A Short Introduction of Modern Speech Foundation Models

Cardiff NLP Workshop

2nd July 2024

Asahi Ushio

- HP: https://asahiushio.com/
- X: https://x.com/asahiushio
- GitHub: https://github.com/asahi417

About Me

Past

- PhD in NLP at Cardiff University, UK (Oct 2020-Dec 2023)
    - Representation Learning, Question Generation, Social Media
    - Research Internship: Google (MusicLM), Snapchat (Computational Social Science), Amazon (Search Technology).

Now

- Applied Scientist at Amazon, Japan (Jan 2024-)
    - Information Retrieval.
- Research Collaborator at Kotoba Technology, Japan (Mar 2024-)
    - Japanese and English bilingual speech foundation model and its application.

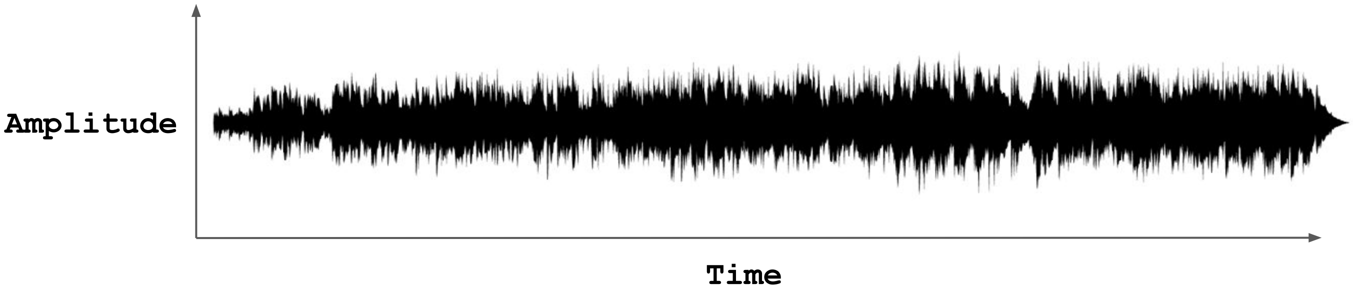
Topics

- Small introduction of speech foundation models
- ToC
    - Basics of audio data
    - Speech foundation model
    - Speech-text downstream tasks
    - Audio tokenizer
    - Representation learning
    - Future works

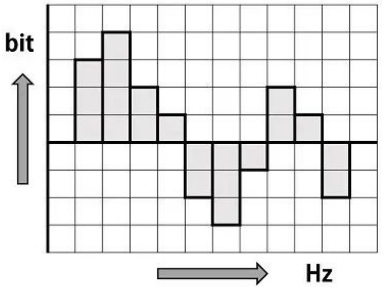
Audio Data

Audio Signal

- Audio is continuous **wave of amplitude over time**.

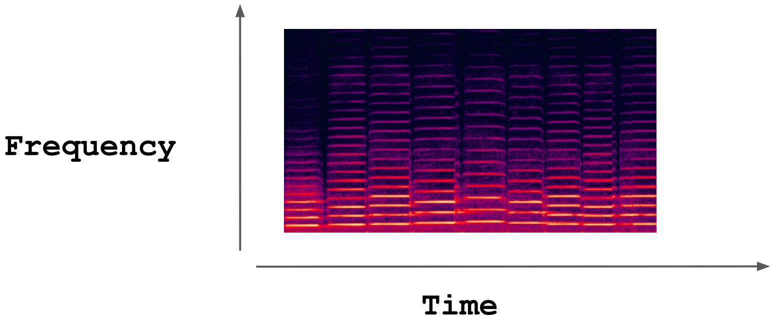


- Digital audio is **quantization** of raw audio.
  - Bit depth (bps): Resolution of each sample.
  - Sampling Rate (Hz): Resolution, *N* Hz means sampling every 1/N second.



Spectrogram

- Spectrogram is **the power distribution** over different frequency level within a **short time window**.



- The digital audio is referred as **raw audio** in contrast to spectrogram.
- Length of spectrogram is much **smaller** than the raw audio.
- Commonly used as an **input feature** to speech task (speech classification or recognition).

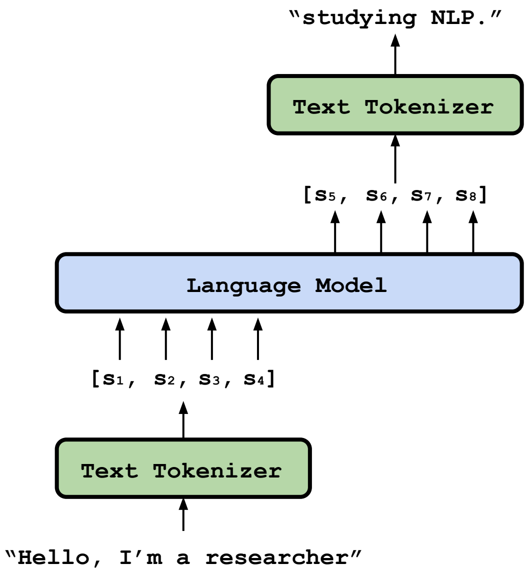
Speech & Text Supervised Tasks

Task	Input	Output
Automatic Speech Recognition (ASR)	Speech (audio)	Transcription (text)
Speech Style Classification	Speech (audio)	Label (text)
S2T translation	Speech (audio)	Translation (text)
Speech-to-speech audio generation (S2S)	Speech (audio)	Speech (audio)
Text-to-speech (TTS)	Transcription (text)	Speech (audio)
S2S translation	Speech (audio)	Translation (audio)

Modelling Speech with LMs

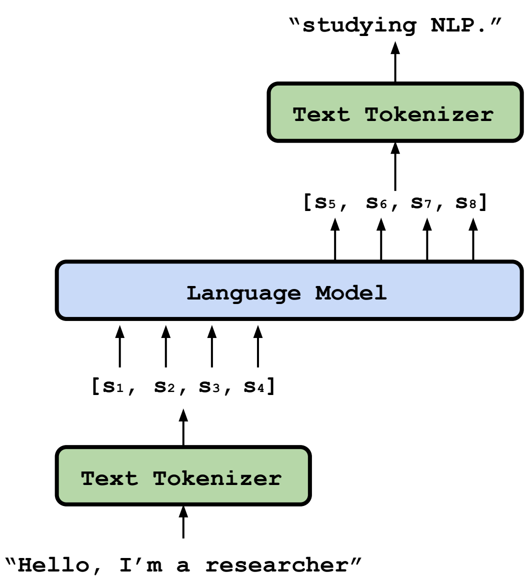
Language Model

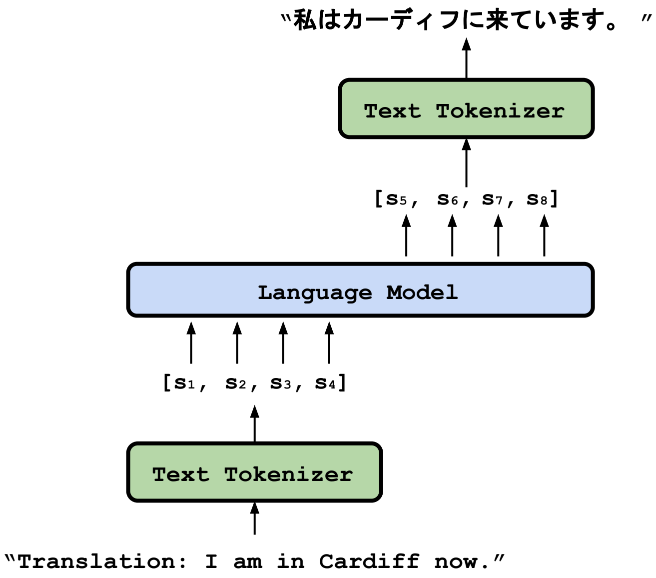
- Predict succeeding text given the precedent text.



