ML 2026 Spring HW10
Spoken Language Model

TA: 陳竣瑋、陳思齊、鄭安妤、尹廷安
ntu-ml-2026-spring-ta@googlegroups.com

Deadline: 2026/**06/18** 23:59:59 (UTC+8)

Outline

- Links
- Task description
- Grading
- Final grade release
- Appendix (Task questions with its options)

Links

- Model-Related Materials:
  - TWIST paper
  - Audio LM paper
  - LLaMA-Omni 2 paper
  - Mimi Model
  - Mimi codec paper
  - Moshi paper
  - GLM-4-Voice paper
- HW Related Materials:
  - ML 2026 Course Website
  - NTU Cool HW10 作業討論區
  - Colab Sample Code
- Extra Materials:
  - 【生成式人工智慧與機器學習導論2025】第 10 講
  - 【生成式AI時代下的機器學習2025】第十二講

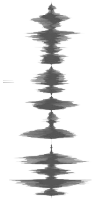
Goals of This Homework

- Section1:
  - Learn different type of Spoken Language Model.
- Section2:
  - Learn how speech can be **represented as tokens** through the **Mimi model**.
  - Learn the initialization, pretrain and interleaving steps in Spoken Language Model.
- For this assignment, you only need to **answer 12 multiple-choice** questions on **NTU COOL**.

