
Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

**Chengyi Wang* Sanyuan Chen* Yu Wu* Ziqiang Zhang Long Zhou Shujie Liu
Zhuo Chen Yanqing Liu Huaming Wang Jinyu Li Lei He Sheng Zhao Furu Wei**
Microsoft

<https://github.com/microsoft/unilm>

a.k.a VALL-E

Interpreted by:

John Tan Chong Min





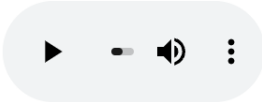
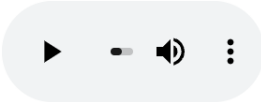

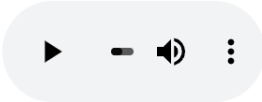
Aim

- To have a flexible generation pipeline to perform various acoustic speech generation based on a single acoustic input
 - Transformer-style learning
- To generate high quality speech
 - Good quality representation via Encodec

Let's hear it

- Some samples from the website

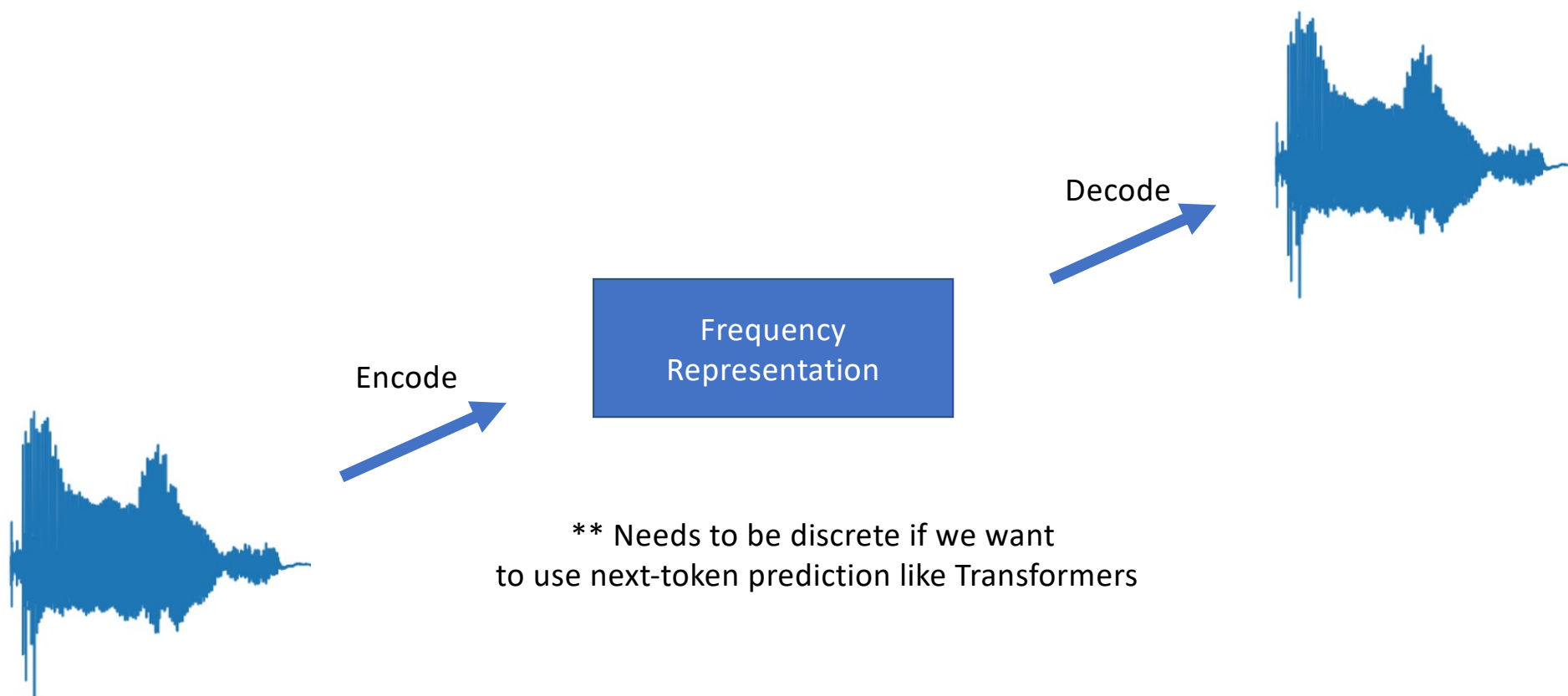
- <https://valle-demo.github.io/>

Text	Speaker Prompt	Ground Truth	Baseline	VALL-E
They moved thereafter cautiously about the hut groping before and about them to find something to show that Warrenton had fulfilled his mission.				
And lay me down in thy cold bed and leave my shining lot.				

Why it works

- **Part 1:** Large pre-training data
 - Pre-training stage, scale up the text-to-speech synthesis (TTS) training data to 60K hours of English speech which is hundreds of times larger than existing systems
- **Part 2:** Discrete Pipeline
 - **Traditional:** phoneme → mel-spectrogram → waveform
 - **New:** phoneme → discrete code → waveform

How to represent sound



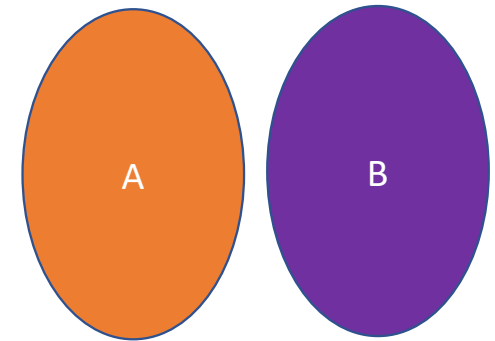
Comparison between normal systems and VALL-E

	Current Systems	VALL-E
Intermediate representation	mel spectrogram	audio codec code
Objective function	continuous signal regression	language model
Training data	≤ 600 hours	60K hours
In-context learning	✗	✓

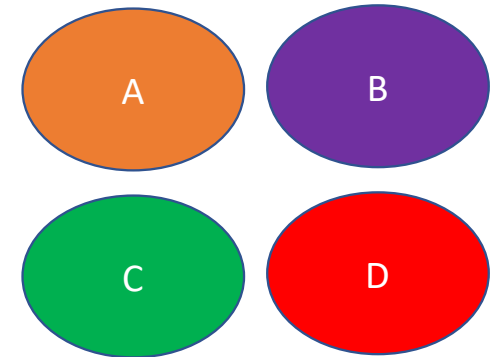
Part 1: Large Data

- Large supervised set
 - Trained with LibriLight, a corpus consisting of 60K hours of English speech with over 7000 unique speakers
 - Original data is audio-only, so speech recognition model was used to generate the transcriptions
- Able to get samples of various speaker profiles
- Can work zero-shot on a new sample