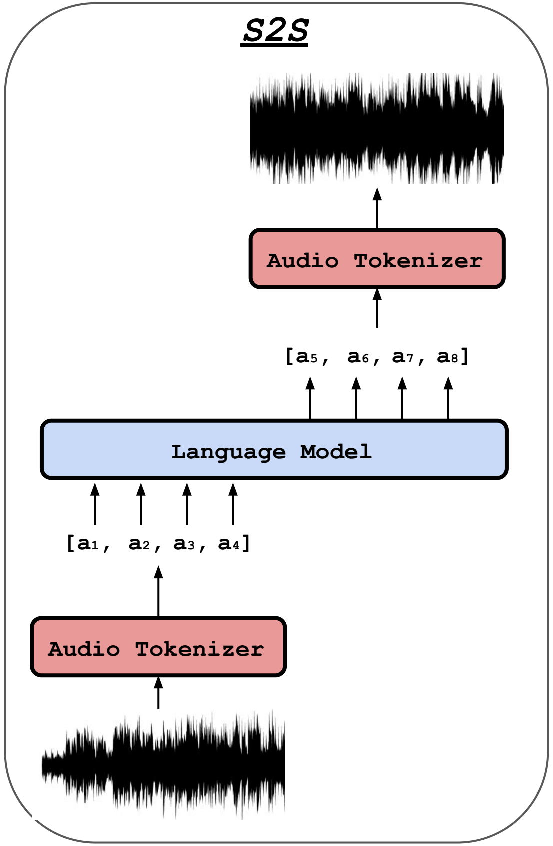
Language Model

- Predict succeeding text given the precedent text.
- Fine-tune on task I/O in text format (text2text).