Task description

Goal: Speech Generation

Input:
speech sequence

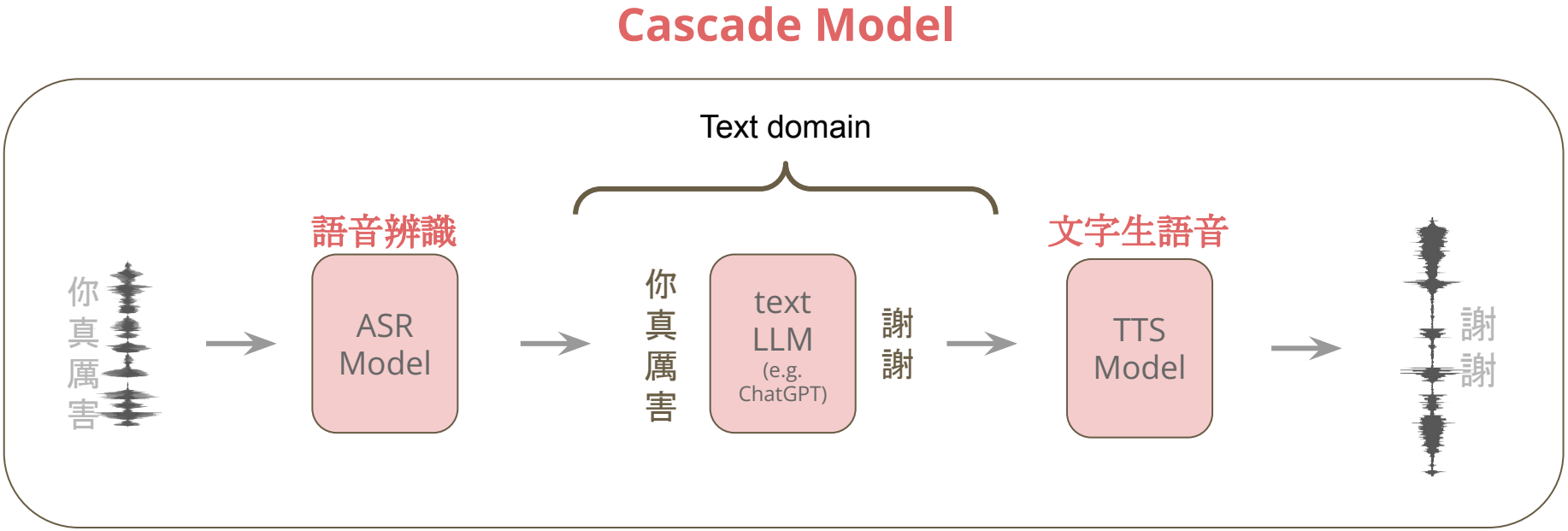


Output:
speech sequence

Section1 - Type comparison

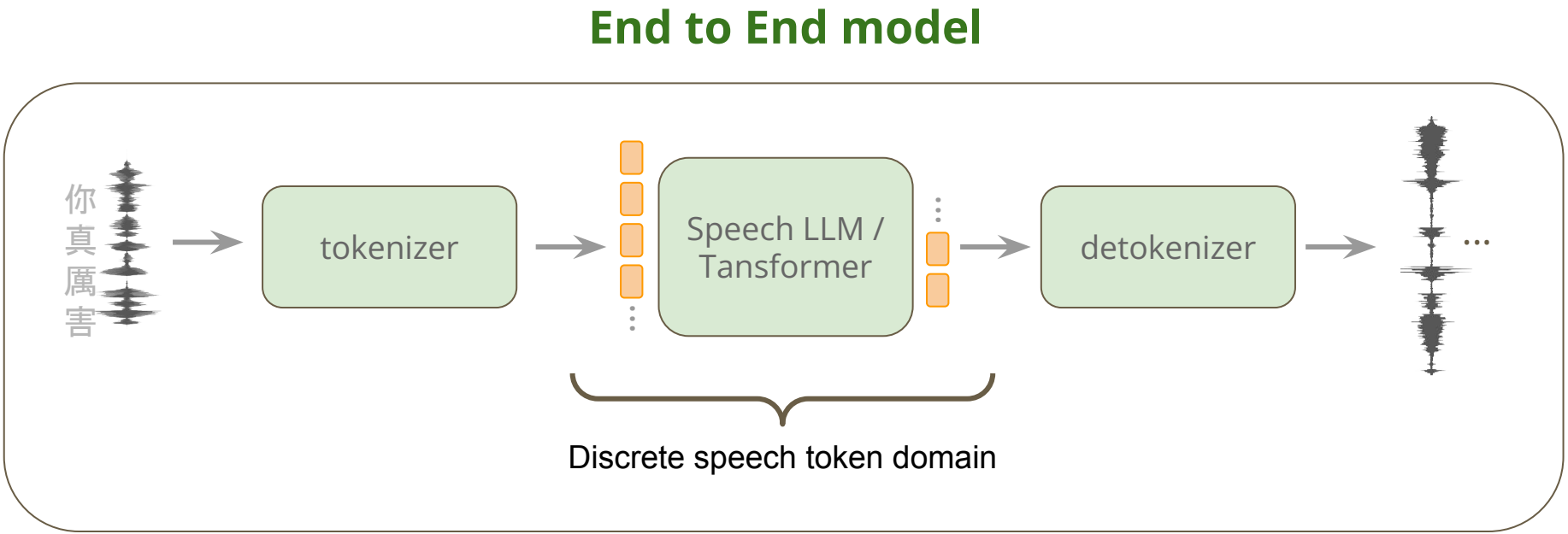
Type 1 - Cascade Model

Speech Generation



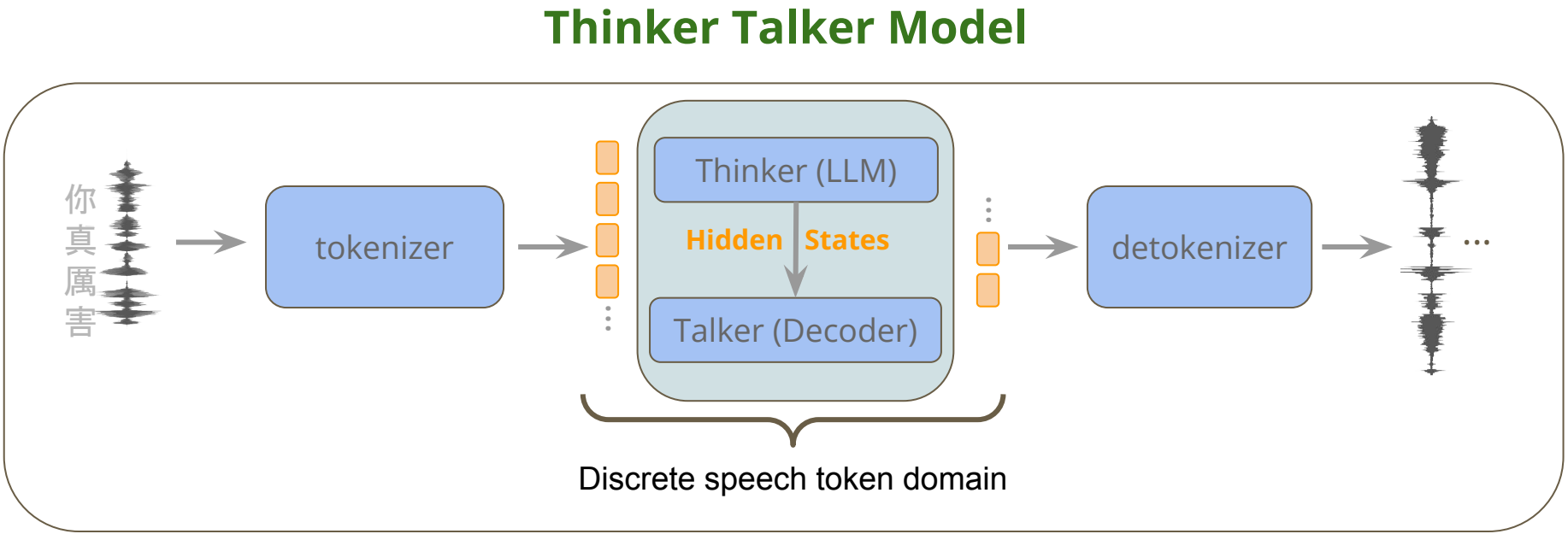
Type2 - End to End model

Speech Generation



Type3 - Thinker Talker Model

Speech Generation



Cascade v.s. Non-cascade

(1%) Q1. Regarding the comparison between cascade speech models and non-cascade speech models, which of the following is correct?

