ETH zürich

VoiceCraft: Zero-Shot Speech Editing

Klejdi Sevdari
Seminar in Deep Neural Networks
Mentor: Luca Lanzendörfer

Oppenheimer: Great movie, terrible audio quality

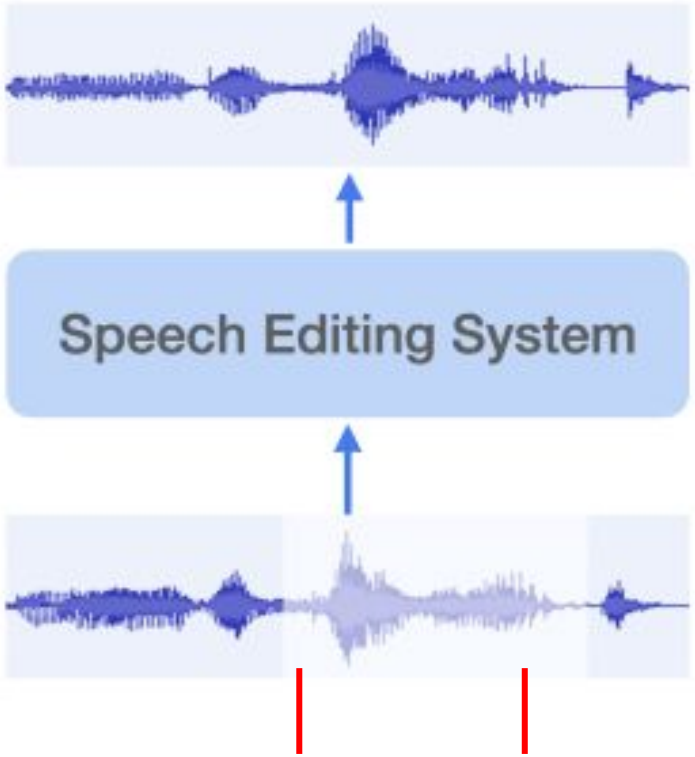
“I like to use the performance that was given in the moment rather than the actor revoice it later”

- Christopher Nolan

Sometimes the subtitles do not help…

Speech Editing

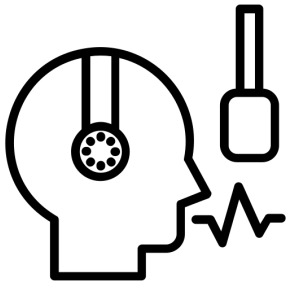
“Now I am become Death, the destroyer of worlds”



“Now I am become Death, the destruction of worlds”

Issues with rerecording:

background noise



subject availability

prosody mismatch

cost

Prosody refers to the patterns of stress and intonation in a language.

Naive Approach

loses almost all information
ideally we would never leave the audio domain

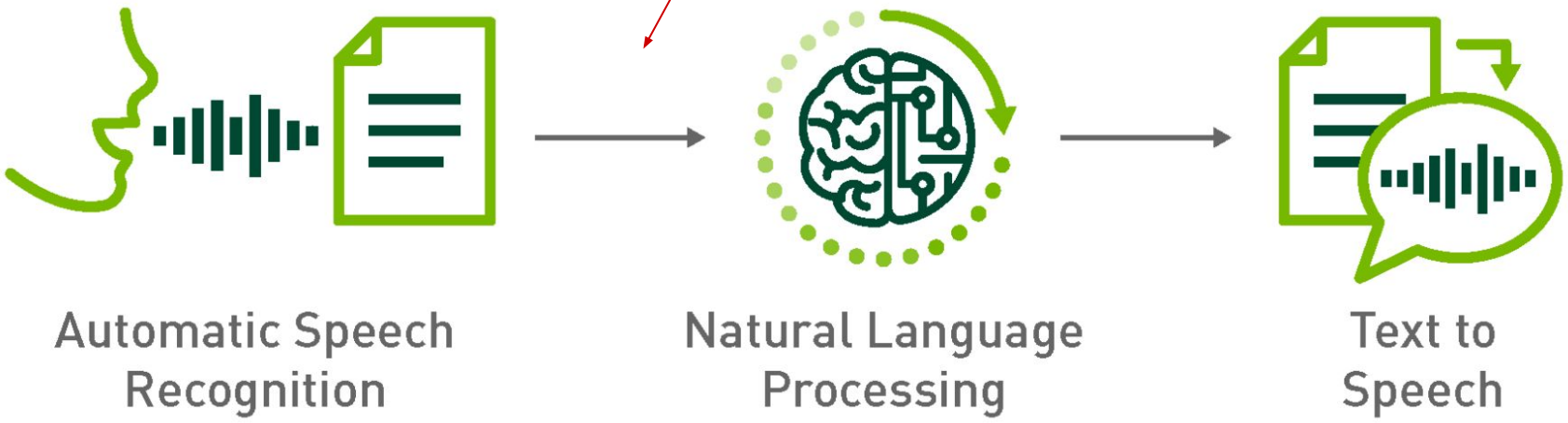
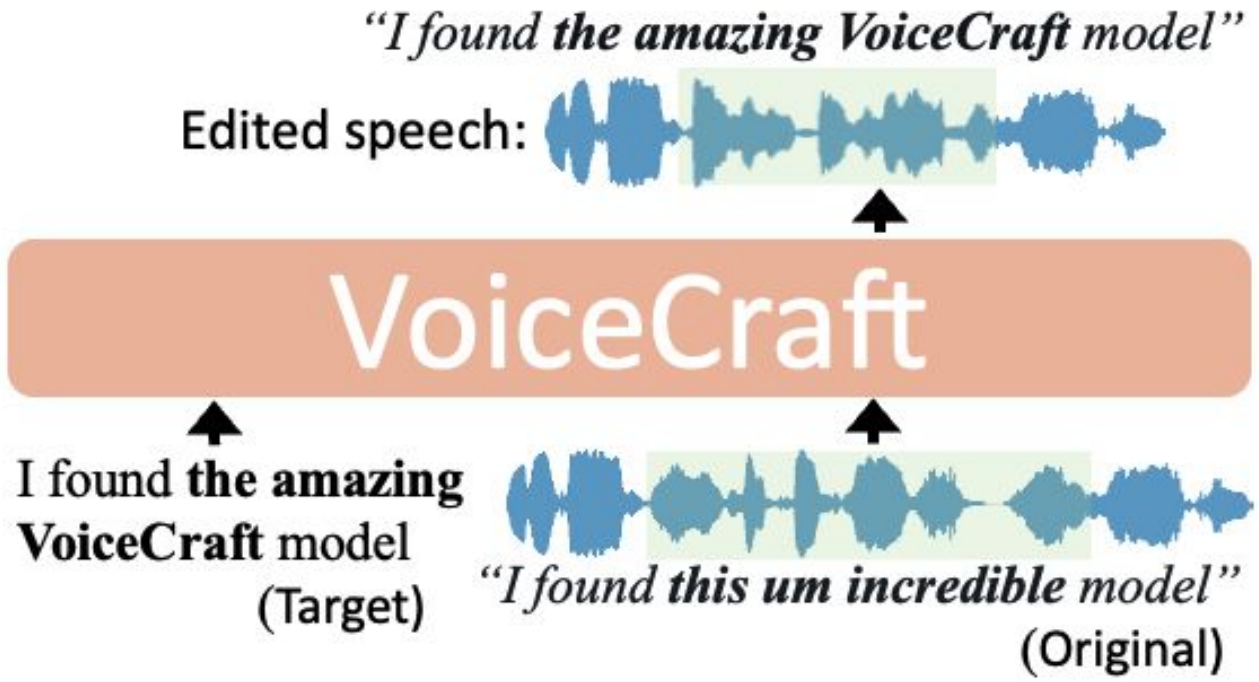


image source: https://developer.nvidia.com/blog/how-to-deploy-real-time-text-to-speech-applications-on-gpus-using-tensorrt/