Manifold of having only 2 classes

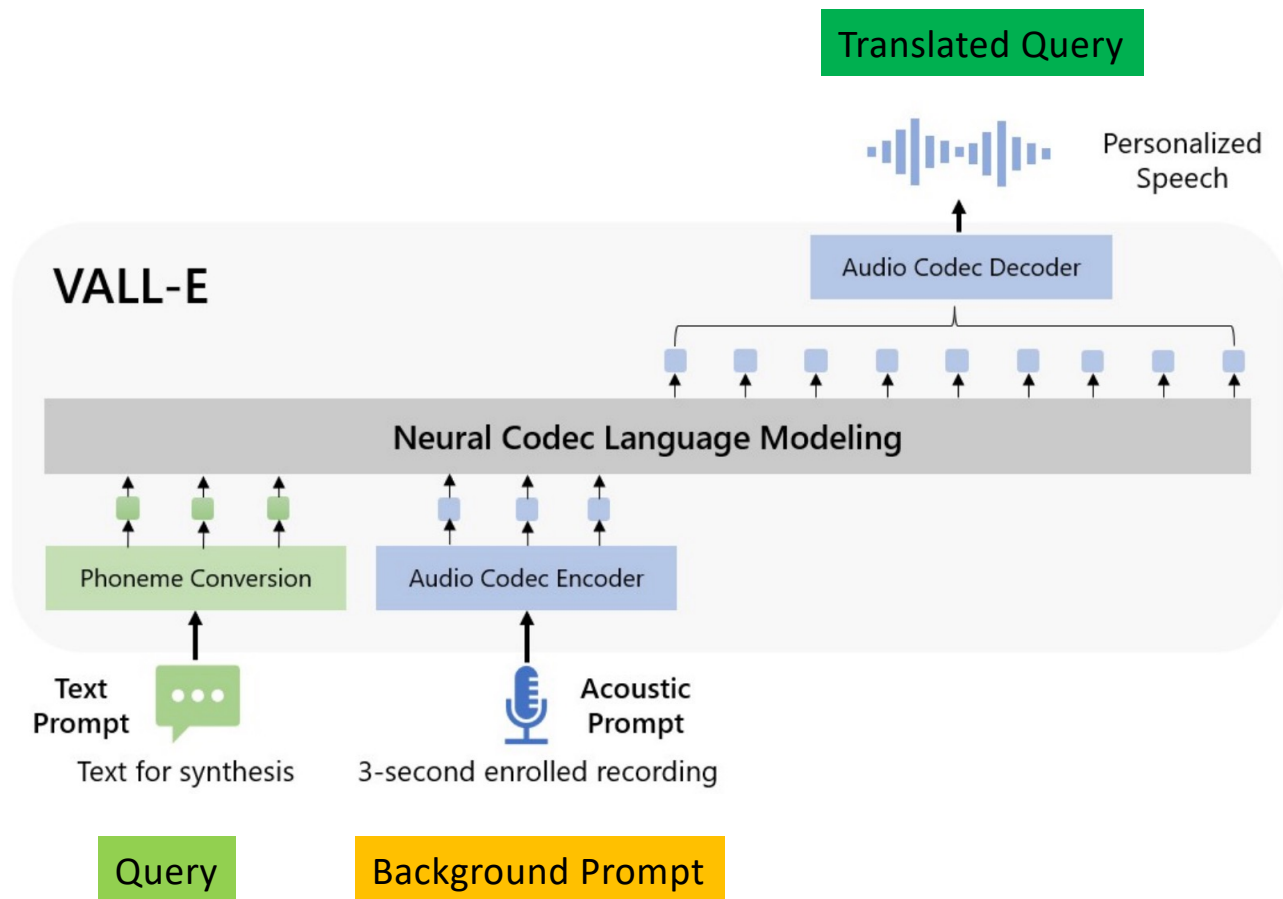


Manifold of having 4 classes



Part 2: Data Representation

- Converting audio into discrete acoustic tokens and back to audio
 - Audio Codec Model
 - Based off Mel Spectrogram
- Transformer-like next-token prediction
- In-context learning capability and enables prompt-based approaches for zero-shot TTS
 - Does not require additional structure engineering, pre-designed acoustic features, and fine-tuning as in previous work



Prompt Information

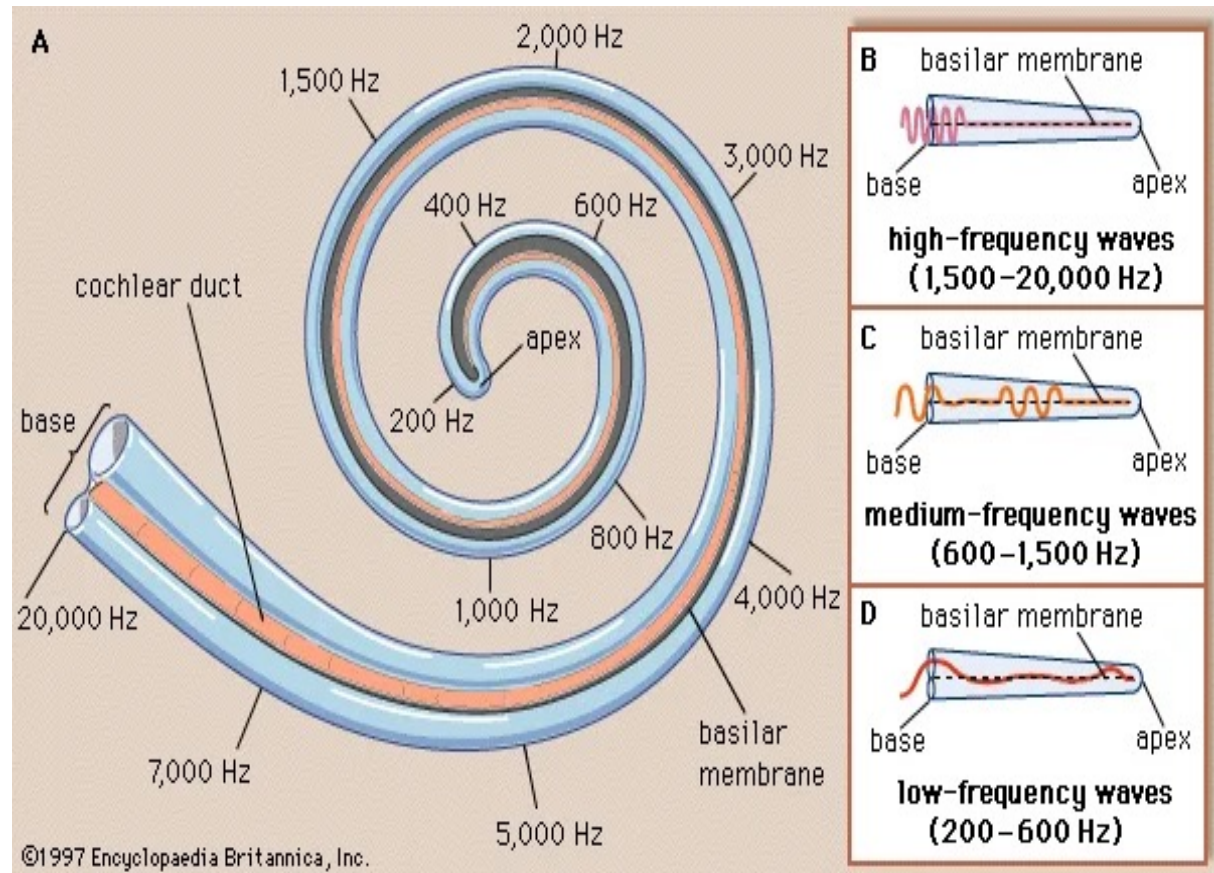
- 3-second prompt
- Constrains speaker information
- Constrains background information

Mel Spectrogram

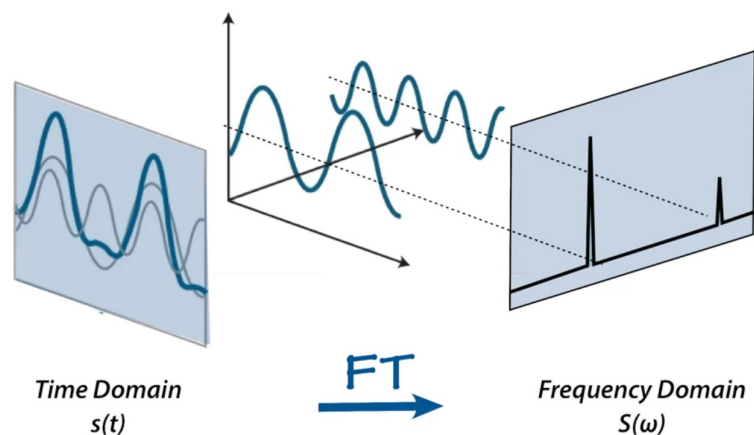
What is in a good representation?