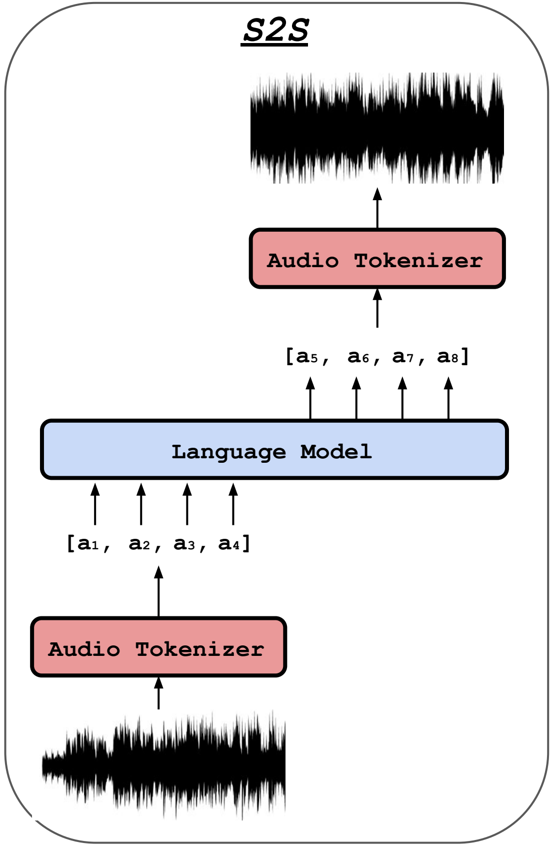


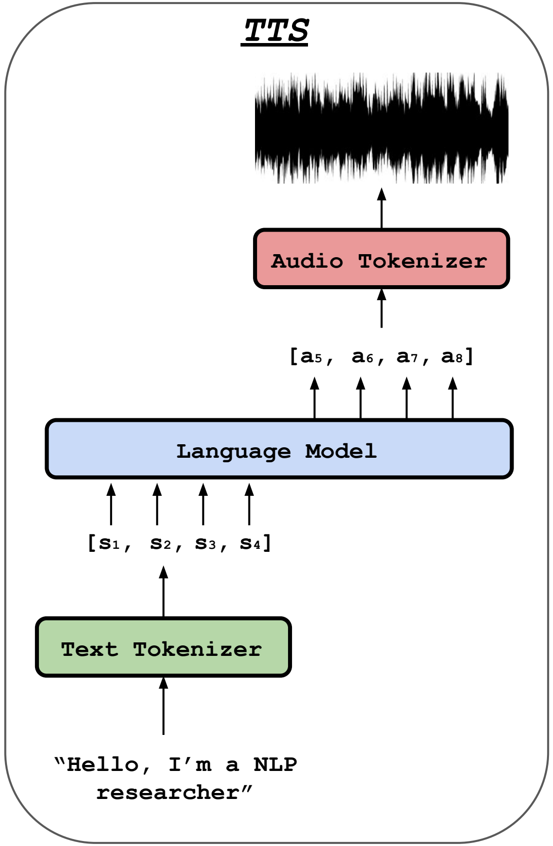


Speech Modelling

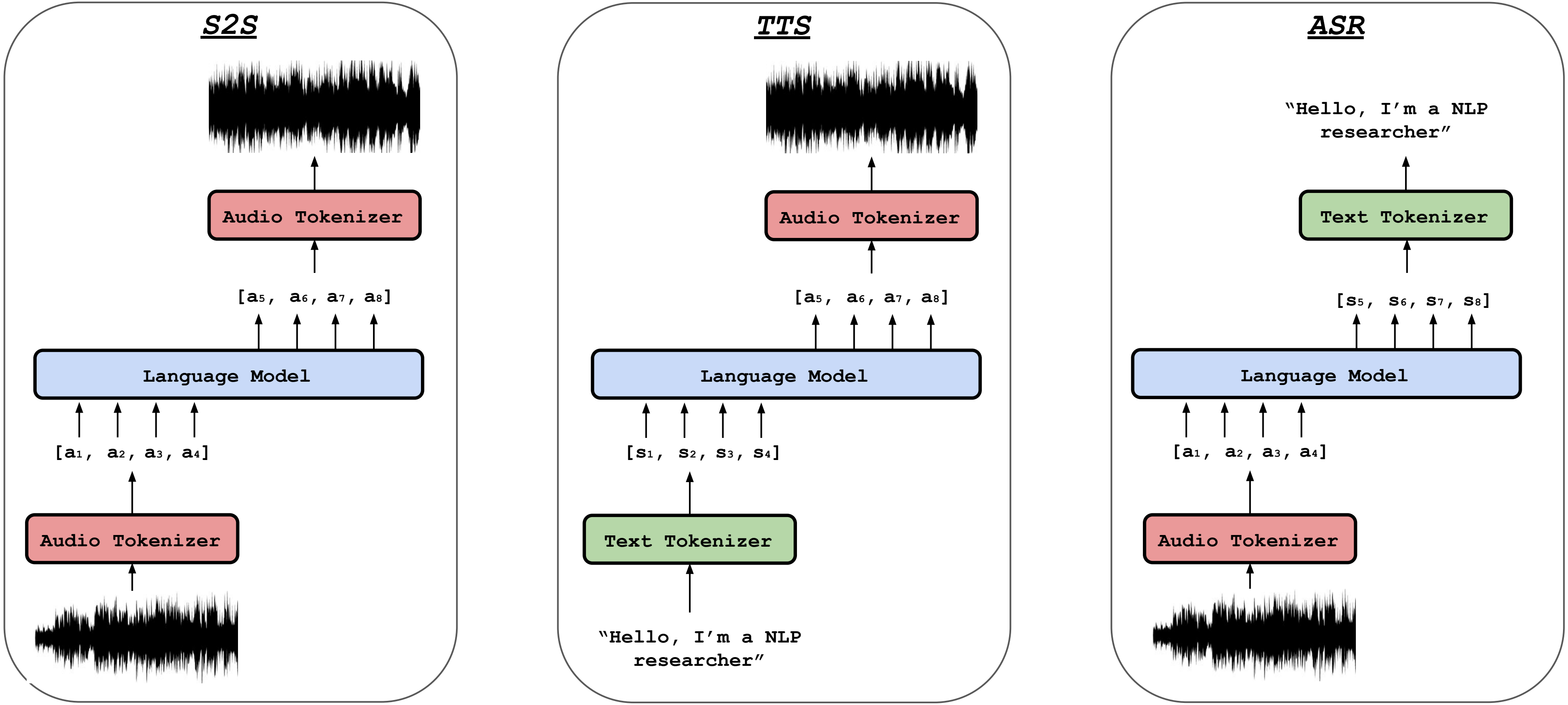


Speech Modelling





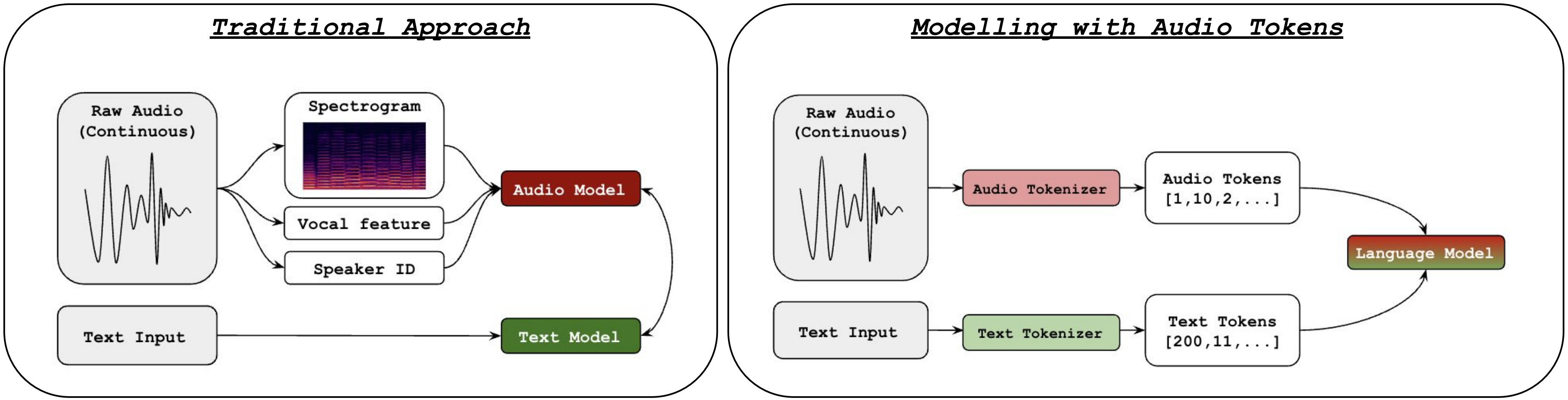
Speech Modelling



Audio Tokenizer

Audio Tokens

- Audio tokenizer opened up a new direction for audio (speech) modelling.
- Seamless integration of audio to LMs (AudioPaLM, AudioGen).



Audio Tokenizer

- Discrete tokens of lower frequency than the raw audio.
  - **Acoustic** Feature: pitch, noise, accent.
  - **Semantic** Feature: meaning, grammar, bpm, melody.
- Challenges of **Tokenizer**.
  - Audio is mixture of different artifacts: speech, background noise, etc.
  - Seuqnece length can be large.
    - Eg) 320 tokens per second.
- Challenges of **De-tokenizer**.
  - De-tokenizer has to be a generative model of raw audio wave.
  - High fidelity and adhesive to the tokens.

Different Types of Tokenizers

- ***<u>Neural Codec based Tokenizer</u>***: SoundStream, Encodec
  - Model-based audio codec (compression).
  - Encoder (tokenizer) and decoder (de-tokenizer) architecture.
    - **Pros:** Joint training, Acoustic feature.
    - **Cons:** Lack of semantic feature.
- ***<u>Embedding based Tokenizer</u>***: w2vBERT, HuBERT, XLS-R
  - Unsupervised model trained on contrastive loss + α.
  - Tokenizer: clustering embeddings (eg. k-means).
  - Detokenizer: Vocoder trained separately on the audio token.
    - **Pros:** Semantic feature, Acoustic feature.
    - **Cons:** Separate training.

Acoustic & Semantic Tokens

- ***Neural Codec based Tokenizer***: SoundStream, Encodec — Acoustic Audio Token
  - Model-based audio codec (compression).
  - Encoder (tokenizer) and decoder (de-tokenizer) architecture.
    - **Pros:** Joint training, Acoustic feature.
    - **Cons:** Lack of semantic feature.
- ***Embedding based Tokenizer***: w2vBERT, HuBERT, XLS-R — Semantic Audio Token
  - Unsupervised model trained on contrastive loss + α.
  - Tokenizer: clustering embeddings (eg. k-means).
  - Detokenizer: Vocoder trained separately on the audio token.
    - **Pros:** Semantic feature, Acoustic feature.
    - **Cons:** Separate training.

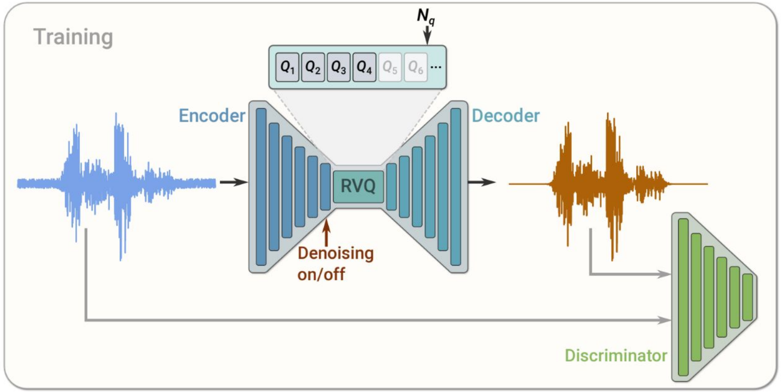
Neural Codec based Tokenizer

- ***Neural Codec based Tokenizer***: SoundStream, Encodec
  - Model-based audio codec (compression).
  - Encoder (tokenizer) and decoder (de-tokenizer) architecture.
    - **Pros:** Joint training, Acoustic feature.
    - **Cons:** Lack of semantic feature.
- ***Embedding based Tokenizer***: w2vBERT, HuBERT, XLS-R
  - Unsupervised model trained on contrastive loss + α.
  - Tokenizer: clustering embeddings (eg. k-means).
  - Detokenizer: Vocoder trained separately on the audio token.
    - **Pros:** Semantic feature, Acoustic feature.
    - **Cons:** Separate training.