(1%) Q2. The following notebook contains an implementation of a cascade speech model and an end to end speech model, along with ten sample audio files for the models to identify the gender of the speaker. Please answer the following question based on the experimental results. **On Colab**

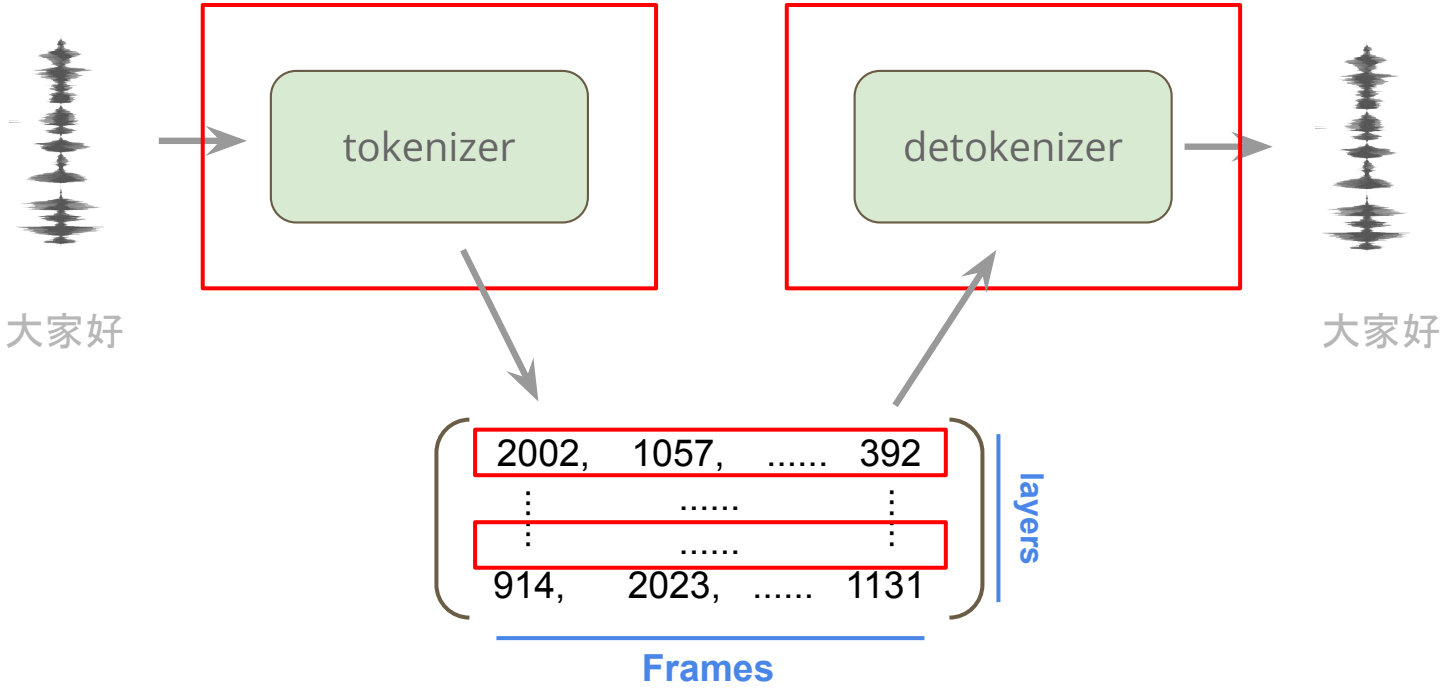
Thinker Talker v.s. End to End

(1%) Q3. Which of the following statements accurately describe the Thinker–Talker architecture of LLaMA-Omni 2

(1%) Q4. Consider LLaMA-Omni 2 as a Thinker–Talker system and Moshi as a representative of end-to-end modeling. Which of the following statements regarding these two approaches are correct?