VoiceCraft Model



https://arxiv.org/pdf/2403.16973

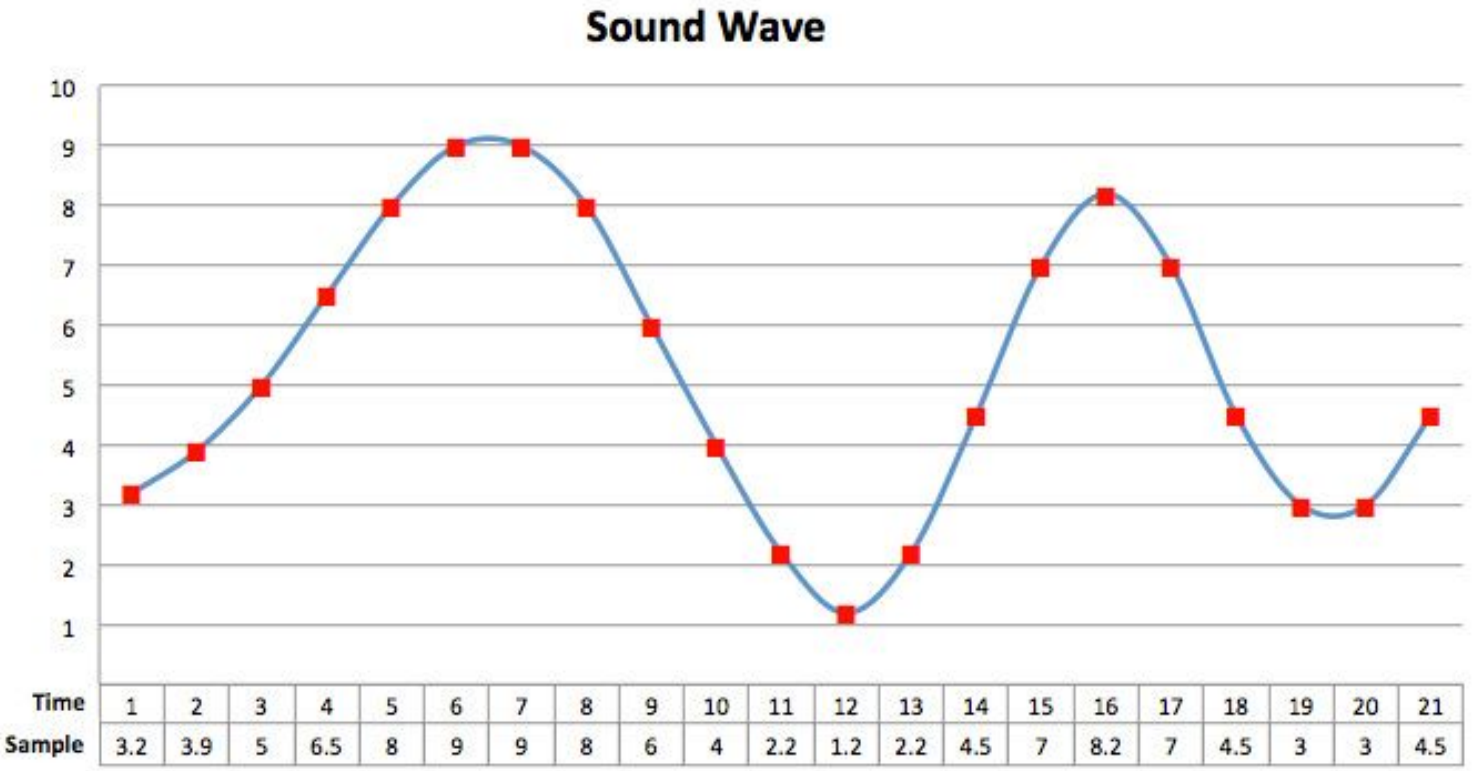
Demo

I saw it, I saw it, and it was beautiful. Who said that I didn't see it? Did Jim say that I didn't see it? He just envies me, I saw it!

The path of the righteous man is beset on all sides by the inequities of the selfish and the tyranny of evil men. Yet, he walks with unwavering resolve, guided by a compass of moral clarity and a heart unyielding in the face of adversity.

https://jasonppy.github.io/VoiceCraft_web/#toc

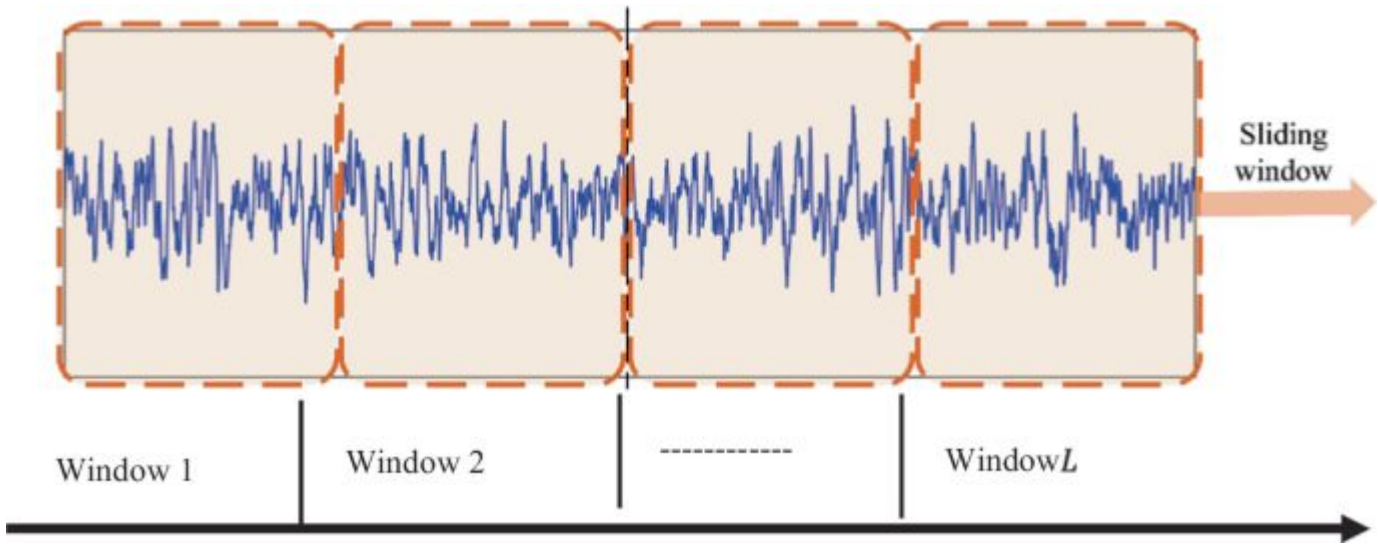
Representing the input



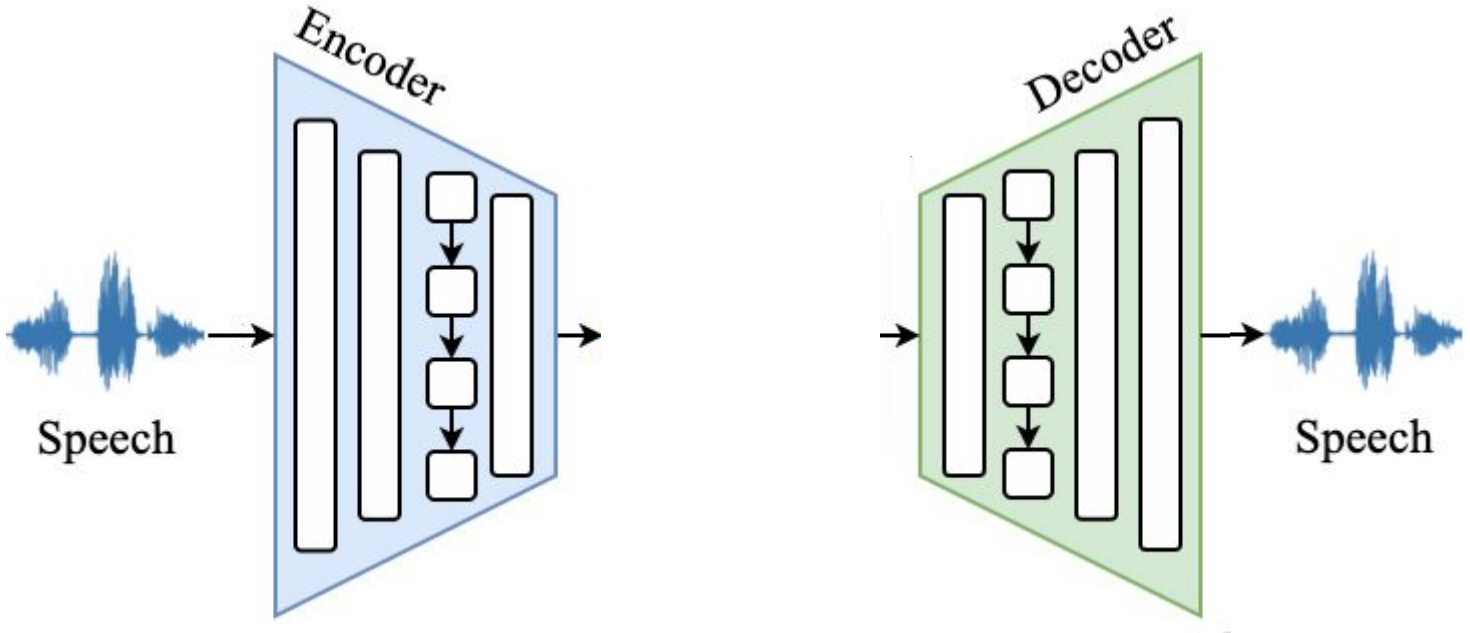
Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sample	3.2	3.9	5	6.5	8	9	9	8	6	4	2.2	1.2	2.2	4.5	7	8.2	7	4.5	3	3	4.5