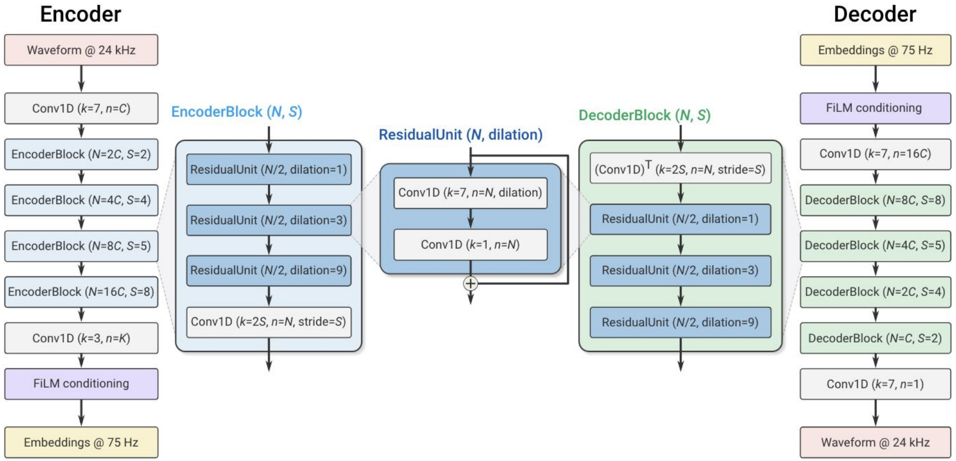
Neural Audio Codec

- Audio codec is a program to encode/decode **high-fidelity audio signal with a minimum number of bits** (eg, flac, mp3).
- Neural audio codec is **encoder-decoder neural network** model trained for audio codec.

SoundStream

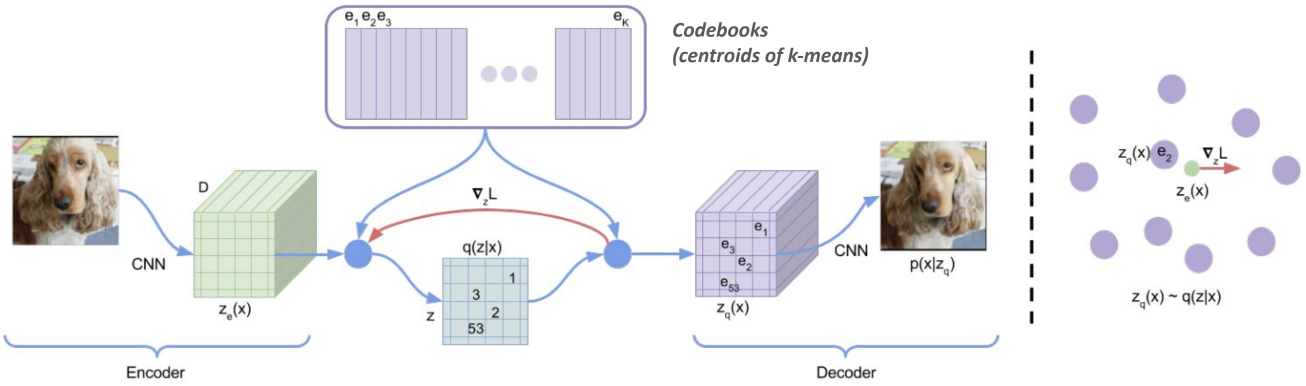


Encodec



Discrete Latent Representation

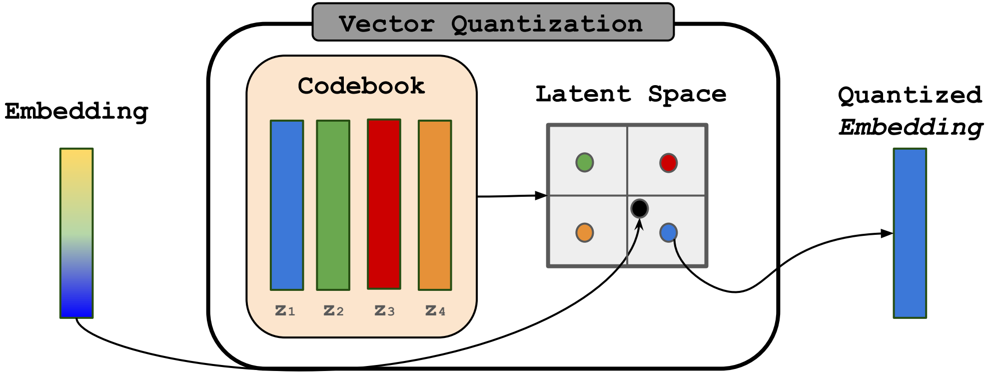
- Neural codec models are built upon **VQ-VAE**.
- VQ-VAE quantizes the latent space to avoid posterior collapse of VAE.
  - Dictionary learning (codebooks update) + Auto-encoding (VAE)



Vector Quantization

- VQ divides the vector space by *k* centroids with minimum error.
  - To represent *N* data point, VQ needs a codebook with *N* codes.

Not scalable…!



Residual Vector Quantization

- RVQ is VQ with multiple codebooks; each codebook model the residual.
- The *L*-layers RVQ quantize a vector *v* as

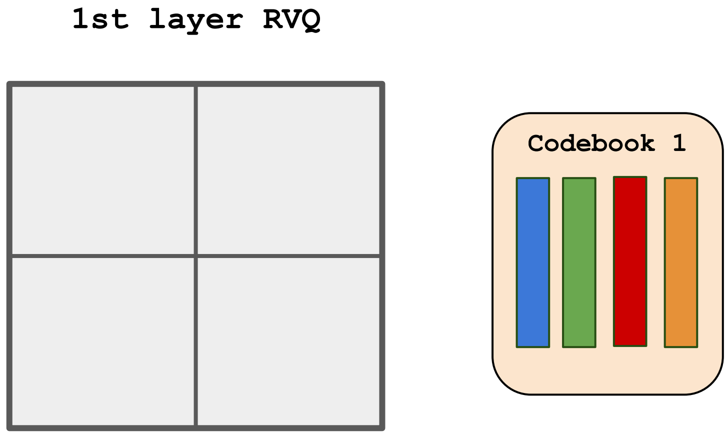
$$
\begin{aligned}
RVQ(v) &= [\hat{v}_1, \ldots, \hat{v}_L] \\
\hat{v}_1 &= Q_1(v) \\
\hat{v}_2 &= Q_2(\hat{v}_1 - v) \\
\hat{v}_3 &= Q_3(\hat{v}_2 + \hat{v}_1 - v) \\
&\vdots \\
\hat{v}_L &= Q_L\left(\sum_{i=1}^{L-1} \hat{v}_i - v\right)
\end{aligned}
$$

  where Q_l is a VQ with *l*-th codebook.

- Later RVQ layers can be ignored in practice (better controllability).

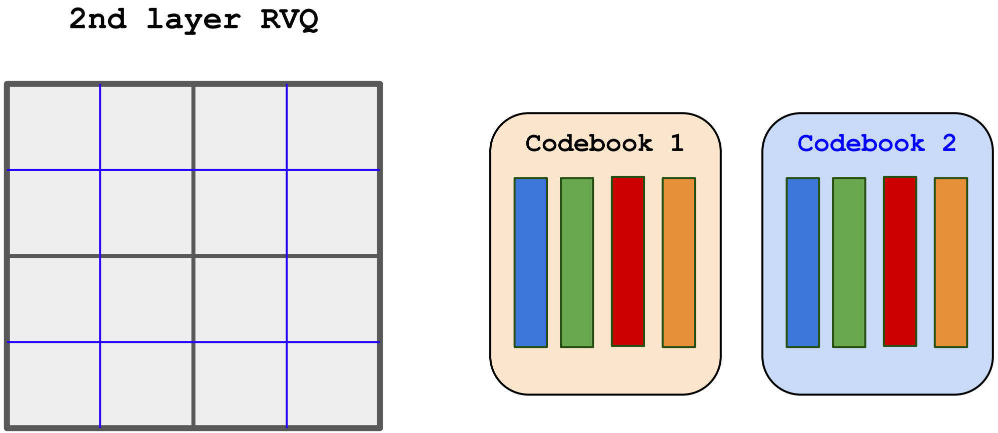
RVQ Visualization

- *RVQ* can represent more bits than VQ with the same number of codes.