Section2 - Spoken Language Model Detail

Spoken LM - Tokenizer / Detokenizer



Ref: 【生成式人工智慧與機器學習導論2025】作業十

Spoken LM - Tokenizer

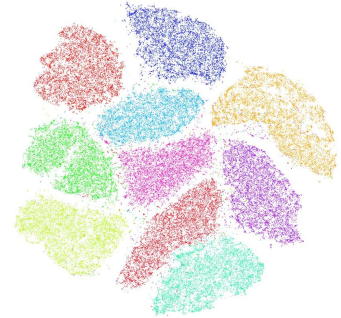
(1%) Q5. Following the procedure on Colab, plot the **UMAP** classification maps of **32 layers** for different **emotion audio** at **layers 0, 6, 16, and 31.** Examine the results and answer the question. **On Colab**

1. Emotion dataset

Use the **EmoV-DB dataset** with predefined emotion categories.

(**Amused**, **Angry**, **Disgusted**, **Sleepy**, **Neutral**)

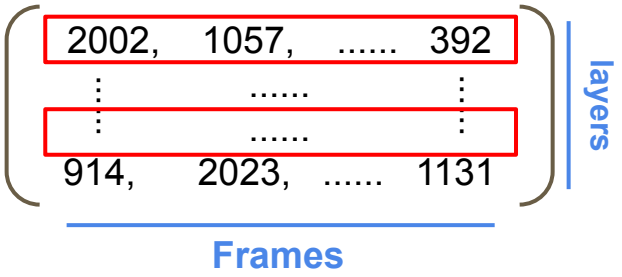
https://www.openslr.org/115/



Mimi tokenizer

UMAP

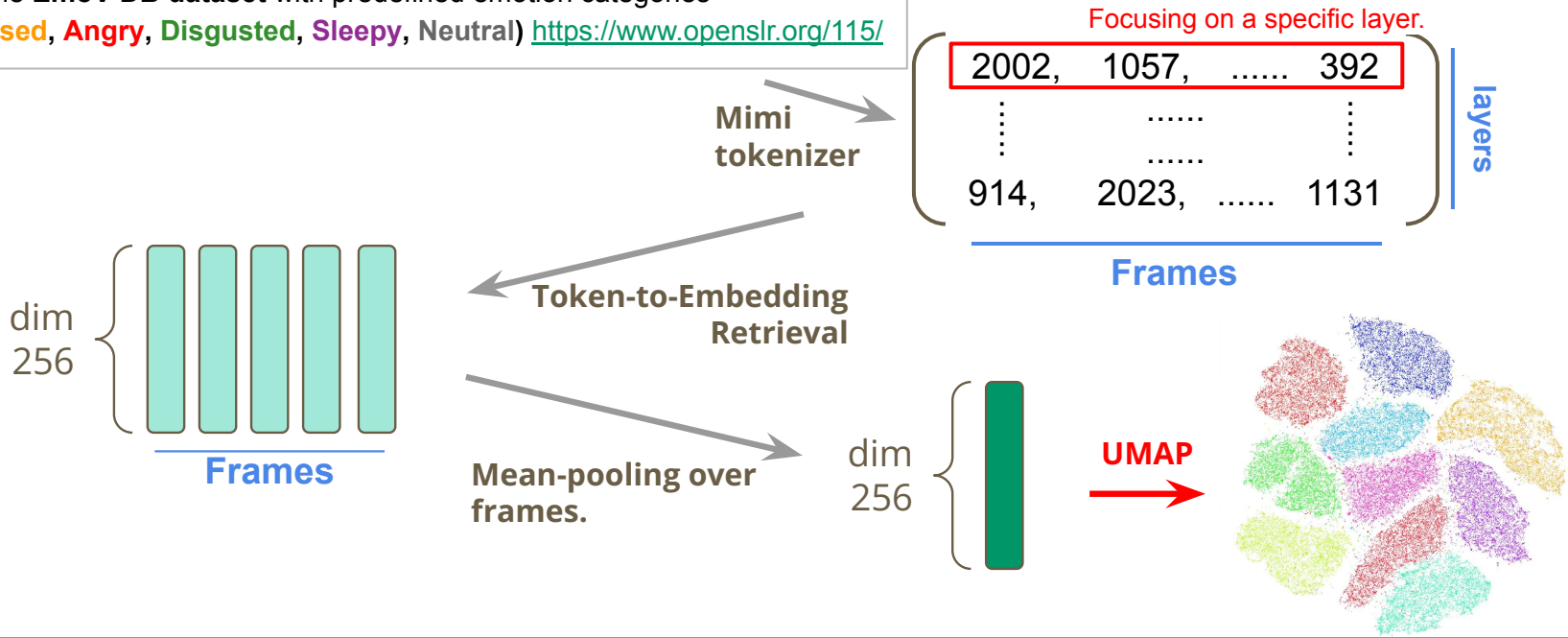
Focusing on layers 0, 6, 16, and 31.



