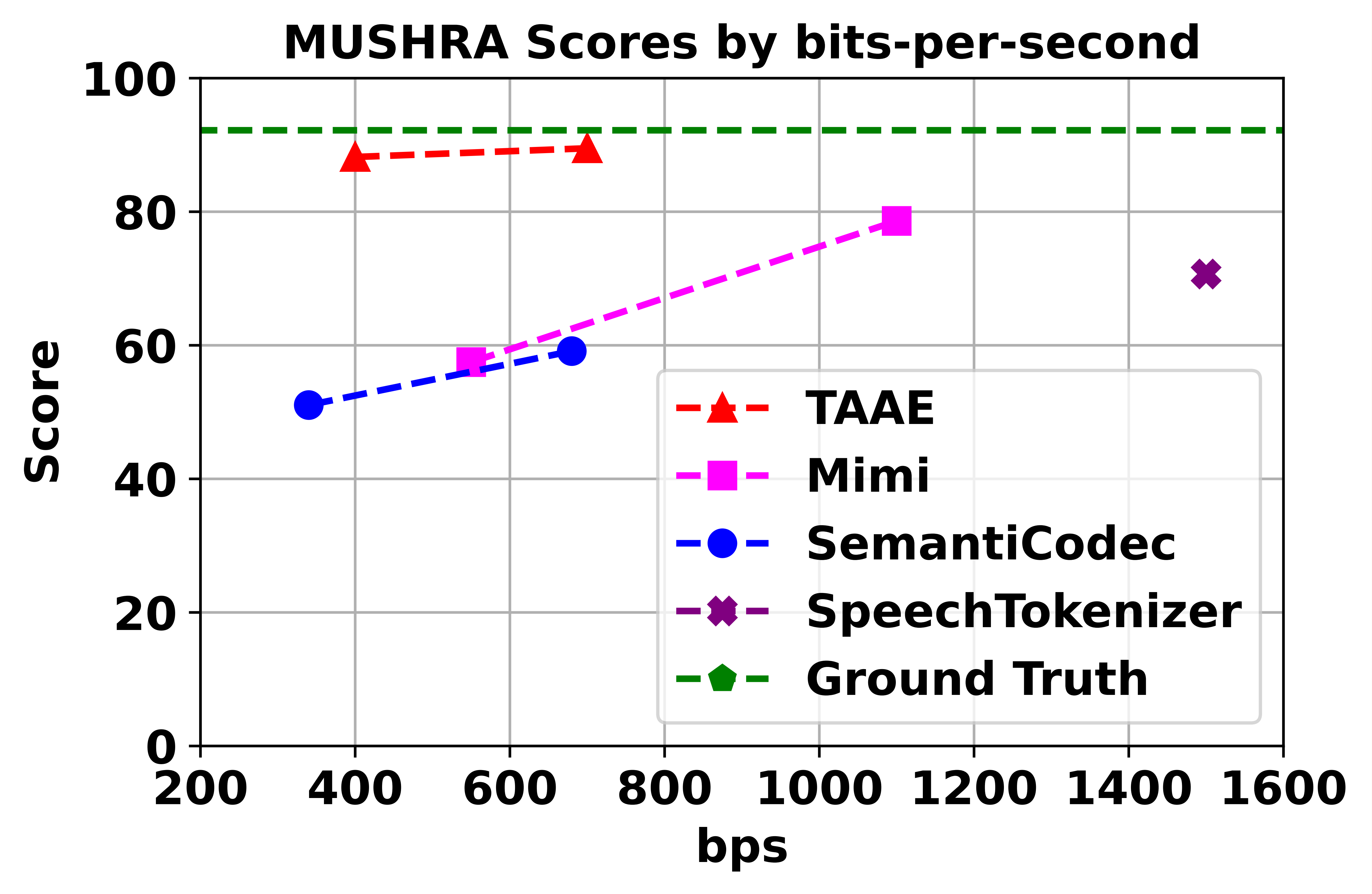
EVALUATION

OBJECTIVE METRICS

Model	BPS	TPF	TPS	SISDR ↑	Mel ↓	STFT ↓	PESQ ↑	STOI ↑	MOSNet ↑
DAC	1000	2	100	−6.51	1.49	1.76	1.64	0.75	2.77
	2000	4	200	−0.37	1.07	1.41	2.29	0.85	2.95
Encodec	1500	2	150	−0.22	1.14	1.49	2.36	0.85	2.87
	3000	4	300	2.77	0.95	1.33	2.84	0.90	2.98
SpeechTokenizer	1000	2	100	−3.30	1.06	1.37	2.41	0.85	2.94
	1500	3	150	−1.33	0.91	1.25	2.70	0.88	3.10
SemantiCodec	337.5	2	25	–	1.20	1.53	2.21	0.79	3.24
	675		50	–	0.98	1.32	2.65	0.86	3.29
Mimi	550	4	50	−4.45	1.19	1.55	2.48	0.85	3.11
	1100	8	100	2.20	0.94	1.31	3.01	0.90	3.24
TAAE	400	1	25	3.18	0.97	1.35	2.96	0.90	3.36
	700	2	50	4.73	0.86	1.26	3.09	0.92	3.36
+ no quant.	—	—	—	5.08	0.85	1.25	3.12	0.92	3.36

SUBJECTIVE TESTS

- **MUSHRA** methodology without anchor.
- Approx **25 participants**, mostly experts/ researchers.
- **Clear preference** for our model - very close to ground truth.
- Preference seems greater than expected from objective metrics - improvements in naturalness?