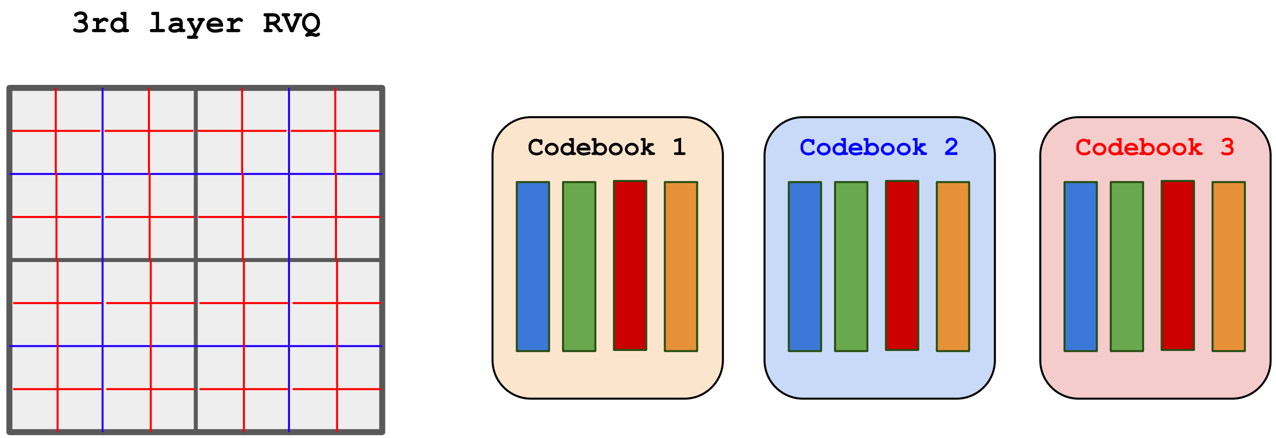
RVQ Visualization

- *RVQ* can represent more bits than VQ with the same number of codes.



RVQ Visualization

- *RVQ* can represent more bits than VQ with the same number of codes.

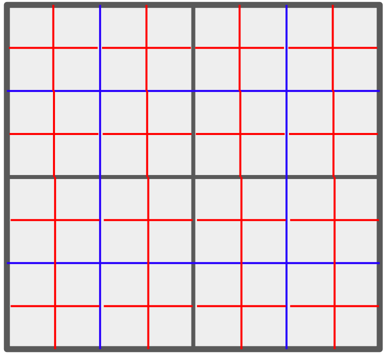


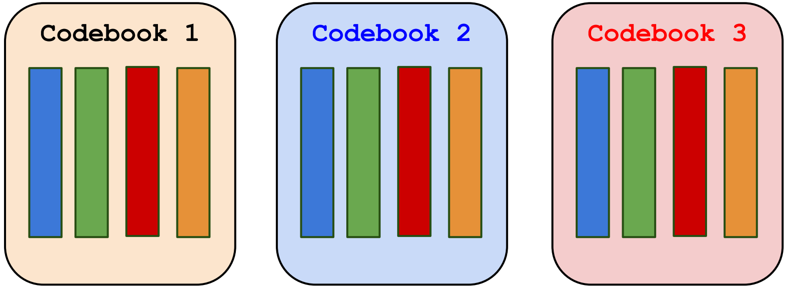
RVQ Visualization

- *RVQ* can represent more bits than VQ with the same number of codes.
- *L*-layer RVQ with *c* codes = VQ with *cL* codes.

	VQ	RVQ
Total codes	$c \times L$	$c \times L$
Data Points	$c \times L$	c^L

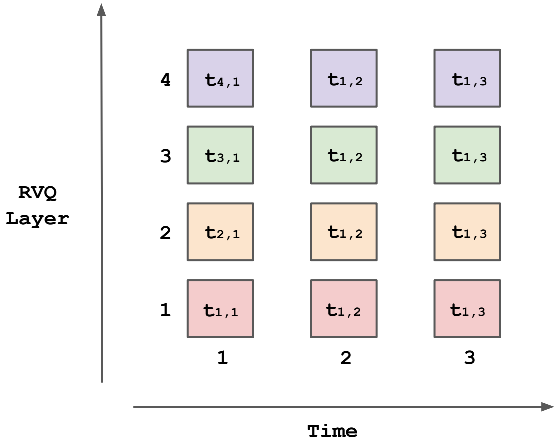
3rd layer RVQ





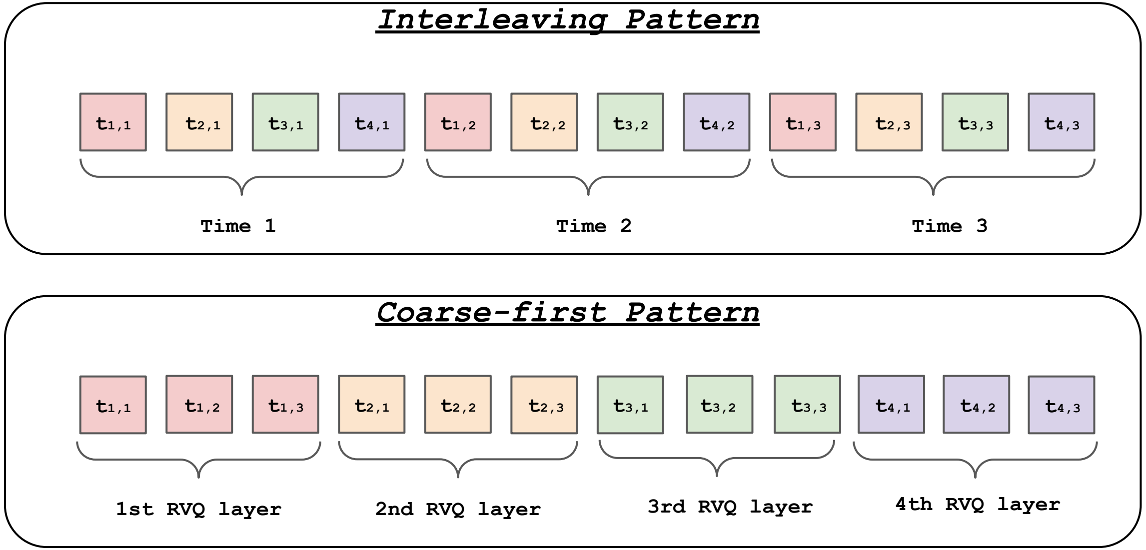
Codebook Interleaving Pattern

- RVQ tokens consists of multiple codes per sample.
- Text token is a single code per sample.