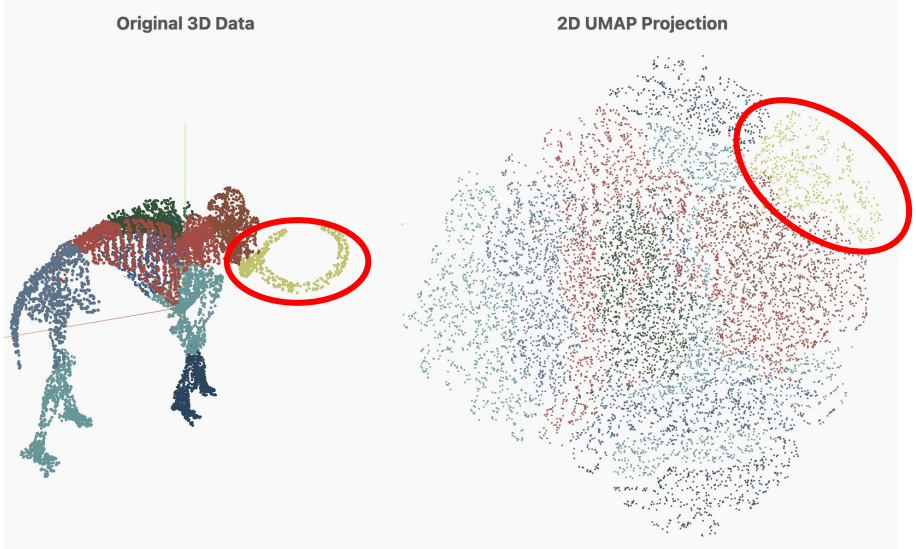
Data Preparation - One Audio Sample, Specific Layer

1. Emotion dataset

Use the **EmoV-DB dataset** with predefined emotion categories (**Amused**, **Angry**, **Disgusted**, **Sleepy**, **Neutral**) https://www.openslr.org/115/



What is UMAP?



UMAP is a **nonlinear dimensionality reduction** tool that is faster than t-SNE and better preserves global structure.
(Ref: Understanding UMAP)

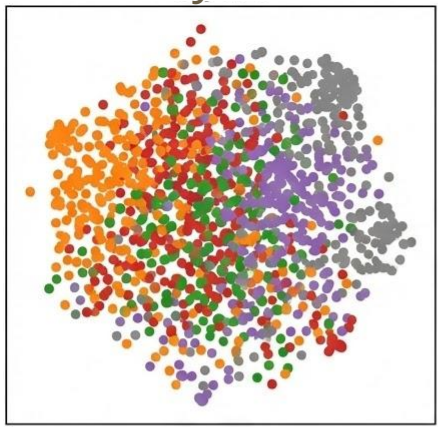
The core goal of UMAP is to **project** complex, high-dimensional data into a **lower-dimensional space** while preserving the essential **neighborhood structure**.

Interpretation

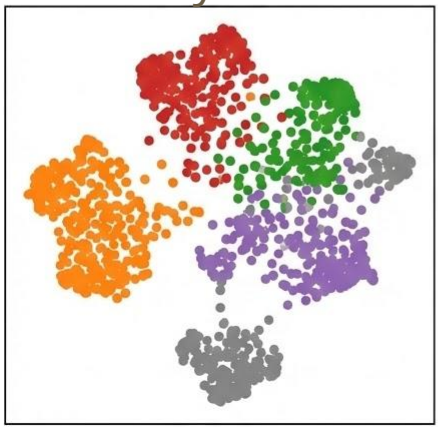
- Each 2D **UMAP figure** visualizes all data points from **a particular layer**.
- Every **point** within the plot is **traceable** to a **specific audio file**.

(**Amused**, **Angry**, **Disgusted**, **Sleepy**, **Neutral**)