At 16kHz, we need 16’000 numbers to represent 1 second of speech

Sliding Window

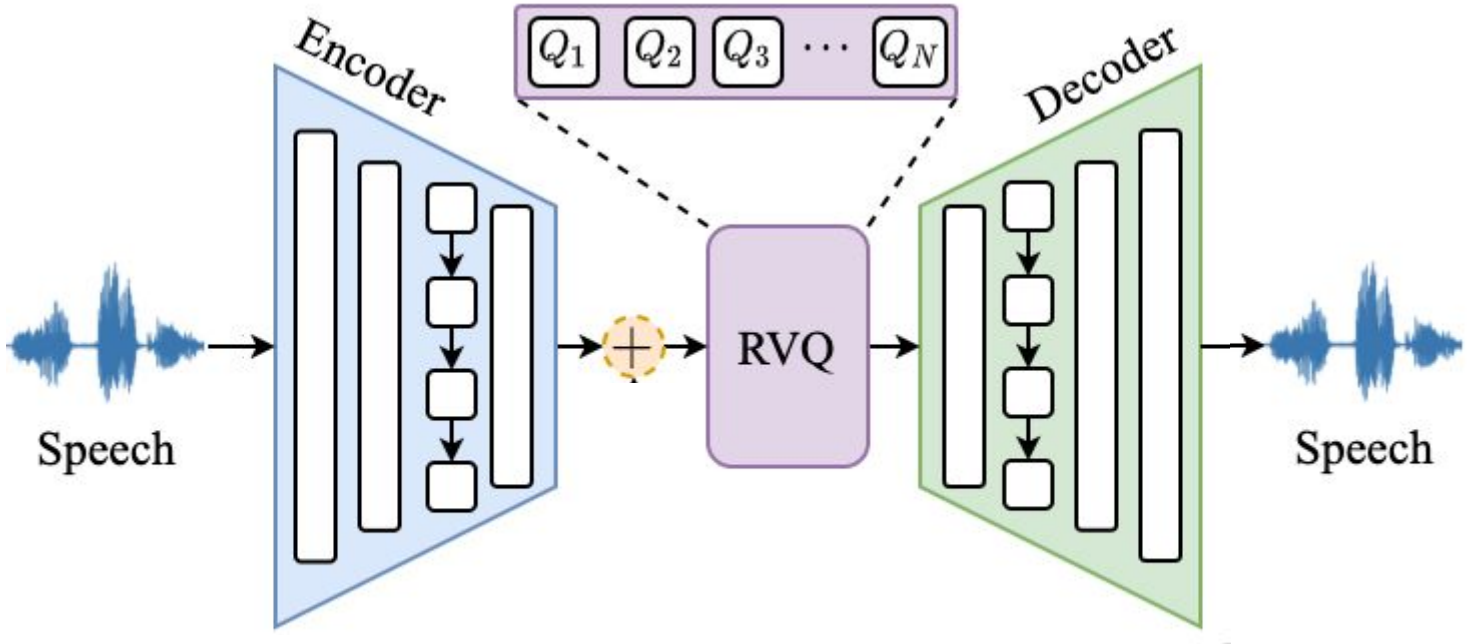


Encodec

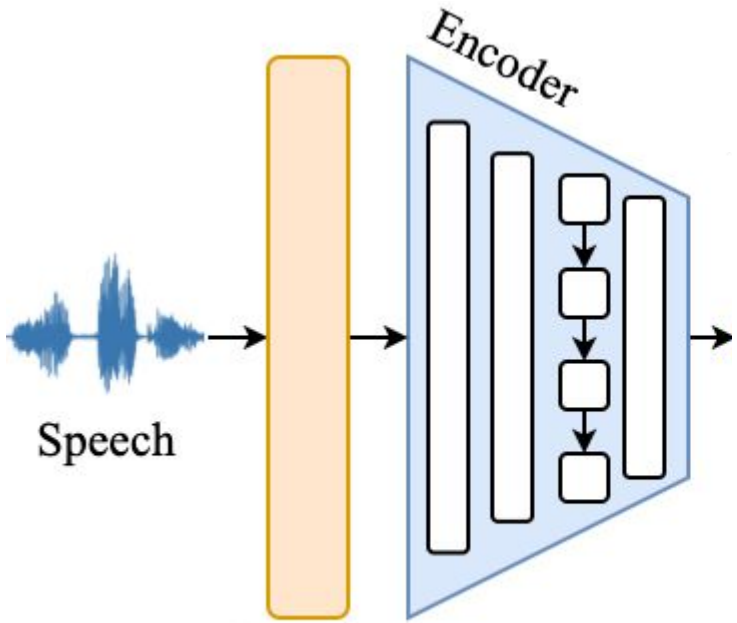


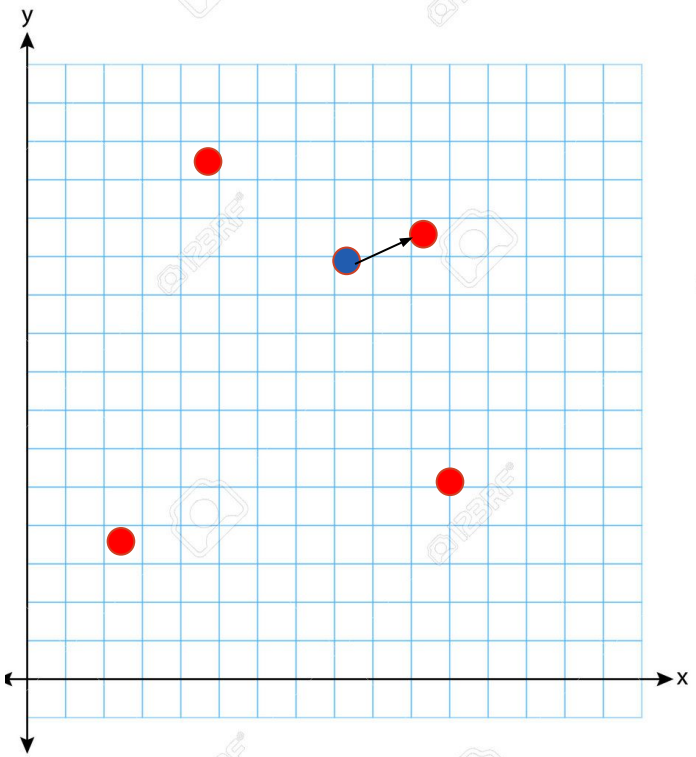
https://audiocraft.metademolab.com/encodec.html