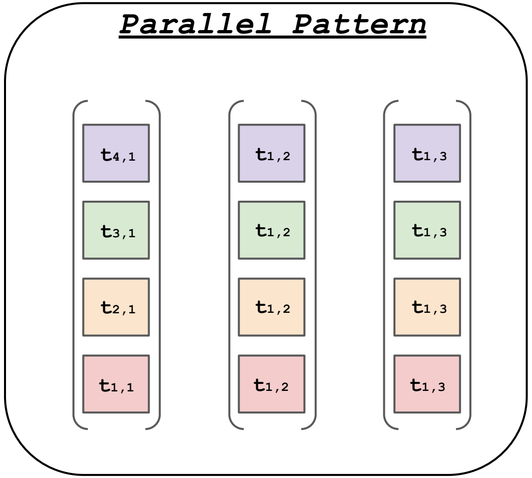
Single-stream Transformer

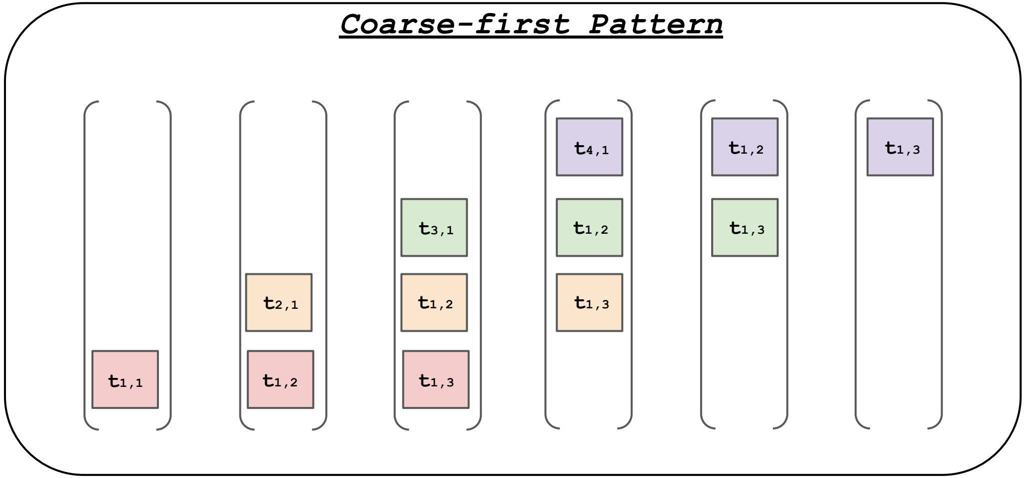
- High latency: sequence length increase linearly with the RVQ depth.
- Better performance (MusicLM).
- Versatility: Extend pre-trained LM (AudioPaLM).



Multi-stream Transformer

- Input/output multiple tokens in a single time frame (Kharitonov 2022).
- Low latency: no increase of sequence length.
- Potential decrease in quality.
- Not compatible with most text pre-trained LMs.

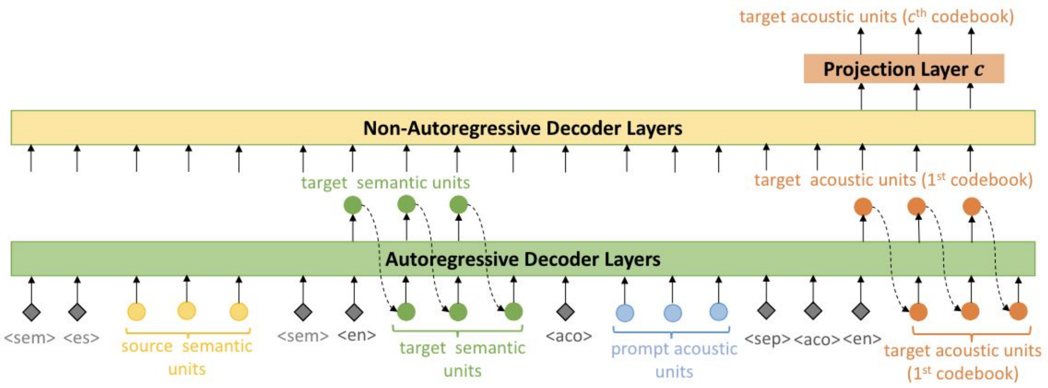




Multi-stage Models

- An autoregressive transformer for the 1st layer RVQ tokens (AR model).
- Non-autoregressive model to predict the rest RVQ tokens (NAR model).
- Pros: Versatility, Low latency, High quality.
- Cons: High complexity (MLOps).

[SeamlessExpressiveLM]()



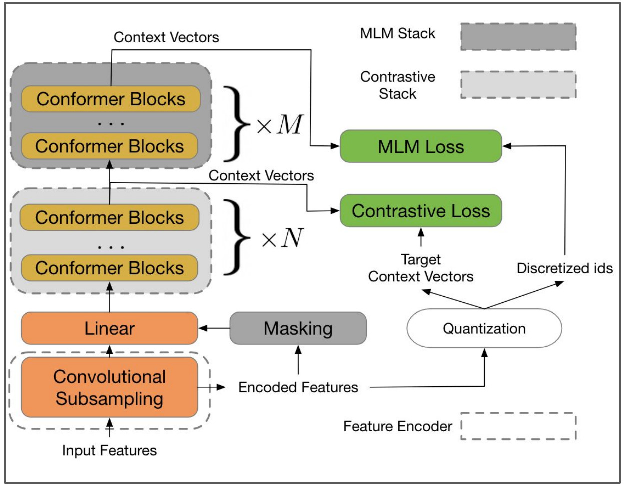
Embedding based Tokenizer

- ***Neural Codec based Tokenizer***: SoundStream, Encodec
  - Model-based audio codec (compression).
  - Encoder (tokenizer) and decoder (de-tokenizer) architecture.
    - **Pros:** Joint training, Acoustic feature.
    - **Cons:** Lack of semantic feature.
- ***Embedding based Tokenizer***: w2vBERT, HuBERT, XLS-R
  - Unsupervised model trained on contrastive loss + α.
  - Tokenizer: clustering embeddings (eg. k-means).
  - Detokenizer: Vocoder trained separately on the audio token.
    - **Pros:** Semantic feature, Acoustic feature.
    - **Cons:** Separate training.

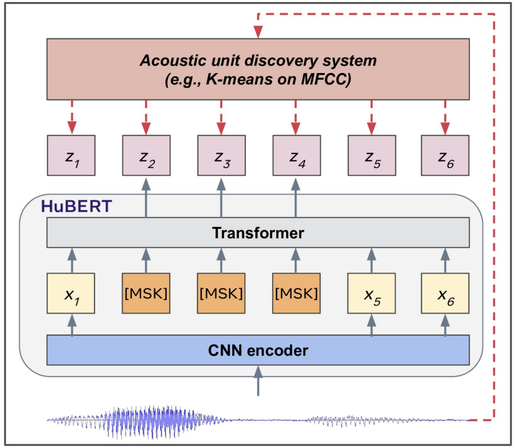
Speech Embedding Model

- Contrastive Loss (CL) + Masked Language Modelling (MLM).
- CL: surrounding tokens as the positive examples.
- MLM: predicting the masked token from the contextual embeddings.

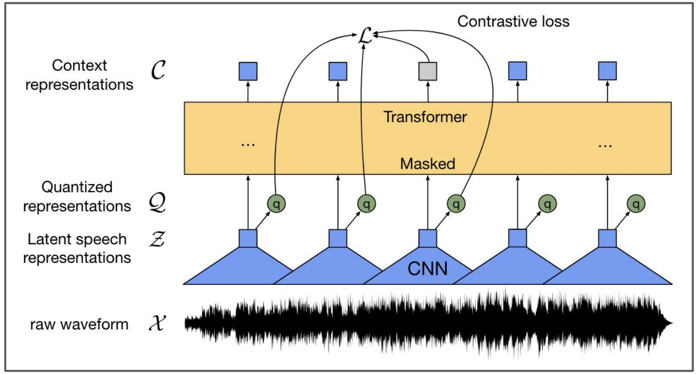
w2v-BERT



HuBERT

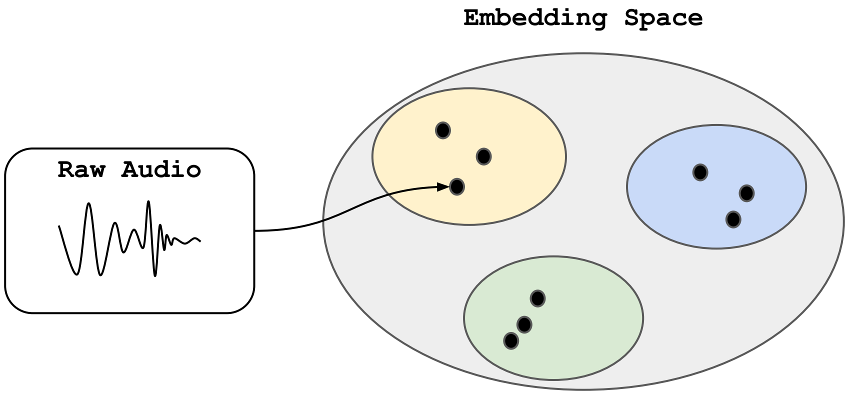


XLS-R (Wav2Vec 2.0)