Layer *X*

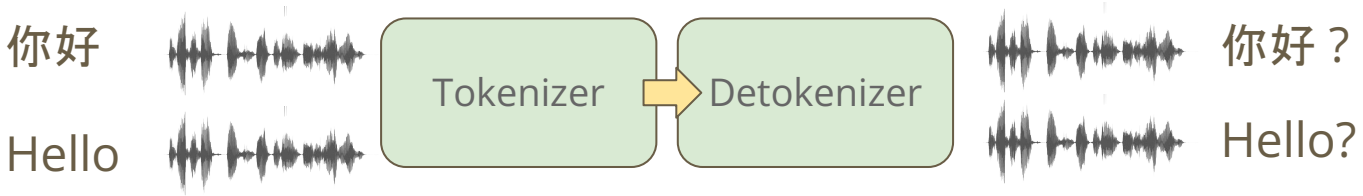


Layer *Y*

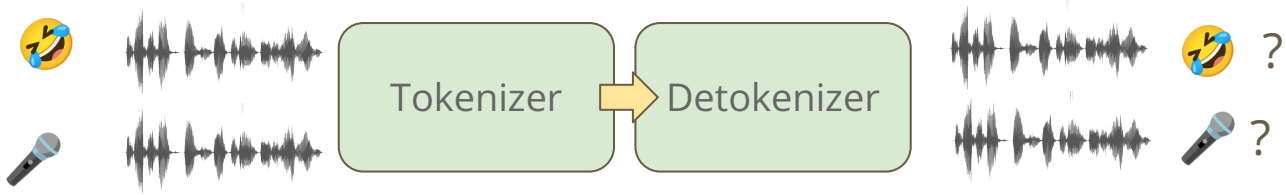


Spoken LM - Detokenizer

(0.5%) Q6. Compare TTS_English_speech.wav and TTS_Chinese_speech.wav; which of the following statements is correct? **On Colab**

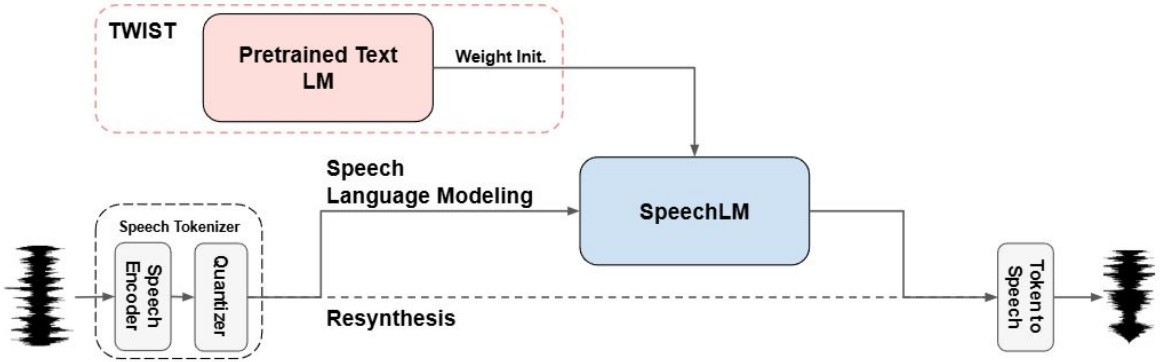


(0.5%) Q7. Compare laughter.wav and music.wav: which audio file yields the worst results after decoding, and what might this signify? **On Colab**



Spoken LM - Initialization / Pretrain / Interleaving

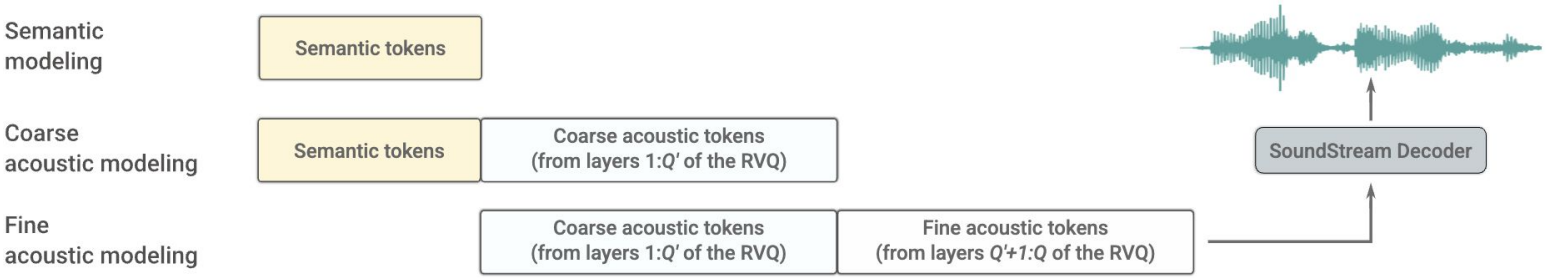
(0.5%) Q8. Regarding the specific methodology and experimental results of TWIST (Text-pretrained Weight Initialization for Speech Transformer), which of the following statements are correct?