The human ear

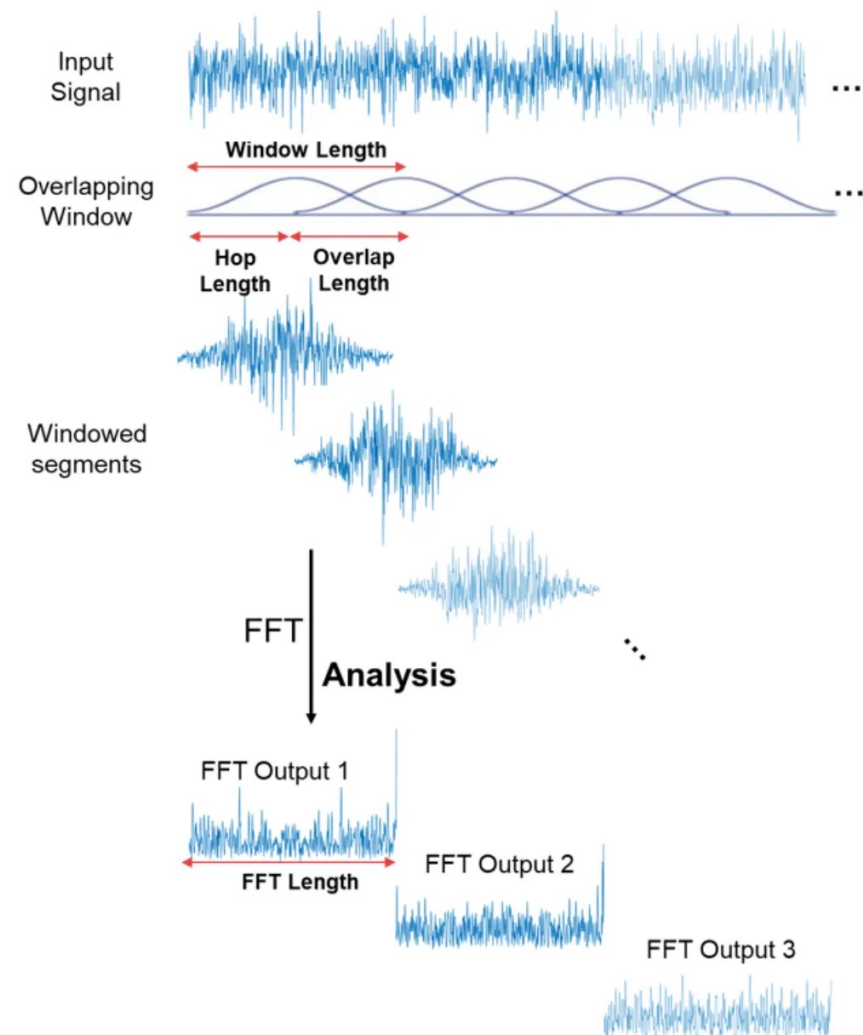
- Cilia on the cochlear convert air vibrations into frequency
- Structural way to in-build bias into the system



Mel Spectrogram



Audio signal converted into frequencies



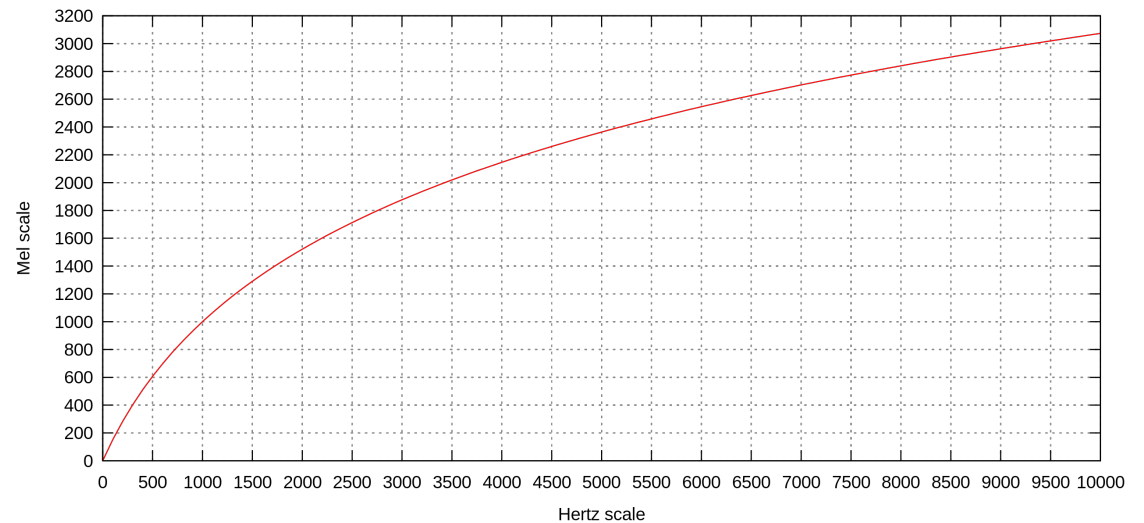
Use overlapping window to model waveform over time

Images taken from: <https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53>