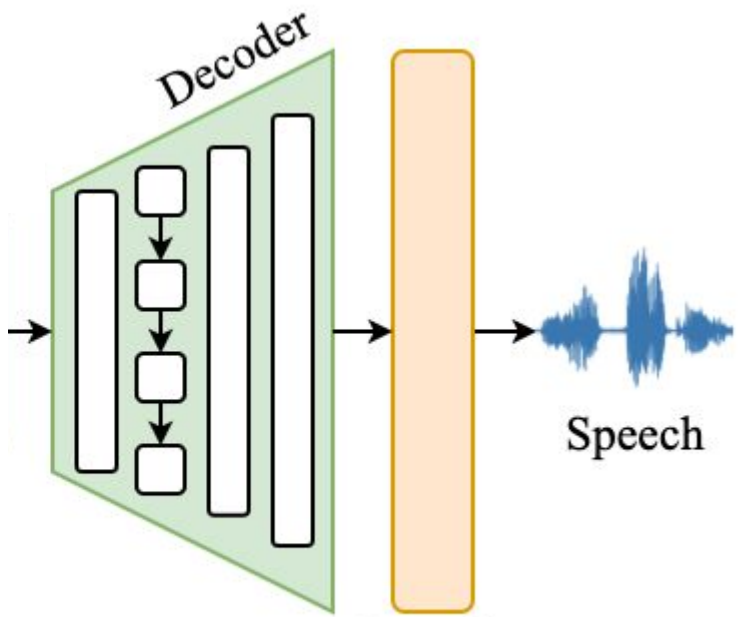
Encodec

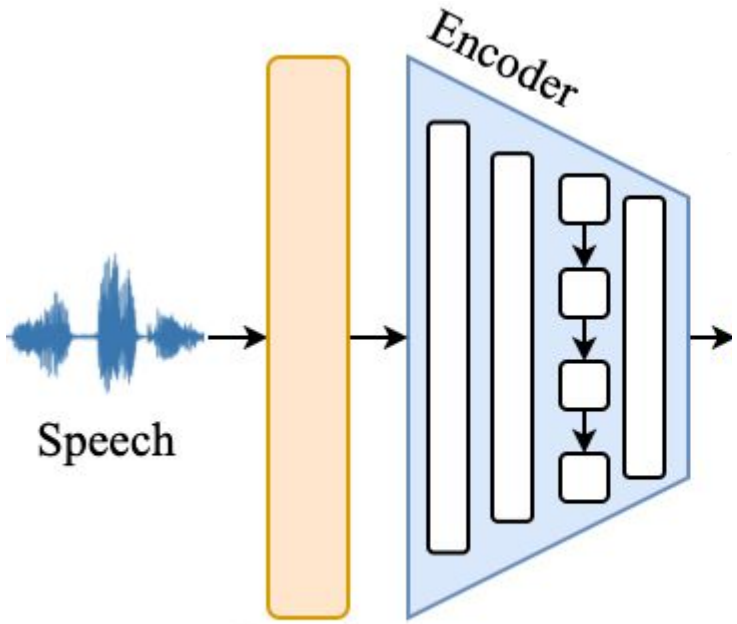


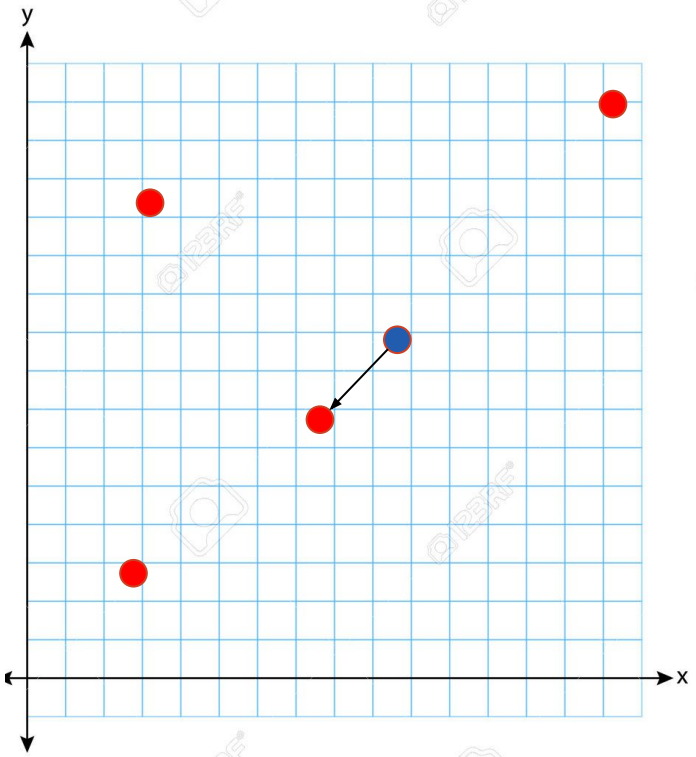
https://audiocraft.metademolab.com/encodec.html