ref: Textually Pretrained Speech Language Models

Spoken LM - Initialization / Pretrain / Interleaving

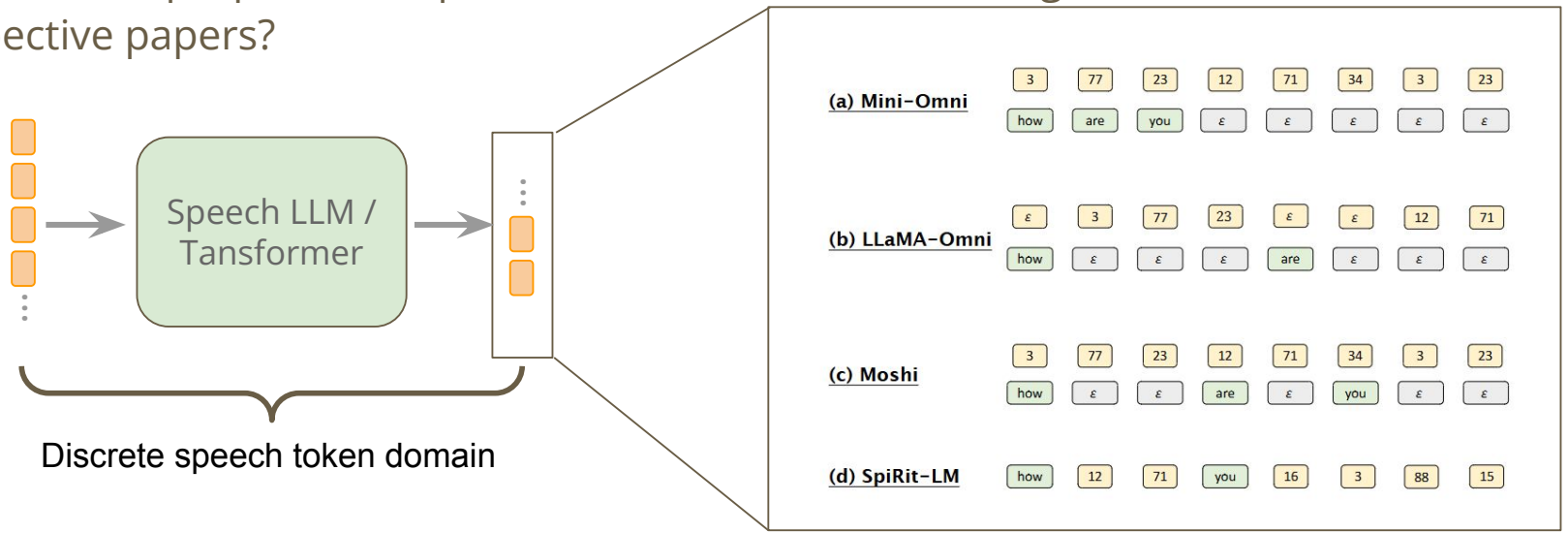
Q9. In Spoken Language Models, audio signals are typically compressed into tokens before being processed by LLM architectures. Which of the following statements accurately describe the common types of discrete speech tokens and their roles in modern SpokenLM pretraining?



ref: [AudioLM: a Language Modeling Approach to Audio Generation]

Spoken LM - Initialization / Pretrain / Interleaving

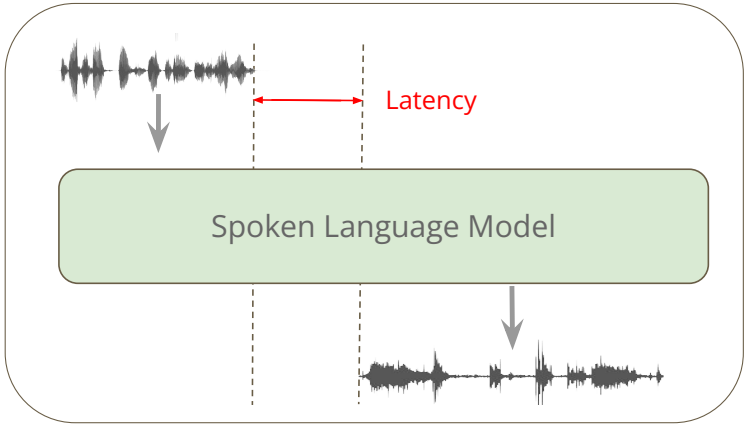
(1%) Q10. In the architectures of Moshi and GLM-4-Voice, the model interleaves discrete text tokens with acoustic (audio) tokens. Which of the following statements correctly describe the purpose or implementation of this interleaving method as detailed in their respective papers?