Mel Spectrogram

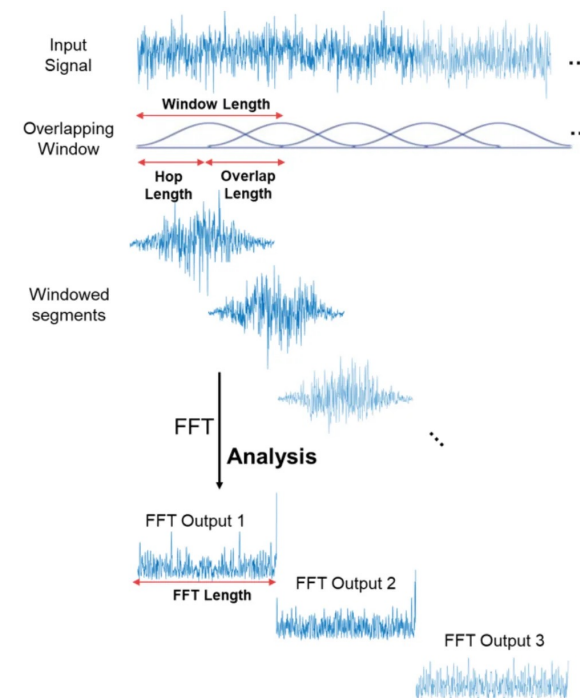
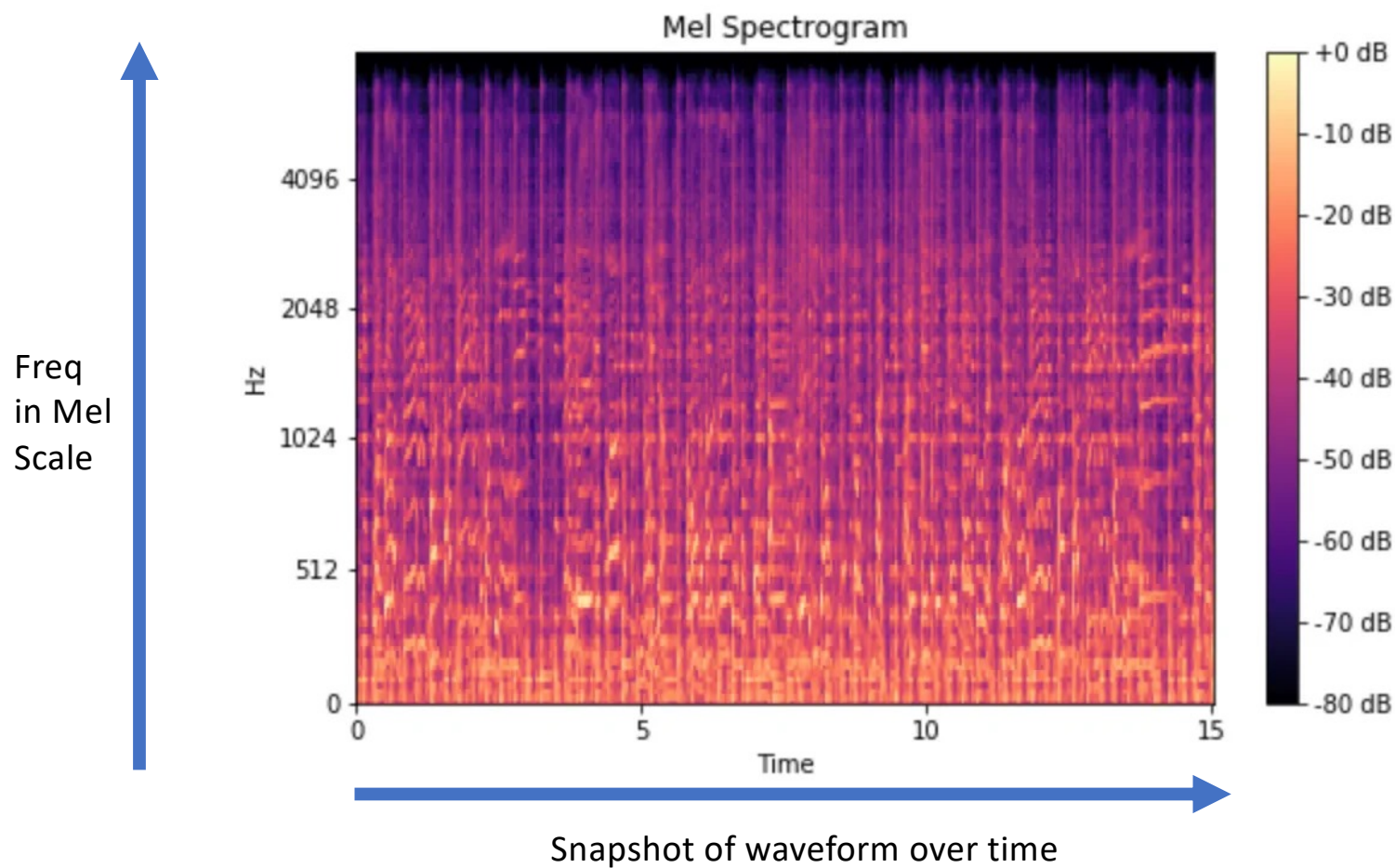
- mel (melody) scale converts frequency into pitch
- Converted to make it equidistant for human cognition
 - Humans perceive frequency logarithmically