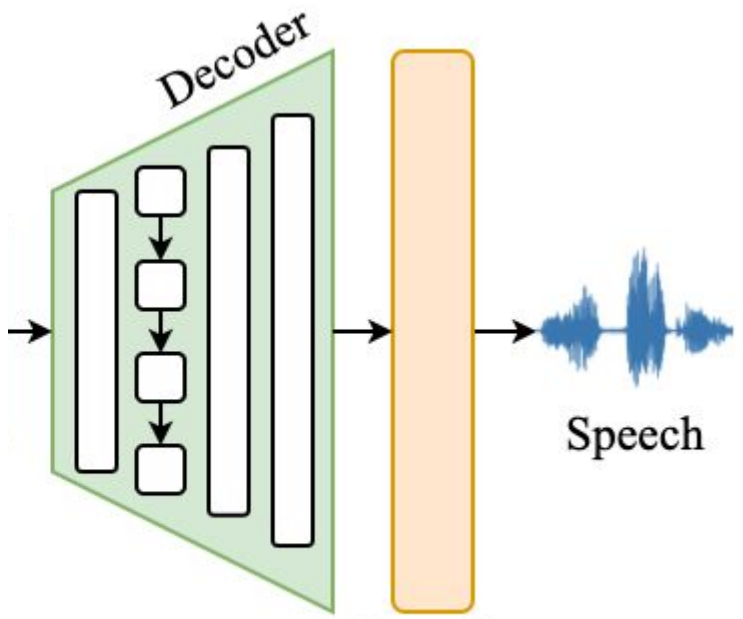

Encoder
Speech


y
x

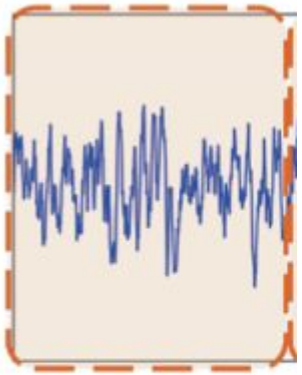

Decoder
Speech


Encoder
Speech


y
x


Decoder
Speech

Vector Quantization

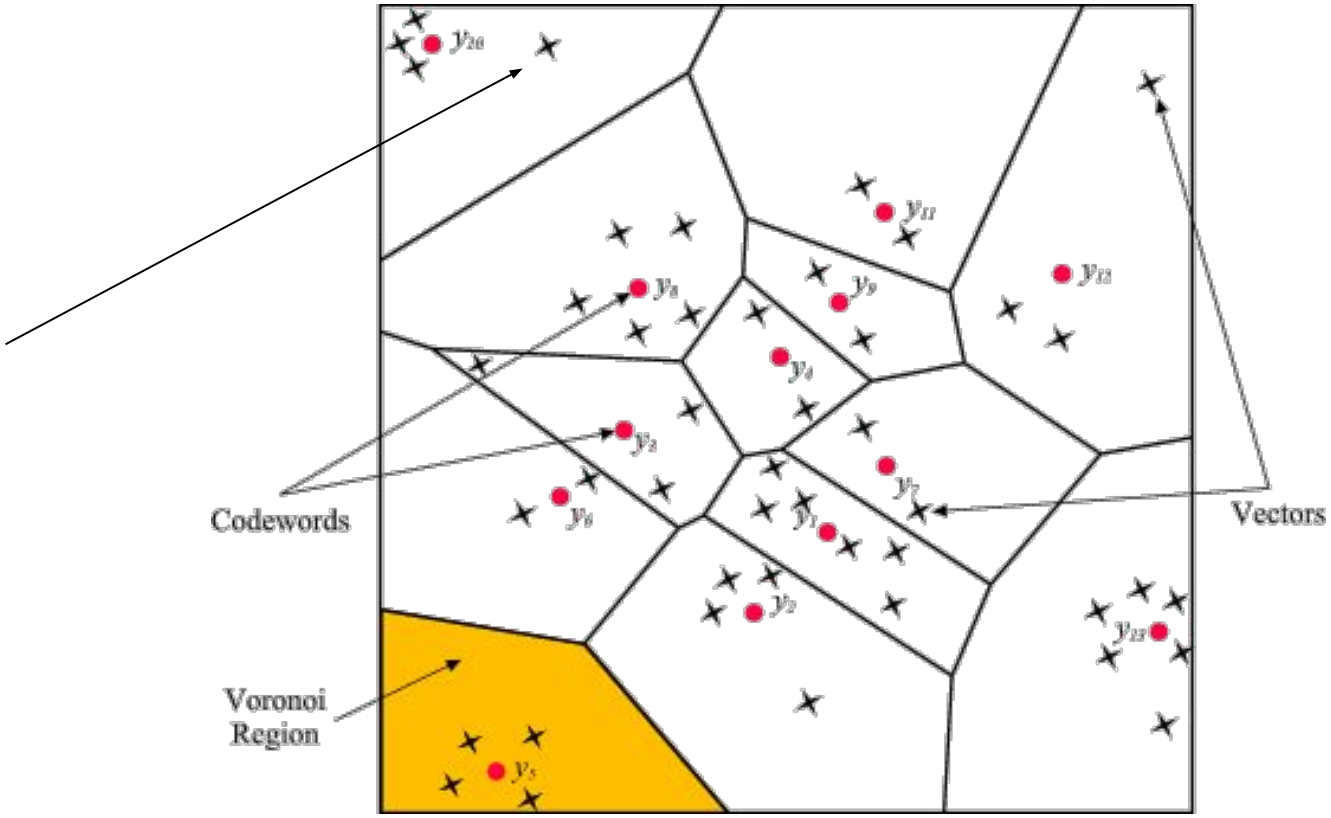


Window 1

0.5

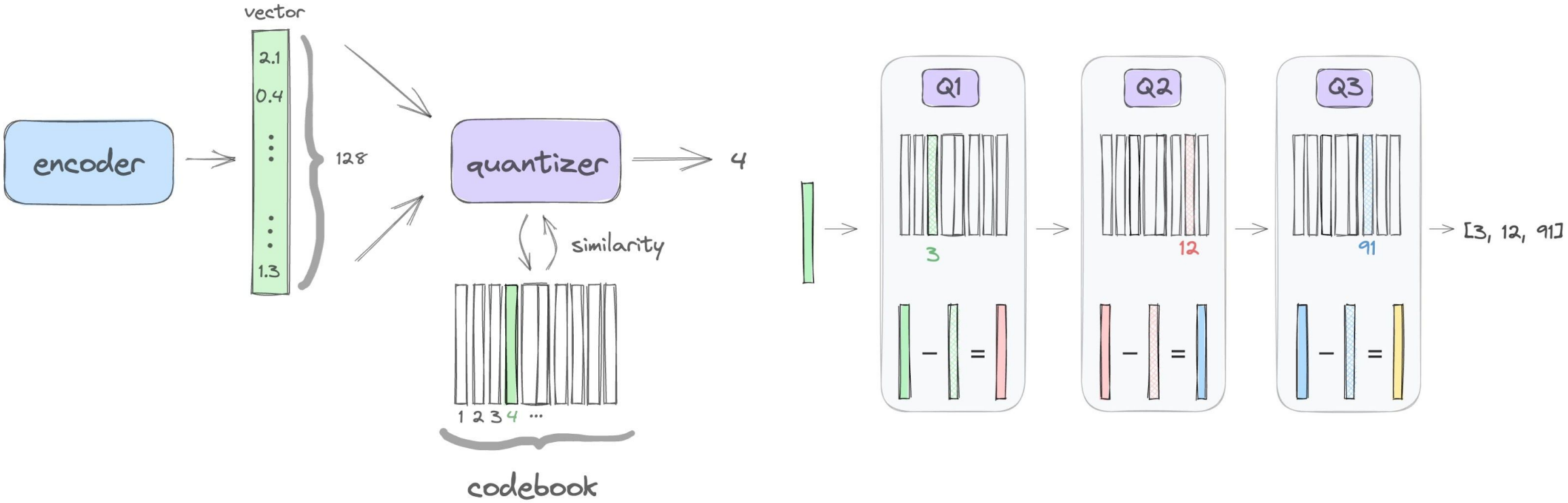
10

represented by the index of its codebook i.e 10



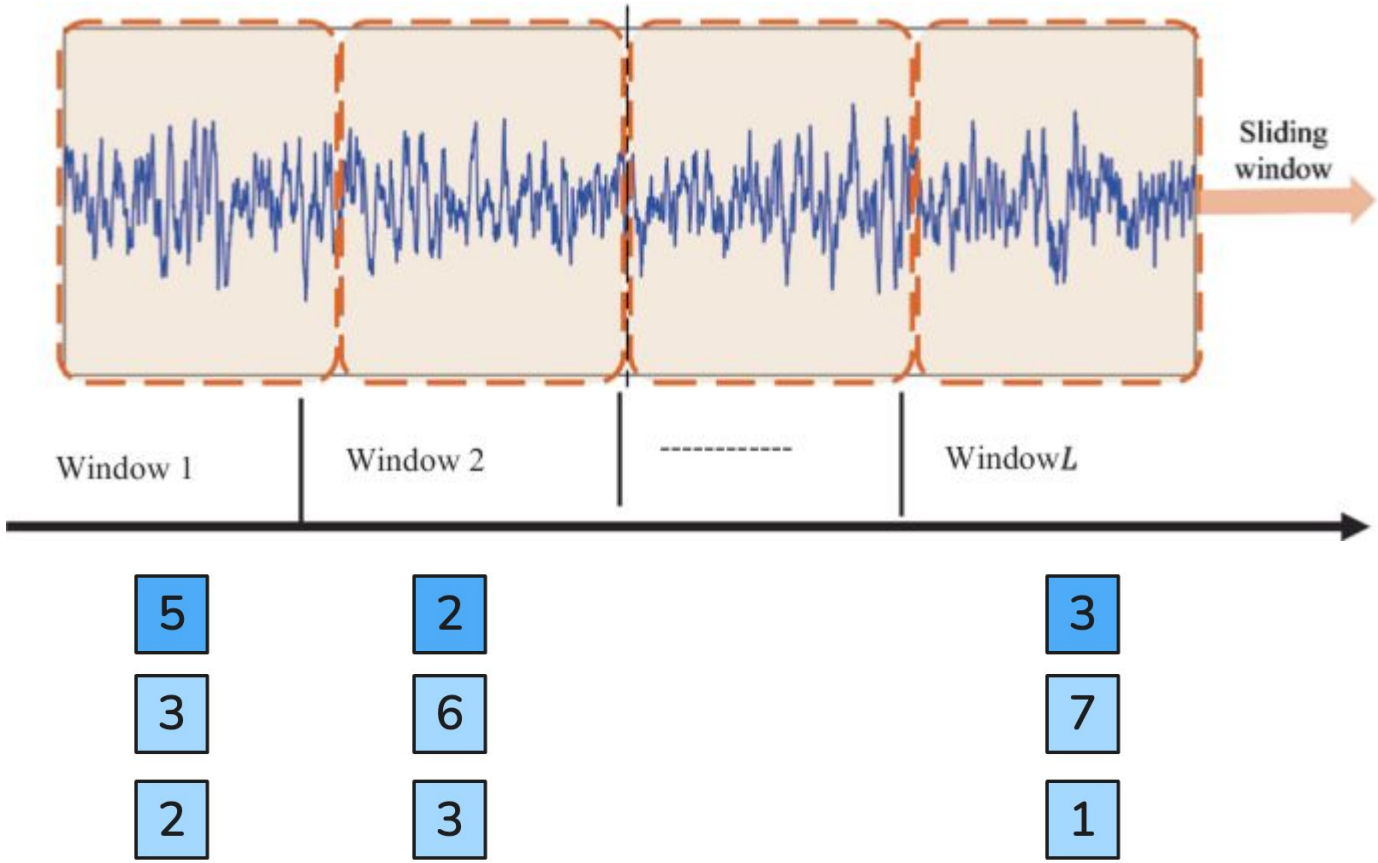
*AlgoLab PTSD trigger warning

Residual Vector Quantization

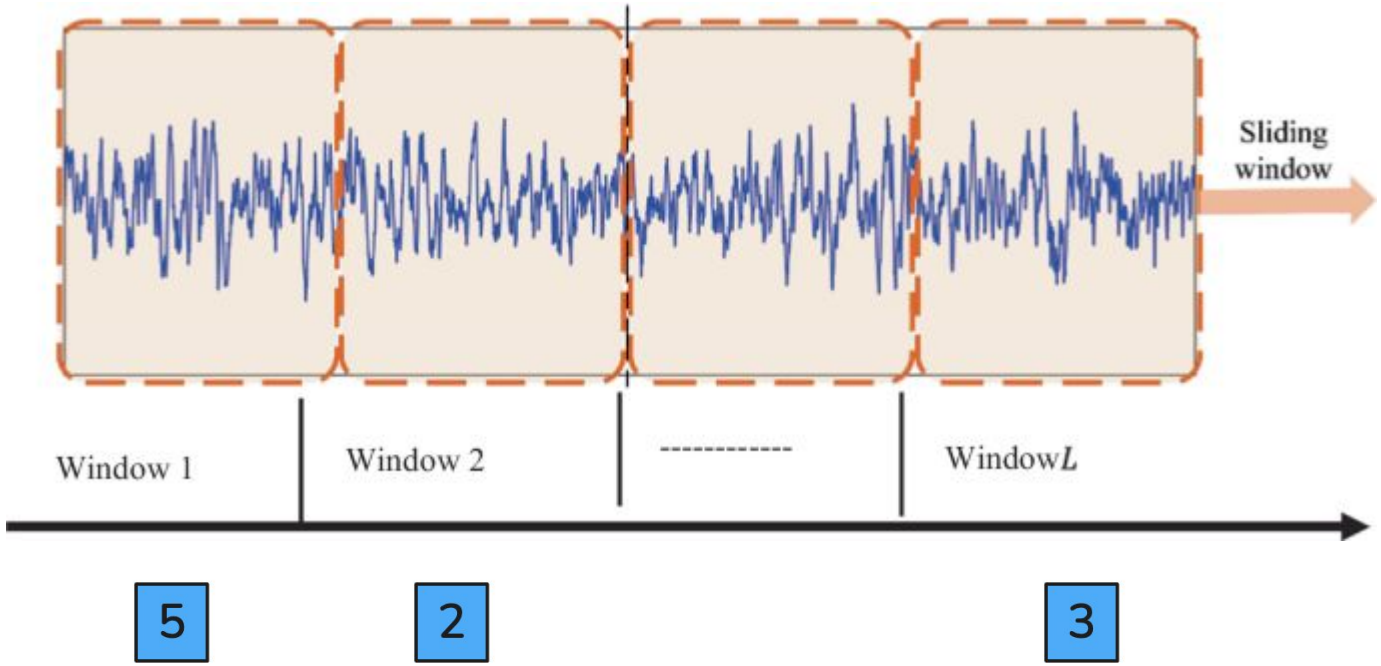


source: https://www.assemblyai.com/blog/what-is-residual-vector-quantization