Semantic Tokens

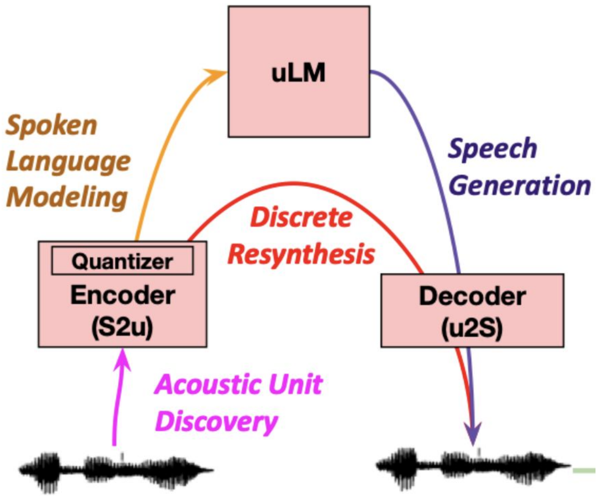
- Apply k-means to obtain discrete tokens (semantic token).
- Train vocoder model (token-to-wave) separately on the semantic tokens.



Examples

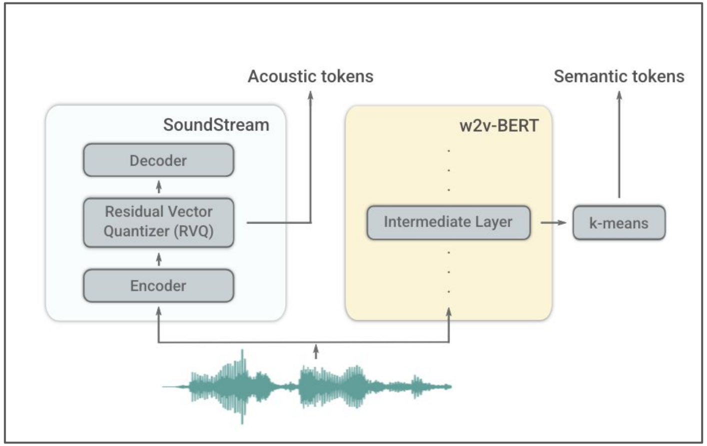
GSLM (Lakhotia 2021)

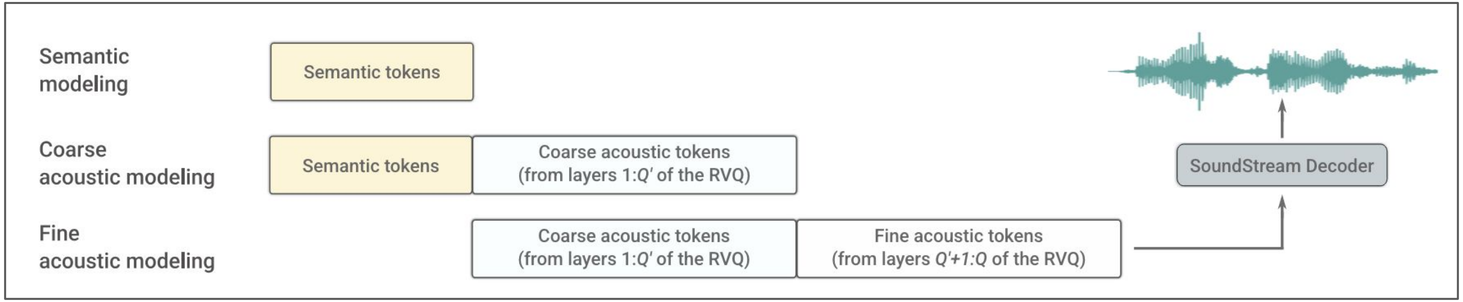
- Generative Spoken Language Modelling (GSLM).
- Very first work attempting LM on raw speech without text.



AudioLM (Borsos 2022)

- Multi-stage autoregressive language modelling of acoustic & semantic tokens.
- 1st model: Semantic tokens.
- 2nd model: coarse acoustic tokens.
- 3rd model: fine acoustic tokens.
- Flatten RVQ code pattern.

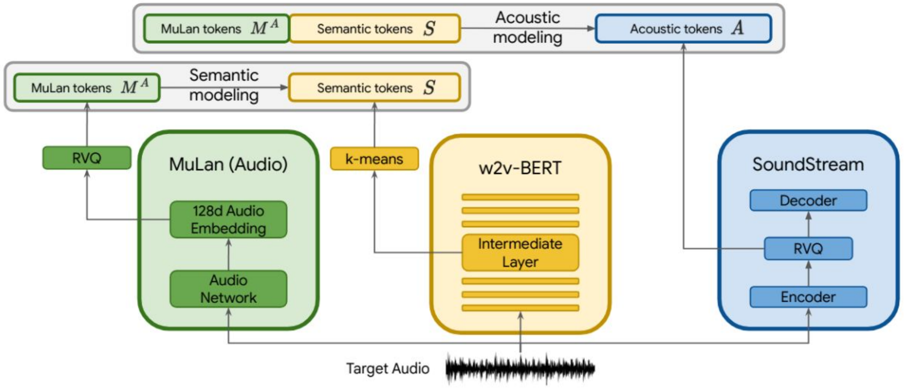




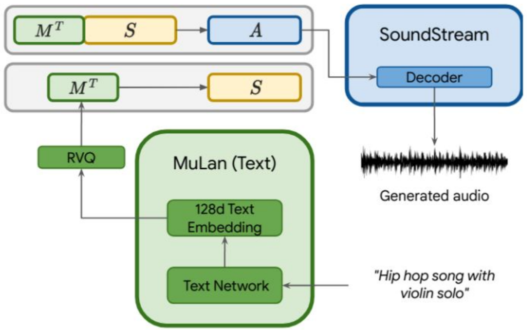
MusicLM (Agostinelli 2022)

- AudioLM + Mulan (music and caption joint embedding model)
- Training: Mulan audio embedding + semantic/acoustic token
- Inference: Mulan text embedding + semantic/acoustic token