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

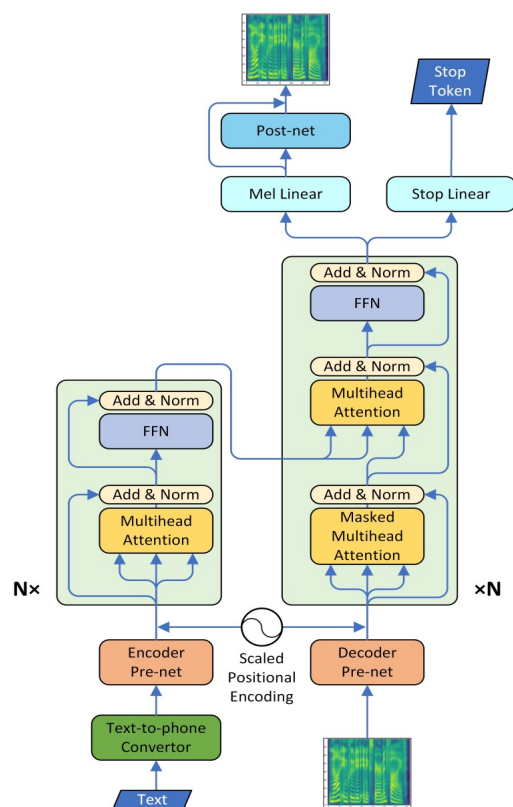


Images taken from: https://en.wikipedia.org/wiki/Mel_scale

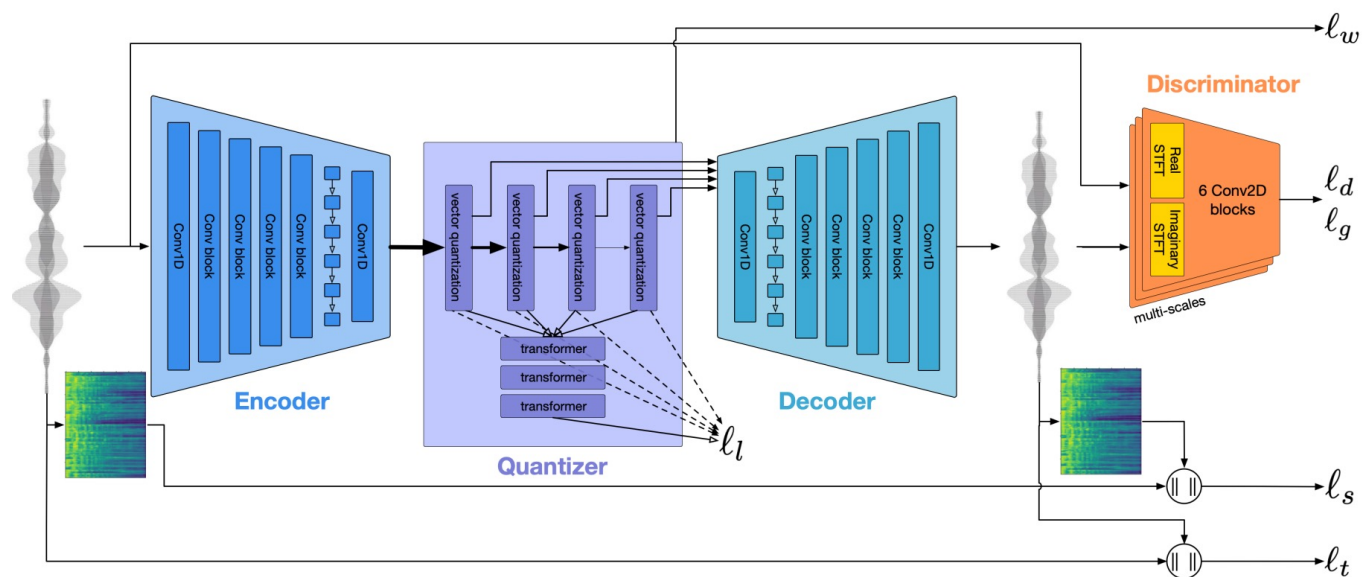
Mel Spectrogram



Modern - Mel Spectrogram approximated by Latent Representation

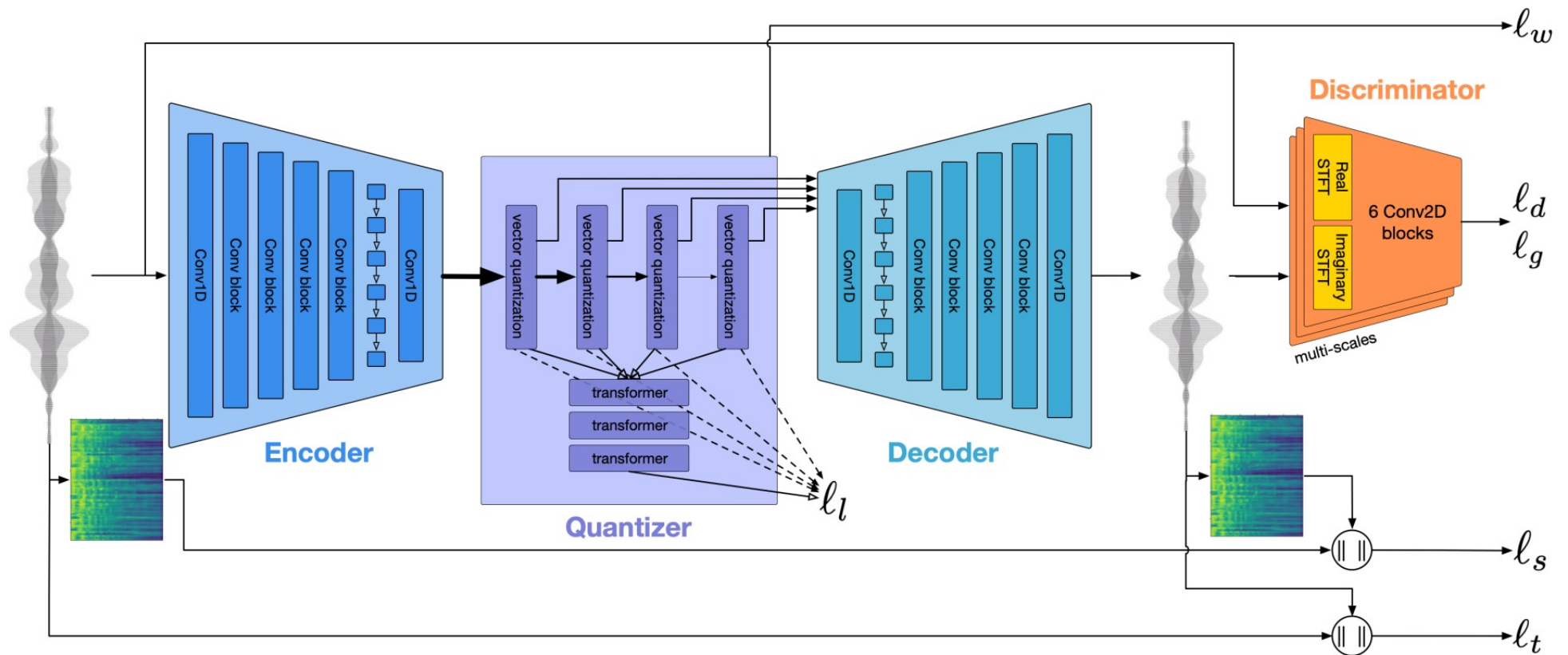


Neural Speech Synthesis with
Transformer Network. Li et al. 2019. AAAI



Encodec (used in VALL-E)
High Fidelity Neural Audio Compression. Défossez et. al. 2022

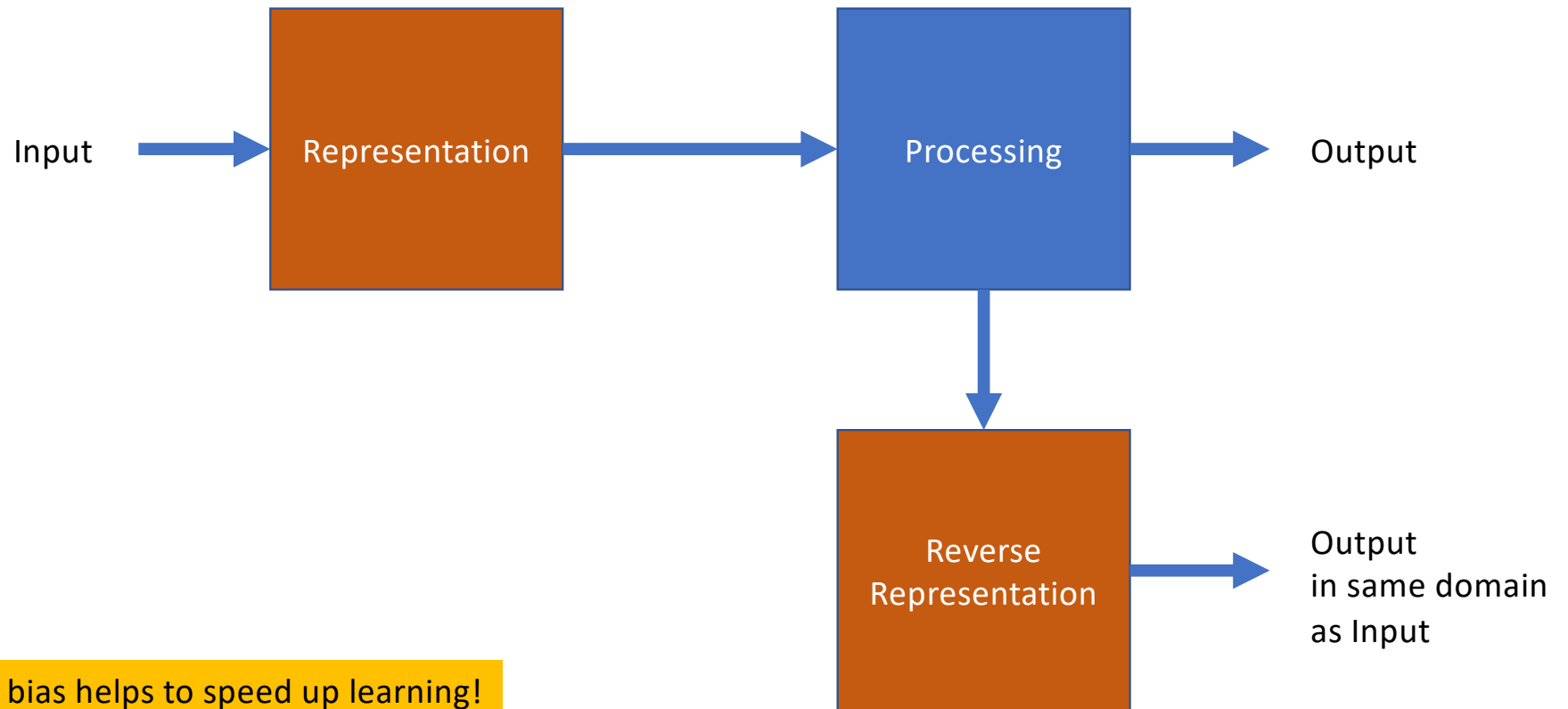
Encodec



Encodec (used in VALL-E)