Representation via Codebooks

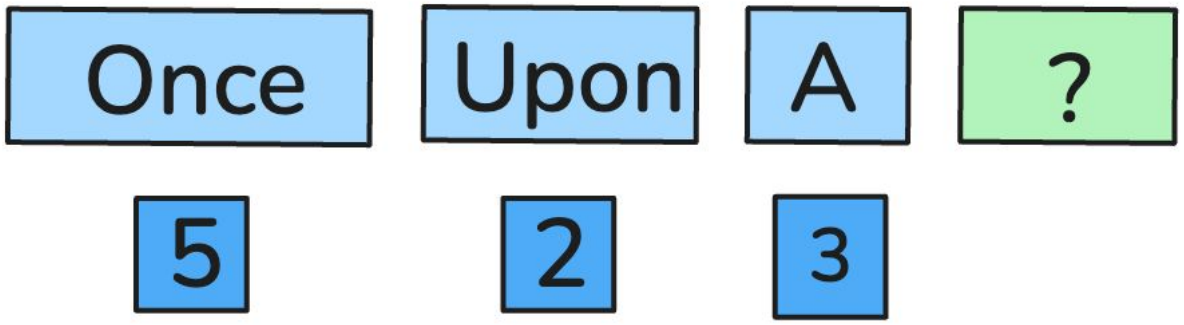


Or even simpler



Does this input representation remind you of something?

Autoregressive language models



Auto-regressive Language Models

$$P(X) = \prod_{t=1}^{T} P\left(x_t \mid x_1, \ldots, x_{t-1}\right)$$

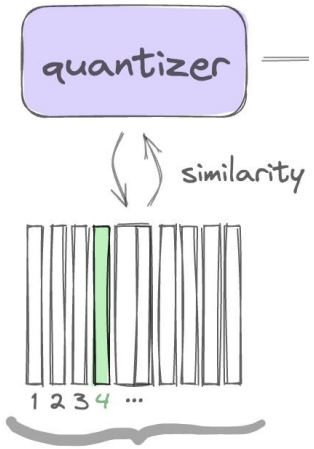
Next Token — Context

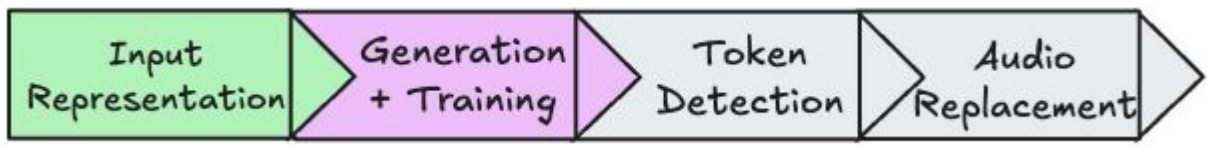
Decomposes sequence modeling into next-token modeling

→ We got audio generation for “free”