REAL **OURS** 0.4KBS **MIMI** 0.55KBS

SCALING

Param. count	SI-SDR ↑	Mel ↓	STFT ↓	PESQ ↑	STOI ↑
240M	3.52	1.24	1.67	2.74	0.87
540M	4.31	1.21	1.66	2.80	0.88
950M	4.80	1.18	1.59	2.82	0.88

- We repeated **pretraining** phase (no WavLM loss) at multiple parameter counts.
- Evidence for **improved reconstruction** with **larger parameter count** is clear.
- Our own later experiments, plus work of others, has shown that this type of architecture scales quite gracefully even to <100M params.

GENERALISATION

- Reviewers expressed **concern** about **English-only dataset** and possibility of overfitting.
- To test this, we **evaluated on Multilingual LibriSpeech**.
- Results show **decent generalisation** to other languages - matching some baselines which are trained on multilingual datasets.

REAL **OURS** 0.7KBS **MIMI** 1.1KBS

Model	BPS	SI-SDR ↑	Mel ↓	STFT ↓	PESQ ↑	STOI ↑
			Italian			
Encodec	1500	0.63	1.20	1.55	2.40	0.85
DAC	2000	−0.13	1.11	1.46	2.23	0.84
SemantiCodec	675	–	1.05	1.41	2.57	0.84
SpeechTokenizer	1000	−2.61	1.07	1.42	2.40	0.84
Mimi	1100	2.69	1.02	1.42	3.00	0.90
TAAE	700	4.54	0.99	1.38	2.89	0.89
			Polish			
Encodec	1500	1.39	1.12	1.49	2.42	0.86
DAC	2000	1.30	1.02	1.40	2.38	0.87
SemantiCodec	675	–	1.08	1.42	2.36	0.85
SpeechTokenizer	1000	−1.70	1.08	1.42	2.36	0.85
Mimi	1100	2.68	1.04	1.46	2.82	0.90
TAAE	700	4.45	0.95	1.36	2.66	0.89
			Dutch			
Encodec	1500	1.18	1.13	1.51	2.59	0.86
DAC	2000	1.30	0.98	1.36	2.55	0.87
SemantiCodec	675	–	1.09	1.42	2.34	0.83
SpeechTokenizer	1000	−5.01	1.09	1.42	2.34	0.83
Mimi	1100	2.84	0.98	1.39	3.01	0.90
TAAE	700	4.03	0.90	1.29	2.93	0.88
			French			
Encodec	1500	3.12	1.16	1.50	2.51	0.85
DAC	2000	2.68	0.98	1.34	2.41	0.87
SemantiCodec	675	–	1.02	1.36	2.54	0.83
SpeechTokenizer	1000	−0.50	1.04	1.36	2.38	0.84
Mimi	1100	4.61	0.98	1.38	2.98	0.89
TAAE	700	6.70	0.94	1.30	2.87	0.88
			Portuguese			
Encodec	1500	−0.46	1.18	1.56	2.49	0.84
DAC	2000	−1.05	1.07	1.44	2.35	0.84
SemantiCodec	675	–	1.04	1.42	2.59	0.83
SpeechTokenizer	1000	−4.15	1.07	1.42	2.43	0.83
Mimi	1100	1.45	0.98	1.42	3.04	0.89
TAAE	700	3.14	0.93	1.33	2.93	0.87
			German			
Encodec	1500	0.04	1.17	1.53	2.40	0.84
DAC	2000	−0.53	1.09	1.44	2.34	0.85
SemantiCodec	675	–	1.07	1.43	2.31	0.83
SpeechTokenizer	1000	−3.86	1.10	1.43	2.31	0.83
Mimi	1100	1.84	1.01	1.42	2.95	0.89
TAAE	700	4.94	0.92	1.32	2.83	0.88
			Spanish			
Encodec	1500	2.32	1.21	1.54	2.42	0.86
DAC	2000	1.93	1.04	1.39	2.36	0.86
SemantiCodec	675	–	1.04	1.39	2.52	0.84
SpeechTokenizer	1000	−0.84	1.07	1.42	2.43	0.85
Mimi	1100	3.82	1.07	1.44	2.93	0.90
TAAE	700	6.15	0.98	1.37	2.80	0.89

POST-PAPER WORK

TTS EXPERIMENTS / WEIGHTS RELEASE

- We wanted to release the model weights publicly for others to experiment with, so we made some tests with most common downstream task - **TTS**
  - Naive **LM** approach had **difficulty modelling token stream** well.
  - Some **precedent** with this in literature.
  - How can we improve this?
- Finetuned model further to regress **force-aligned phonemes** from bottleneck latents using **CTC** head.
  - **Significant improvement** for TTS, **slightly damages reconstruction metrics** - some reports in the wild of it damaging generalization.
  - Released two versions of model, `*stable-codec-speech-16k*` with CTC, and `*stable-codec-speech-16k-base*` without. Available on 🤗.

LIMITATIONS / LEARNINGS / FUTURE WORK

- Directly passing a real world signal into a transformer will always present difficulties.
- A very powerful decoder can present problems as well as advantages.
- How can this arch work properly with music (watch this space).
- How can we eliminate the last elements of convolution?
- Relatively small dataset + large param count means that there's still the possibility that existing model is overfit. Future work should scale up data.
- Is FSQ the best practical choice? Not sure. It's great for optimising reconstruction/bit, but might not be the most practical downstream.

QUESTIONS?