High Fidelity Neural Audio Compression. Défossez et. al. 2022

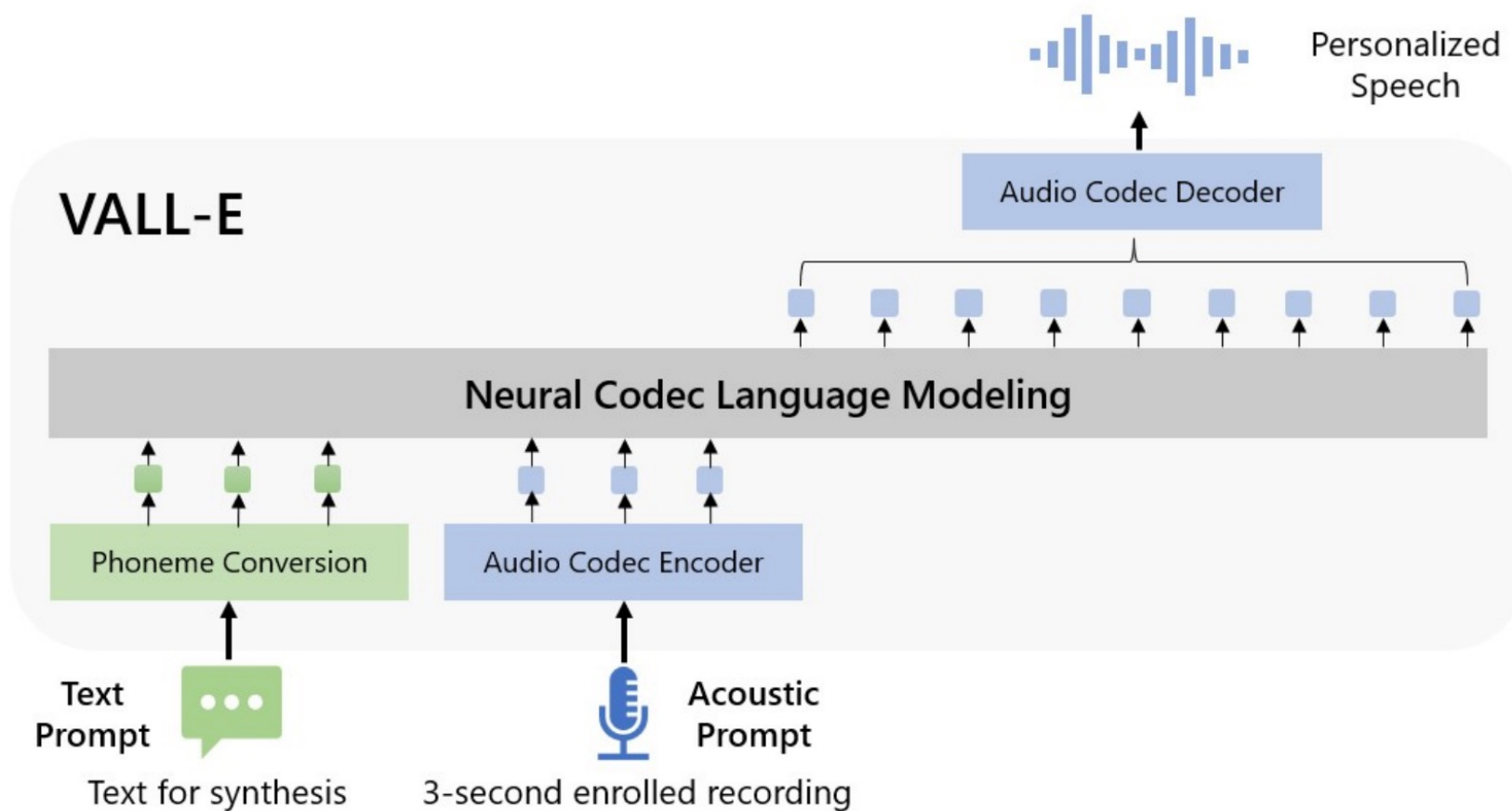
Information Pipeline



Code for Encodec!

VALL-E

Overall Structure



Audio Codec Encoder/Decoder

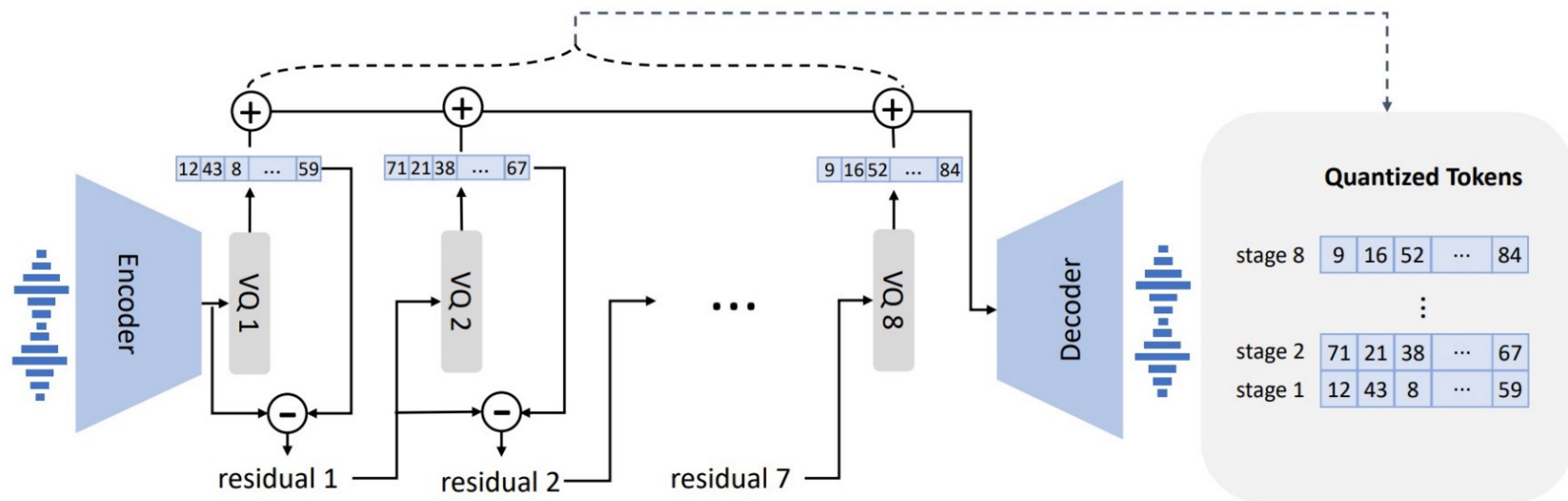


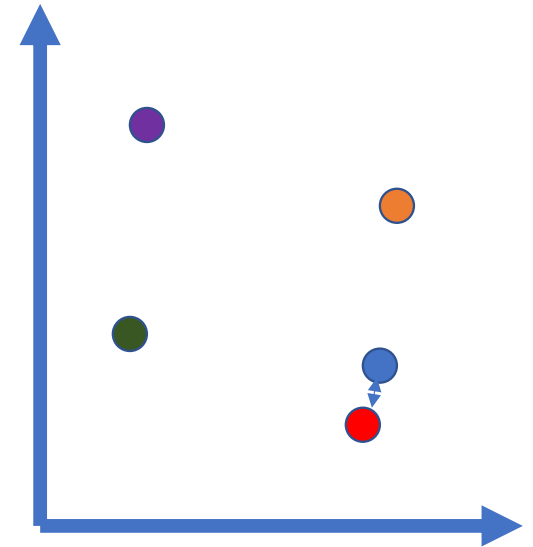
Figure 2: The neural audio codec model revisit. Because RVQ is employed, the first quantizer plays the most important role in reconstruction, and the impact from others gradually decreases.

Residual Vector Quantization (RVQ):

8 hierarchy quantizers (8 different layers) with 1024 codewords

Residual Vector Quantization (RVQ)

- Quantize the output of the encoder
- Vector quantization consists in projecting an input vector onto the closest entry in a codebook of a given size
- RVQ refines this process by computing the residual after quantization, and further quantizing it using a second codebook, and so forth
- Residual vector quantization helps to express more information in fewer vectors
 - E.g. 100 vectors per second at 3 kbps, and using 5 quantizer layers, the codebook size goes from 1 billion to 320



Decoder

- **Auto-Regressive (AR)** used for first quantizer
 - Attend only to previous tokens
 - Generated sequence length can be unknown
 - $O(T)$ time complexity
- **Non-Auto-Regressive (NAR)** used for 2nd to 8th quantizer
 - Attend to all tokens
 - Generated sequence length is known
 - $O(1)$ time complexity
 - Can generate all sequences in parallel
- Separate sinusoidal position embeddings for prompts and output