Now that we have a shared representation, we can borrow techniques from LLMs

Roadmap



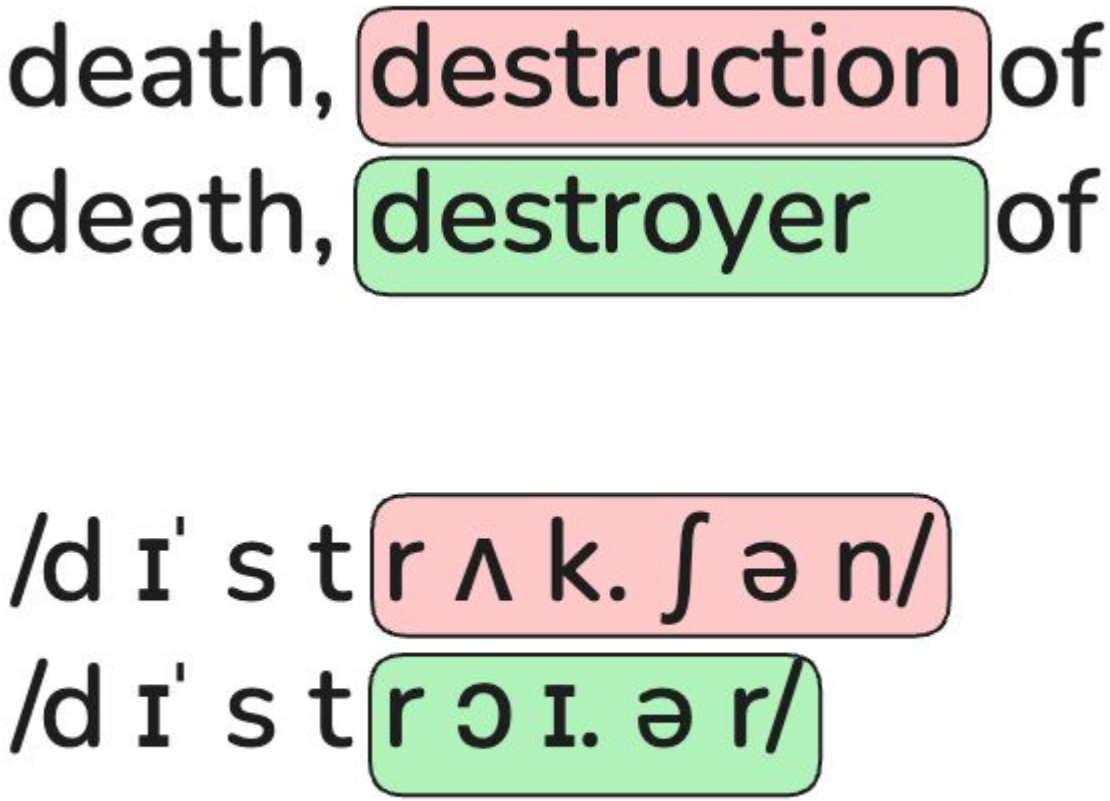


$$P(X) = \prod_{t=1}^{T} P\left(x_t \mid x_1, \ldots, x_{t-1}\right)$$

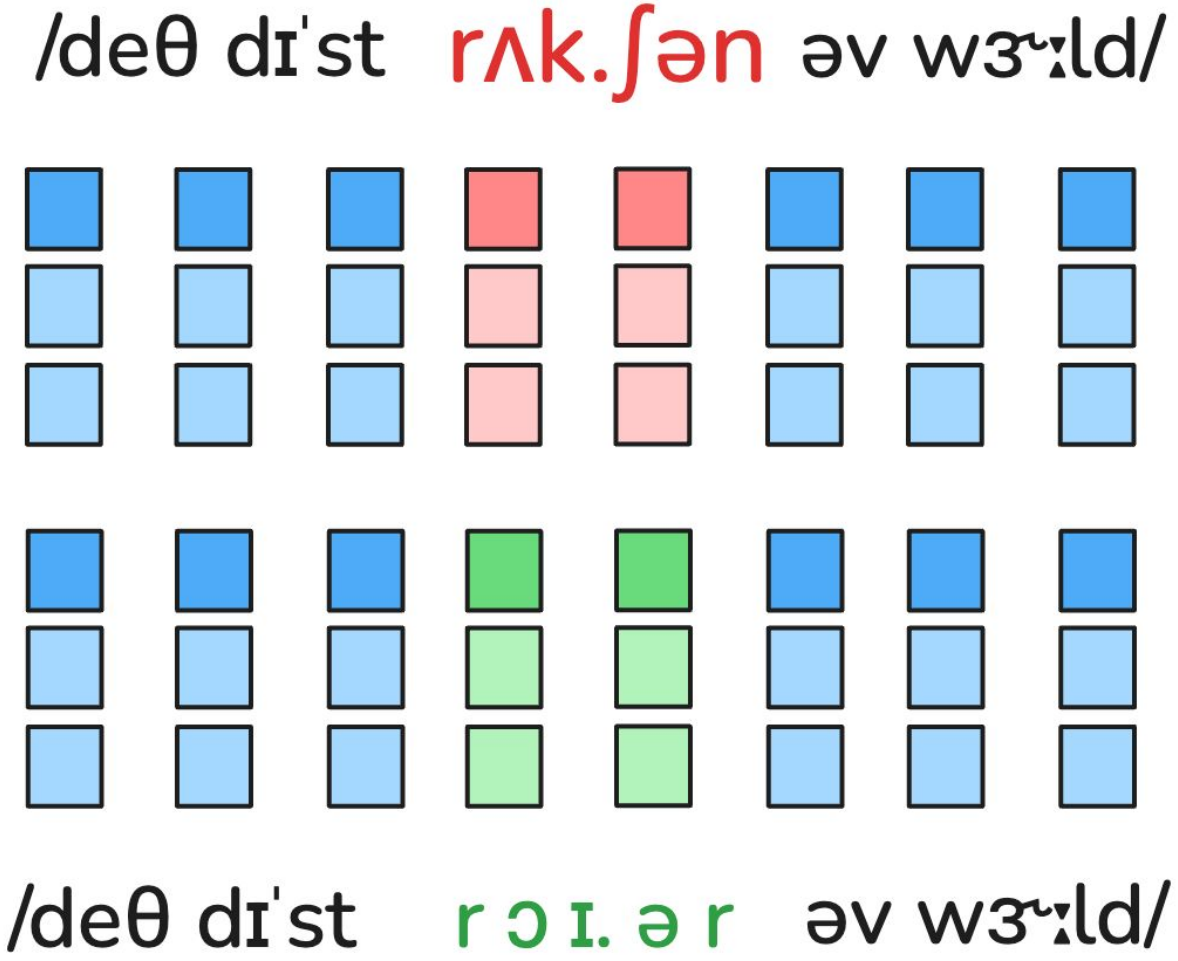
Next Token Context

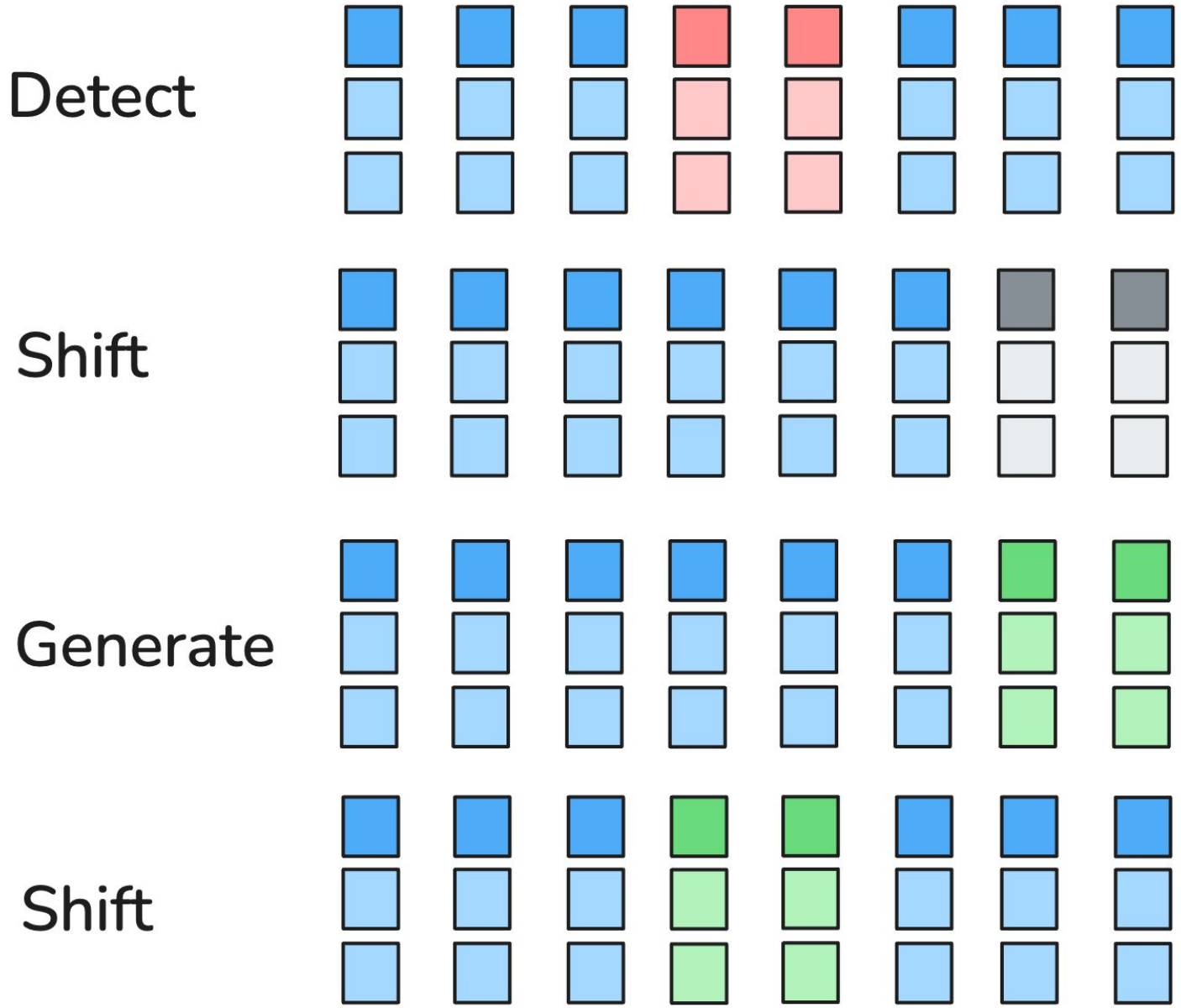
Decomposes sequence modeling into next-token modeling

Detecting what part of the input needs to change



Token replacement




Detect
Shift
Generate
Shift

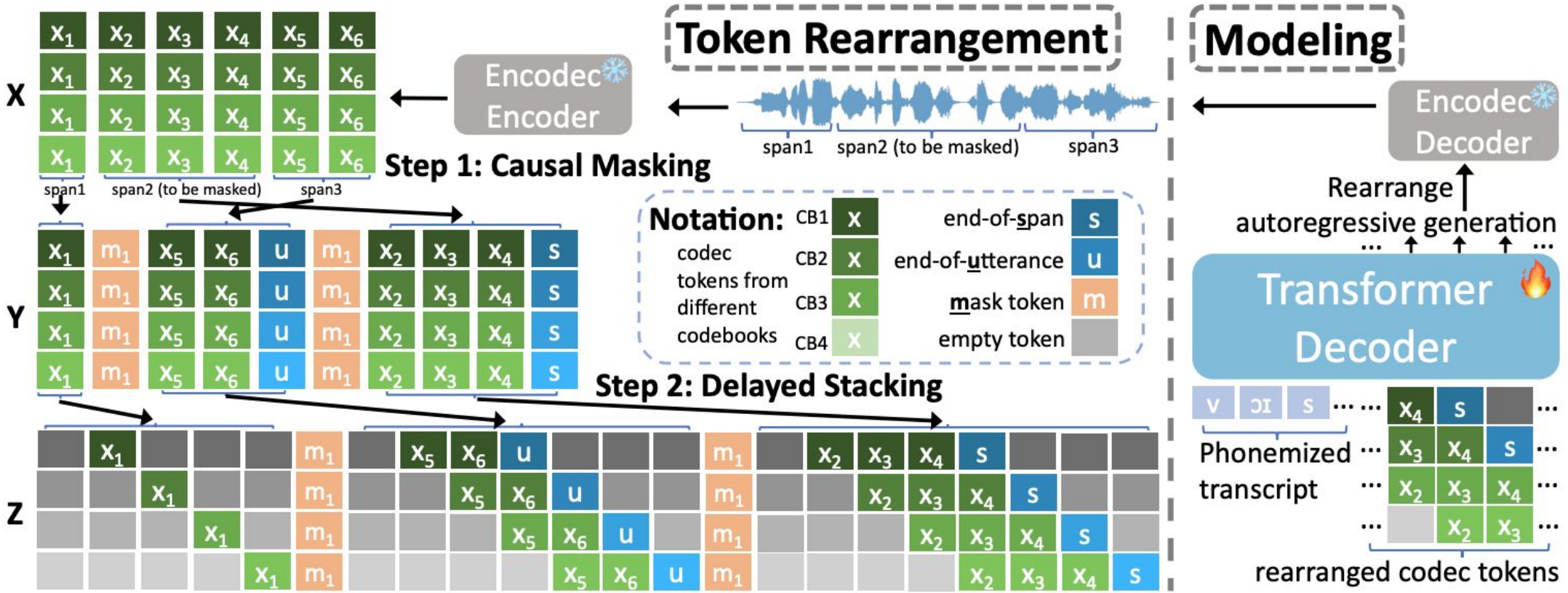


Figure 2: An example of the token rearrangement procedure and modeling framework. The rearrangement procedure involves two steps: (1) *Causal masking*, where masked spans are replaced with mask tokens and moved to the end, and (2) *Delayed stacking*, where tokens are shifted in the time dimension based on their codebook index.