Training

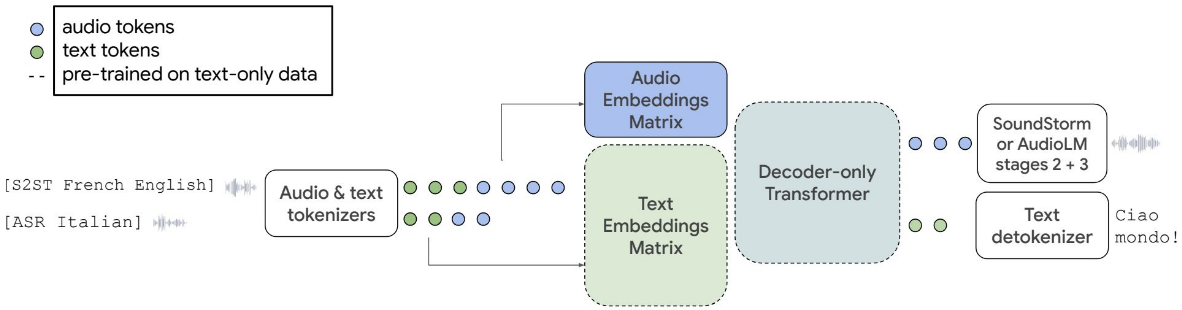


Inference



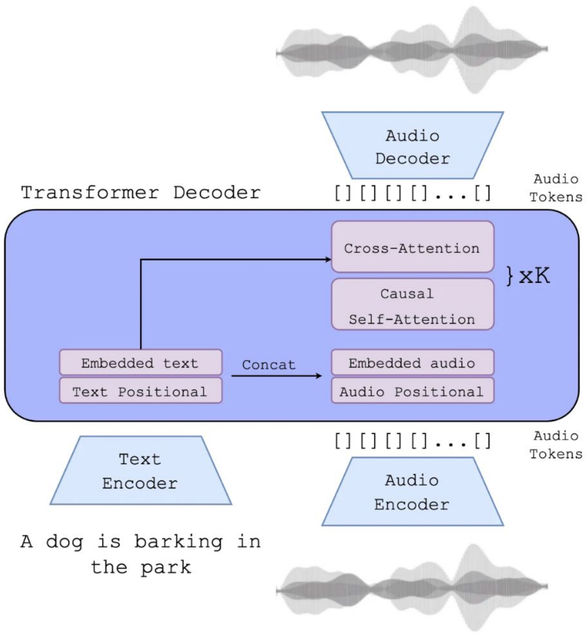
AudioPaLM (Rubenstein 2023)

- Extend the vocabulary of pre-trained LMs to include acoustic tokens.
- Continuous training on audio & text tasks on language modelling.
- Flatten RVQ code pattern.



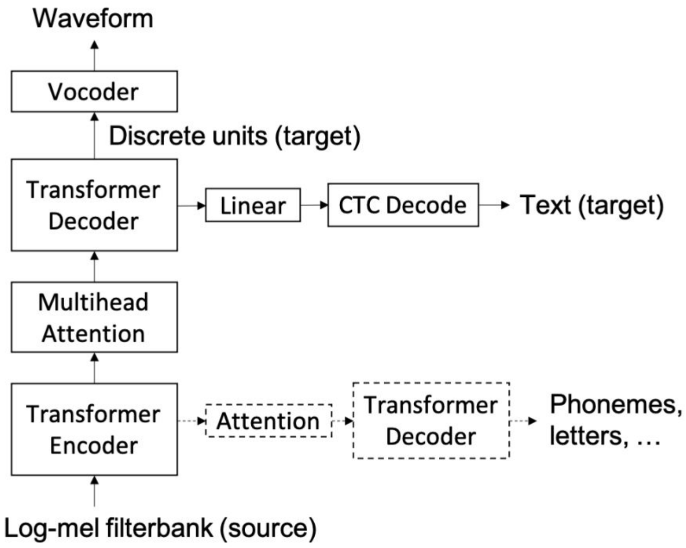
AudioGen (Kreuk 2023)

- Language modelling on acoustic tokens.
- Transformer architecture.
- GAN training (discriminator model).
- Text conditioning by cross attention.
- MusicGen (Copet 2023): Trained on music generation with the similar architecture.



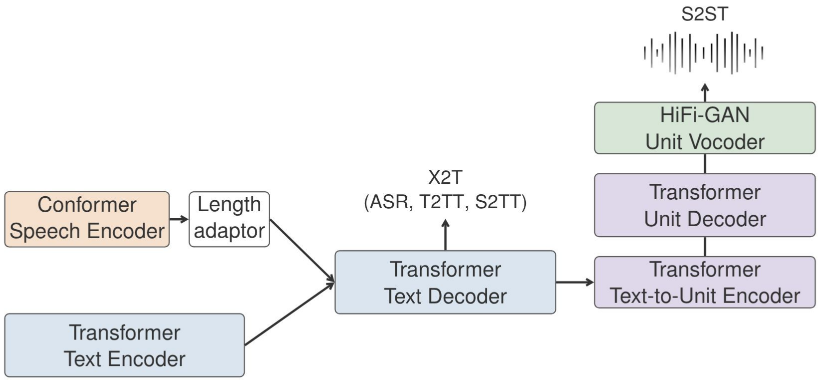
Speech-to-Unit (Lee 2021)

- Direct speech-to-speech translation.
    - Text is used for auxiliary task.
    - HuBERT tokens + Vocoder.
- Textless S2U (Lee 2021): Remove the intermediate auxiliary loss on text.
- pGSLM (Kharitonov 2022): Encodec RVQ tokens with multi-stream transformer.



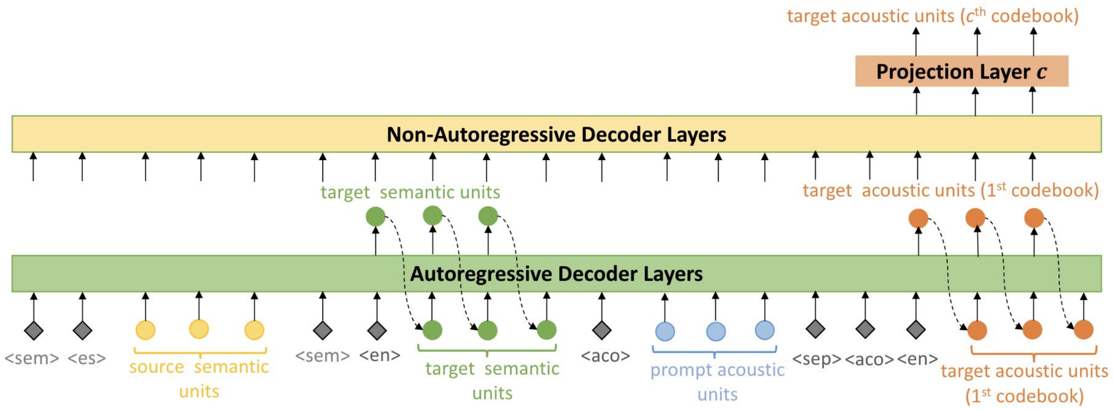
SeamlessM4T (2023)

- Multilingual (100 languages) translation model.
- Multimodal: {text->speech, speech->text, text->text, speech->speech}.
- w2vBERT for input feature (not token).
- XLS-R for output tokens + vocoder.



SeamlessExpressiveLM (Gong 2024)

- Expressive S2S translation model.
- Encodec RVQ tokens with multi-stage models.
- HuBERT semantic tokens to control the characteristics of the output speech.



Future Works

Audio Tokenizer

- Better way to integre of **RVQ codes pattern** into LM.
    - Enable to leverage pre-trained models.
    - Low latency.
- The relationship between semantic and acoustic tokens.
- Are tokens better than embedding?
    - Cross-attention (eg. Flamingo) instead of prompting with audio tokens?
- Joint (Neural Codec) vs Independent (Audio Embedding + Vocoder)
- Better quantization than RVQ
    - Finate Scalar Quantization

Speech Representation

- Many speech embeddings (w2vBERT, XLS-R, HuBERT, etc).
  - Which aspect do they represent…?
    - Pitch, Sentiment, Noise, etc.
- Expressive Speech Generation: controllable speech generation
  - SeemlessExpressiveLM conditions the generation on HuBERT tokens.
- Text-speech joint embedding:
  - LASER (SONAR): Speech and Transcription
  - CLAP: Audio and Caption
  - Speech description
    - Eg.) Female speaking slowly with low tone.

Speech Generation

- Voice-cloning
  - Read transcription in the reference voice.
  - Conditioning generation on the speaker embedding.
- Expressive speech generation.
  - Control the characteristic of the generated speech.
  - Sentiment (sad/happy), pitch (gender, age), speed.
- LM probing studies in NLP for S2S foundation model…?
  - CommonSense, Factuality, Relational Knowledge.

QA