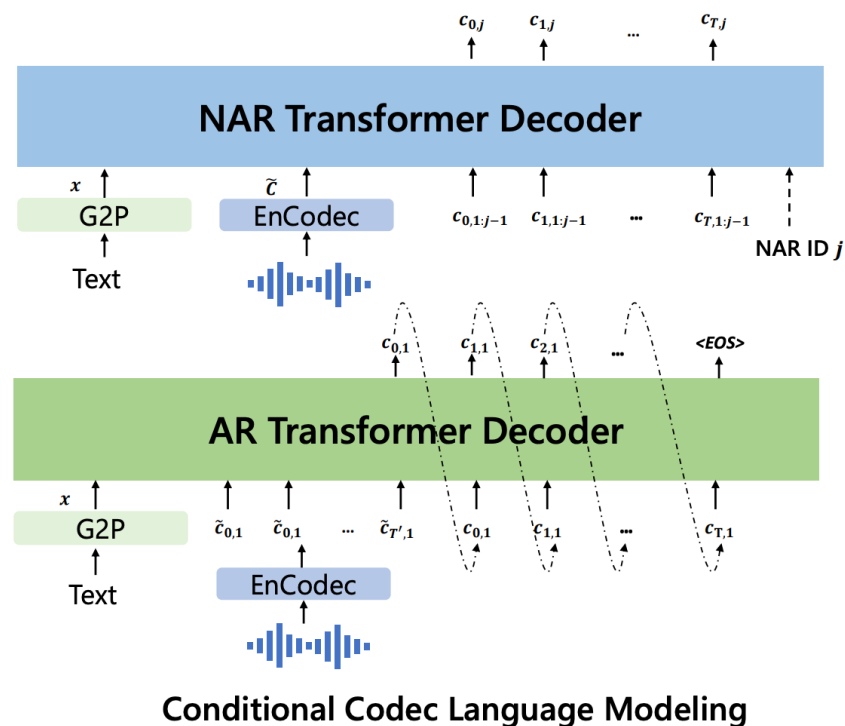
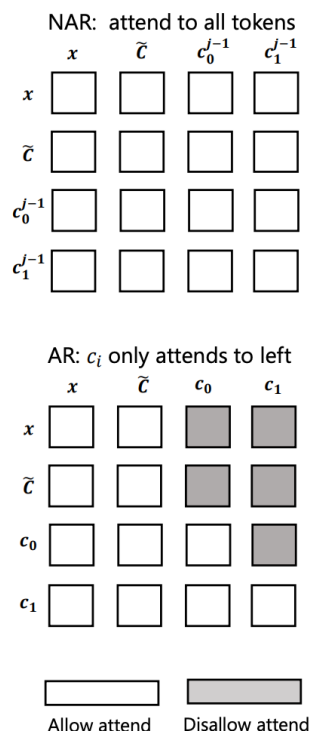


Figure 3: The structure of the conditional codec language modeling, which is built in a hierarchical manner. In practice, the NAR decoder will be called seven times to generate codes in seven quantizers.

Loss Function

- Token embeddings e are summed up from all 8 layers
- AR Loss is autoregressive loss of predicting next token for layer 1
- NAR Loss is summed loss of predicting all tokens for a chosen layer between 2-8
 - Which layer to predict is randomly sampled
- **Next-token prediction is done in embedding space (though may not have semantic meaning unlike text-based Transformers)**

$$e_{c_t,j} = W_a^j \odot c_{t,j}$$
$$\mathbf{e}_{c_t} = \sum_{j=1}^{i-1} e_{c_t,j}$$

Benefits of Neural Audio Codec

- Uses vector quantization: Contains abundant speaker and acoustic information
 - Possibly more information than mel spectrogram
- Large reduction in sampling rate
 - The encoder produces embeddings at 75 Hz for input waveforms at 24 kHz, which is a 320-fold reduction in the sampling rate
- Off-the-shelf Codec decoder to convert discrete tokens to waveform

Results

Results – Program-based Evaluation

- **Word Error Rate** evaluated by using an Automatic Speech Recognition program
- **Speaker Similarity** predicted by WavLM-TDNN (Chen et. al, 2022)
- **Previous SOTA (YourTTS, GSLM, AudioLM) not trained on VALL-E's large dataset
- Overall, VALL-E performs the best

	Word Error Rate	Speaker Similarity [-1, 1]
model	WER	SPK
GroundTruth	2.2	0.754
Speech-to-Speech Systems		
GSLM	12.4	0.126
AudioLM*	6.0	-
TTS Systems		
YourTTS	7.7	0.337
VALL-E	5.9	0.580
VALL-E-continual	3.8	0.508

Results – Human Evaluation

	SMOS	CMOS (v.s. VALL-E)
YourTTS	$3.45_{\pm 0.09}$	-0.12
VALL-E	$4.38_{\pm 0.10}$	0.00
GroundTruth	$4.5_{\pm 0.10}$	+0.17

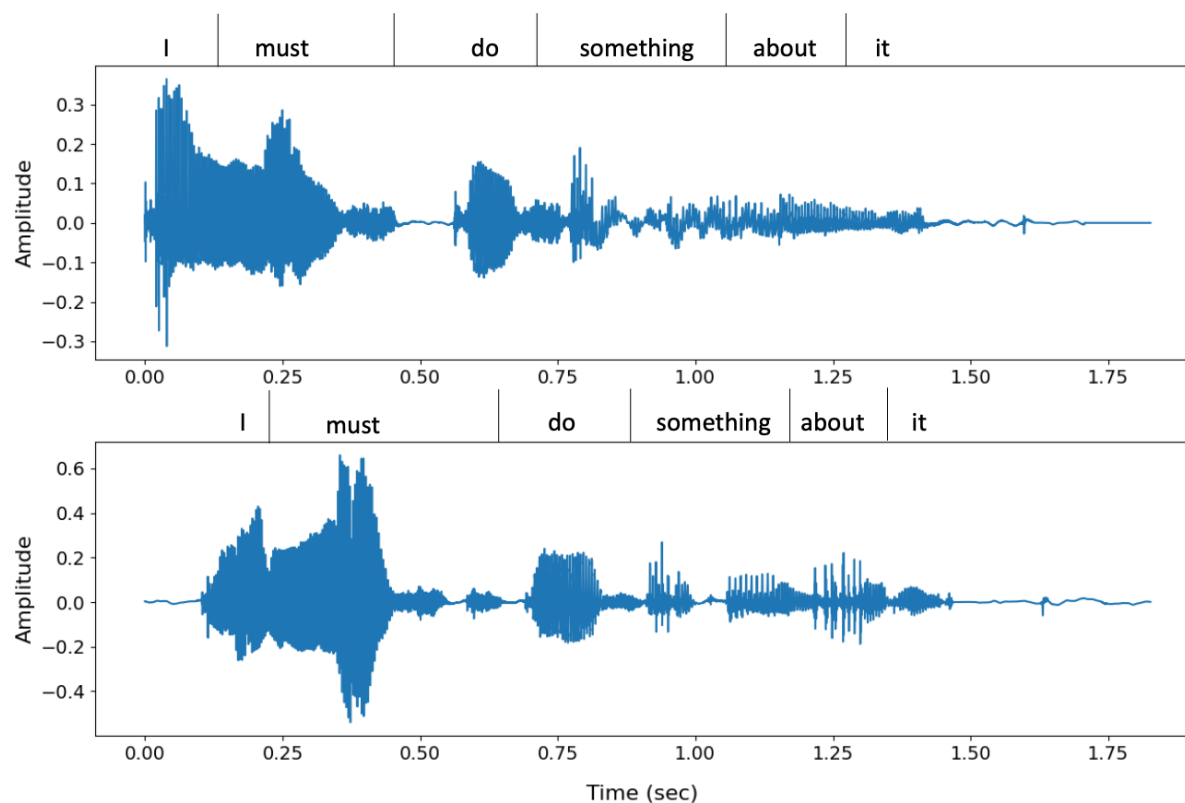
- Comparative mean option score (CMOS):
 - Indicator of speech naturalness, ranked -3 to 3 w.r.t. VALL-E
- Similarity mean option score (SMOS)
 - Indicator of speaker similarity, ranked 1-5

VALL-E has good zero-shot speaker similarity

Table 6: Automatic evaluation of speaker similarity with 108 speakers on VCTK. *YourTTS has observed 97 speakers during training, while VALL-E observed none of them.