Note that the prediction is conditioned on transcript W, and all tokens in Z before $Z_{s,t}$, denoted as $H_{s,t}$. Mathematically, the Transformer decoder models the factorized conditional distribution of Z:

$$\mathbb{P}_\theta(Z|W) = \prod_s \prod_t \mathbb{P}_\theta(Z_{s,t}|W, H_{s,t}) \tag{2}$$

$$= \prod_s \prod_t \prod_{k=1}^{K} \mathbb{P}_\theta(Z_{s,t,k}|W, H_{s,t}) \tag{3}$$

Results on Voice Editing

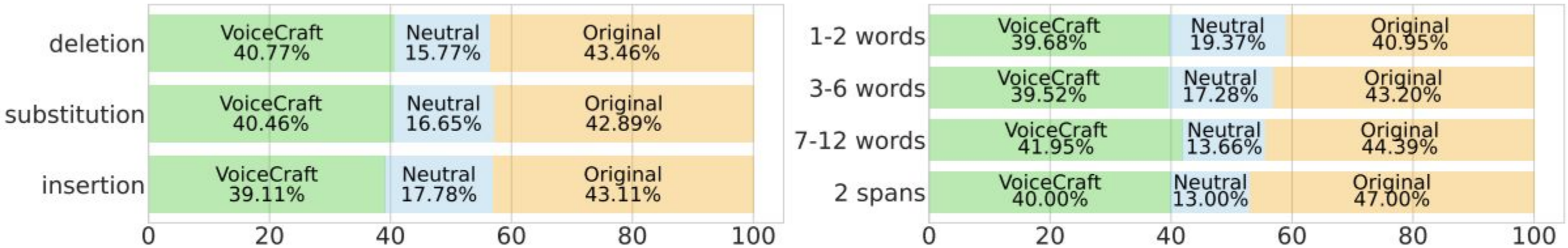
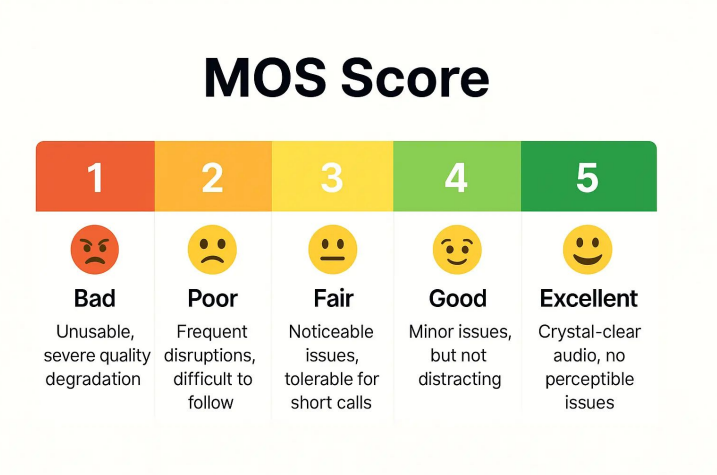


Figure 3: Breakdown of side-by-side human preference on naturalness comparing VOICECRAFT edited speech and the original speech. Grouped by edit type (left) and edit span length (right).

Results on Text to Speech

Table 6: On the zero-shot TTS task, comparing VOICECRAFT with other models.

			Intelligibility MOS			Naturalness MOS			Speaker Similarity MOS		
Model	WER	SIM	Libri.	YouTube	Total	Libri.	YouTube	Total	Libri.	YouTube	Total
YourTTS	6.6	0.41	$3.28_{\pm 0.11}$	$3.01_{\pm 0.12}$	$3.14_{\pm 0.08}$	$2.99_{\pm 0.12}$	$2.59_{\pm 0.12}$	$2.79_{\pm 0.08}$	$3.10_{\pm 0.12}$	$2.49_{\pm 0.12}$	$2.79_{\pm 0.09}$
FluentSpeech	**3.5**	0.47	$3.70_{\pm 0.11}$	$3.65_{\pm 0.12}$	$3.67_{\pm 0.08}$	$3.34_{\pm 0.11}$	$3.43_{\pm 0.12}$	$3.38_{\pm 0.08}$	$4.10_{\pm 0.09}$	$3.92_{\pm 0.11}$	$4.01_{\pm 0.07}$
VALL-E	7.1	0.50	$4.05_{\pm 0.09}$	$3.94_{\pm 0.10}$	$4.00_{\pm 0.07}$	$3.85_{\pm 0.10}$	$3.86_{\pm 0.10}$	$3.86_{\pm 0.07}$	$4.12_{\pm 0.10}$	$4.02_{\pm 0.10}$	$4.07_{\pm 0.07}$
XTTS v2	3.6	0.47	$4.29_{\pm 0.09}$	$3.97_{\pm 0.10}$	$4.13_{\pm 0.07}$	$4.02_{\pm 0.09}$	$3.90_{\pm 0.10}$	$3.96_{\pm 0.07}$	$3.64_{\pm 0.12}$	$3.25_{\pm 0.12}$	$3.44_{\pm 0.08}$
VOICECRAFT	4.5	**0.55**	$\mathbf{4.38}_{\pm 0.08}$	$\mathbf{4.08}_{\pm 0.10}$	$\mathbf{4.23}_{\pm 0.06}$	$\mathbf{4.16}_{\pm 0.08}$	$\mathbf{4.18}_{\pm 0.09}$	$\mathbf{4.17}_{\pm 0.06}$	$\mathbf{4.35}_{\pm 0.08}$	$\mathbf{4.33}_{\pm 0.09}$	$\mathbf{4.34}_{\pm 0.06}$
Ground Truth	3.8	0.76	$4.37_{\pm 0.08}$	$4.42_{\pm 0.08}$	$4.39_{\pm 0.06}$	$4.32_{\pm 0.08}$	$4.64_{\pm 0.06}$	$4.48_{\pm 0.05}$	$4.26_{\pm 0.10}$	$4.62_{\pm 0.08}$	$4.44_{\pm 0.06}$



Ethical Considerations

US NEWS

Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms

By Joshua Rhett Miller

Published March 17, 2022 | Updated March 17, 2022, 1:06 p.m. ET

My Opinion

Strengths

- Token rearrangement - very clever idea.
- Audio quality is impressive and also backed by human tests.
- Everything is open source including the dataset the authors created to train the model.

Limitations

- Watermark mentioned but not addressed.
- Degenerate looping - generate 10 times, discard the 4 longest ones and pick randomly from the remaining 6.

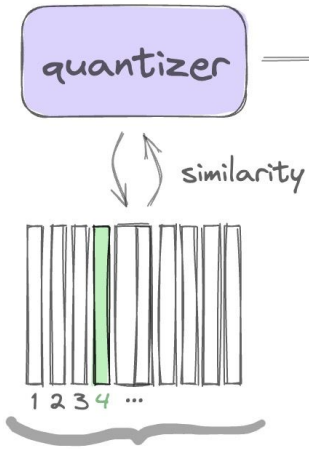
ETH zürich

Thank you for your attention!

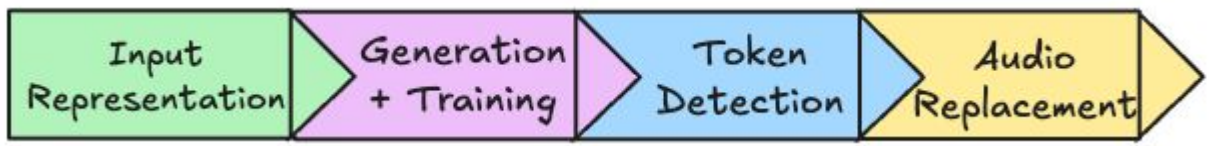
Questions?

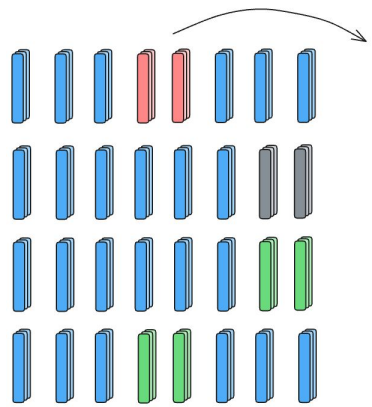
Klejdi Sevdari
MSc Computer Science
ksevdari@ethz.ch

TLDR



/d ɪ ˈ s t r ʌ k. ʃ ə n/
/d ɪ ˈ s t r ɔ ɪ. ə r/



$$P(X) = \prod_{t=1}^{T} P\left(x_t \mid x_1, \ldots, x_{t-1}\right)$$

Next Token Context