	3s prompt	5s prompt	10s prompt
108 full speakers			
YourTTS*	0.357	0.377	0.394
VALL-E	0.382	0.423	0.484
GroundTruth	0.546	0.591	0.620
11 unseen speakers			
YourTTS	0.331	0.337	0.344
VALL-E	0.389	0.380	0.414
GroundTruth	0.528	0.556	0.586

Random Sampling of VALL-E leads to diversity

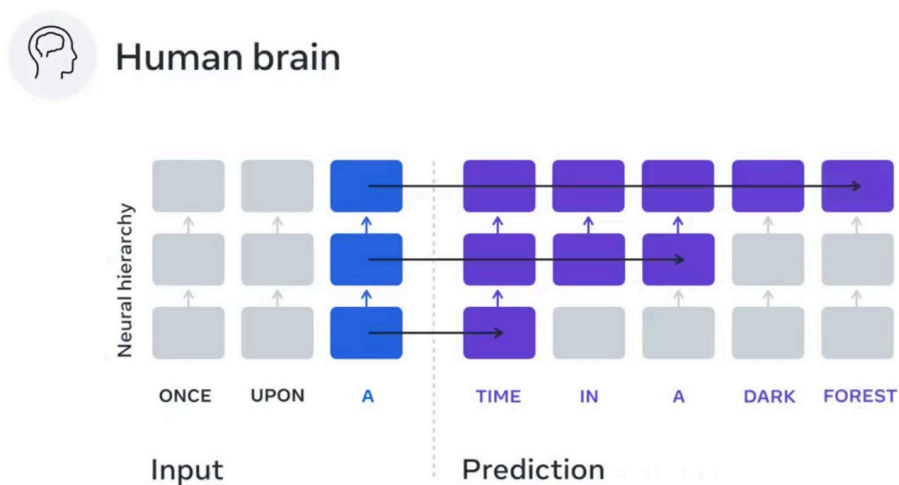


(b) A VCTK sample: I must do something about it.

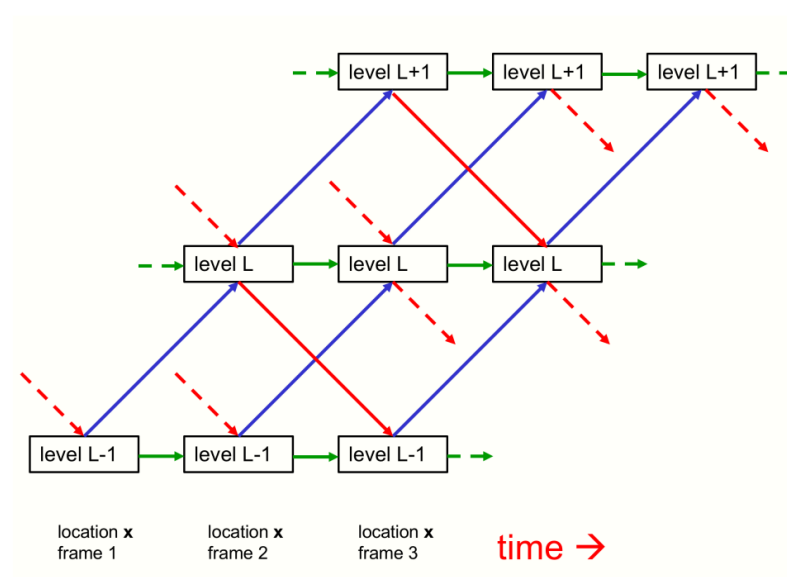
Limitations of model

- Acoustic-to-phoneme tokenization may cut at inappropriate points
 - Words unclear, missed, duplicated in speech synthesis
 - **My thoughts: Maybe we need an overlapping prediction window instead of discrete tokens without overlap**
 - **My thoughts: Might some level of hierarchical prediction to constrain output be required?**
- VCTK speaker similarity can be bad due to insufficient accent speakers even in 60K hours of data
 - **My thoughts: Could the superior performance in speaker similarity be just due to increased training data?**

Hierarchical Prediction?

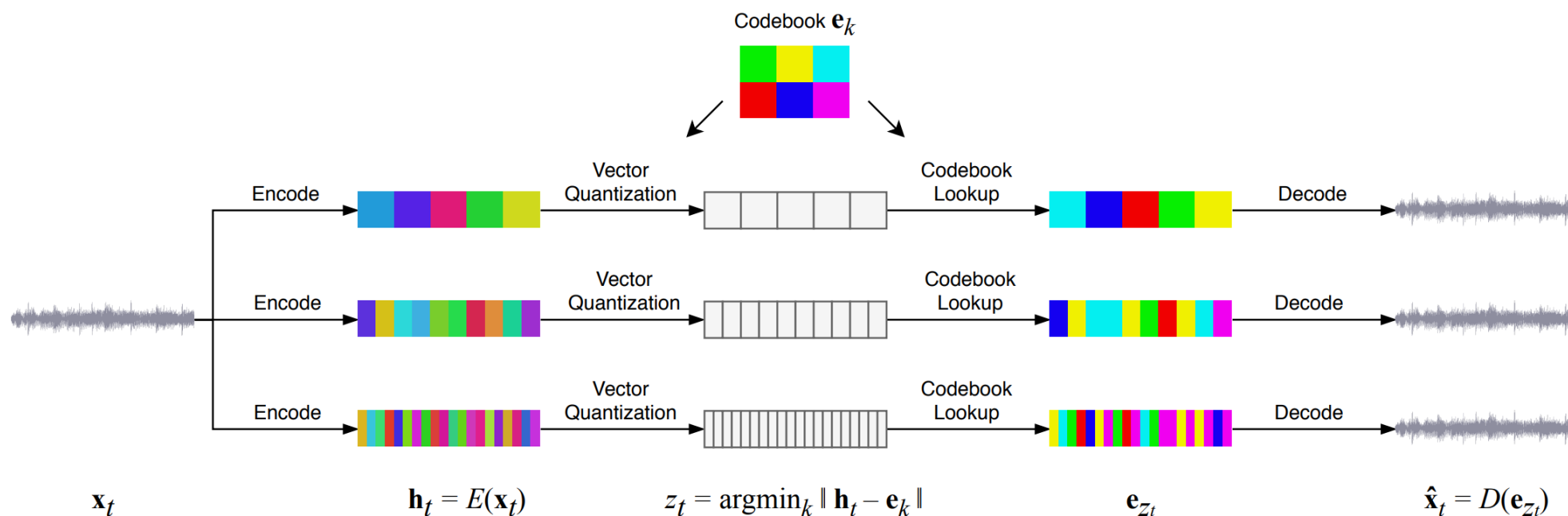


Evidence of a predictive coding hierarchy in the human brain listening to speech.
Caucheteux. 2022. Nature Human Behaviour.



How to represent part-whole hierarchies in a neural network. Hinton. 2021.

Hierarchical Prediction – Jukebox (OpenAI)



Jukebox: A Generative Model for Music. Dhariwal et. al. 2020.

Questions to Ponder (Part 1 – Representation Learning)

- Why not just use the Mel Spectrogram representation instead of Audio Codec?
- Is the way of tokenizing in audio correct? What if the split is unnatural? What about hierarchical prediction?
- Is the embedding space of Encodec semantically meaningful?
- Is reconstruction / Generative Adversarial Network (GANs) loss good for training embedding space representation?

Questions to Ponder (Part 2 – Conditional Generation)

- Since VALL-E can give various samples via random sampling, can there be a CLIP-like ranking to select the best generated speech output?
- How do we prompt VALL-E for different emotions, context etc?
- How do we do efficient online learning for this model?