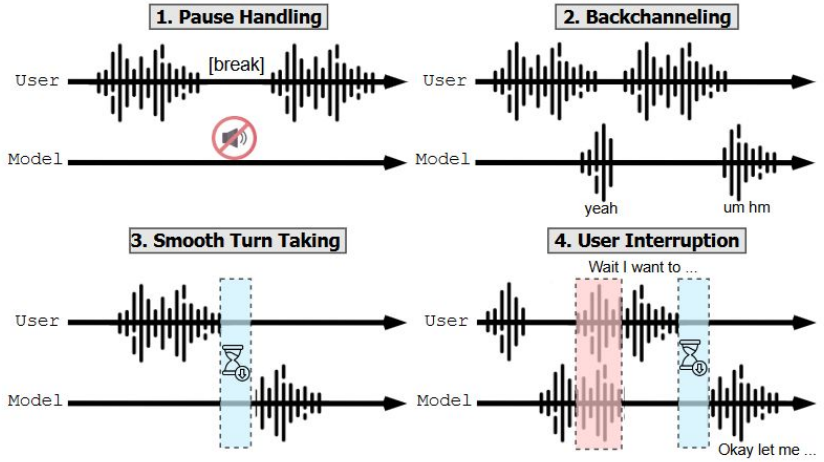


ref: On The Landscape of Spoken Language Models

Spoken LM - Realtime / Full-Duplex

Realtime: Short output latency
Full-Duplex: Listen to voice and talk back at the exact same time.





ref: Full-Duplex-Bench

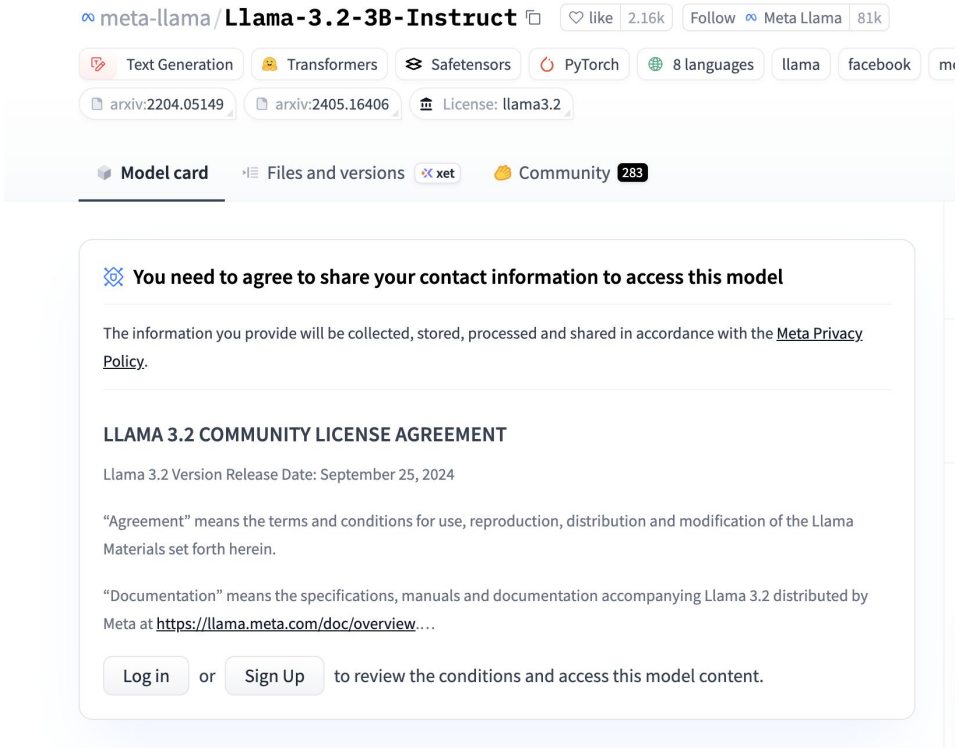
Spoken LM - Realtime

(1%) Q11. Which of the following technical implementations are responsible for Moshi’s low-latency, real-time performance?

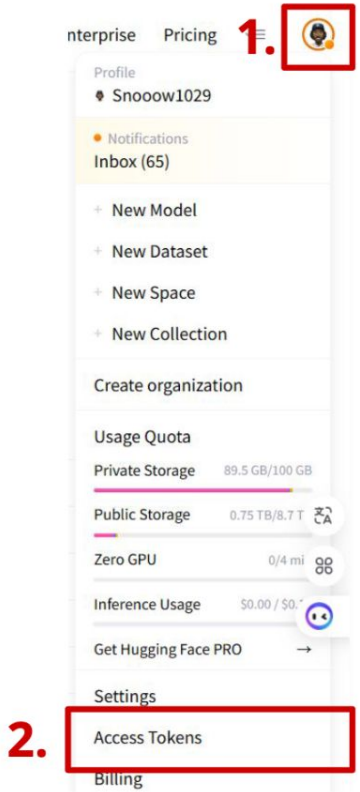
(1%) Q12. In the Moshi architecture, which strategies specifically enable "Full-Duplex" behavior where the model and user can speak simultaneously?

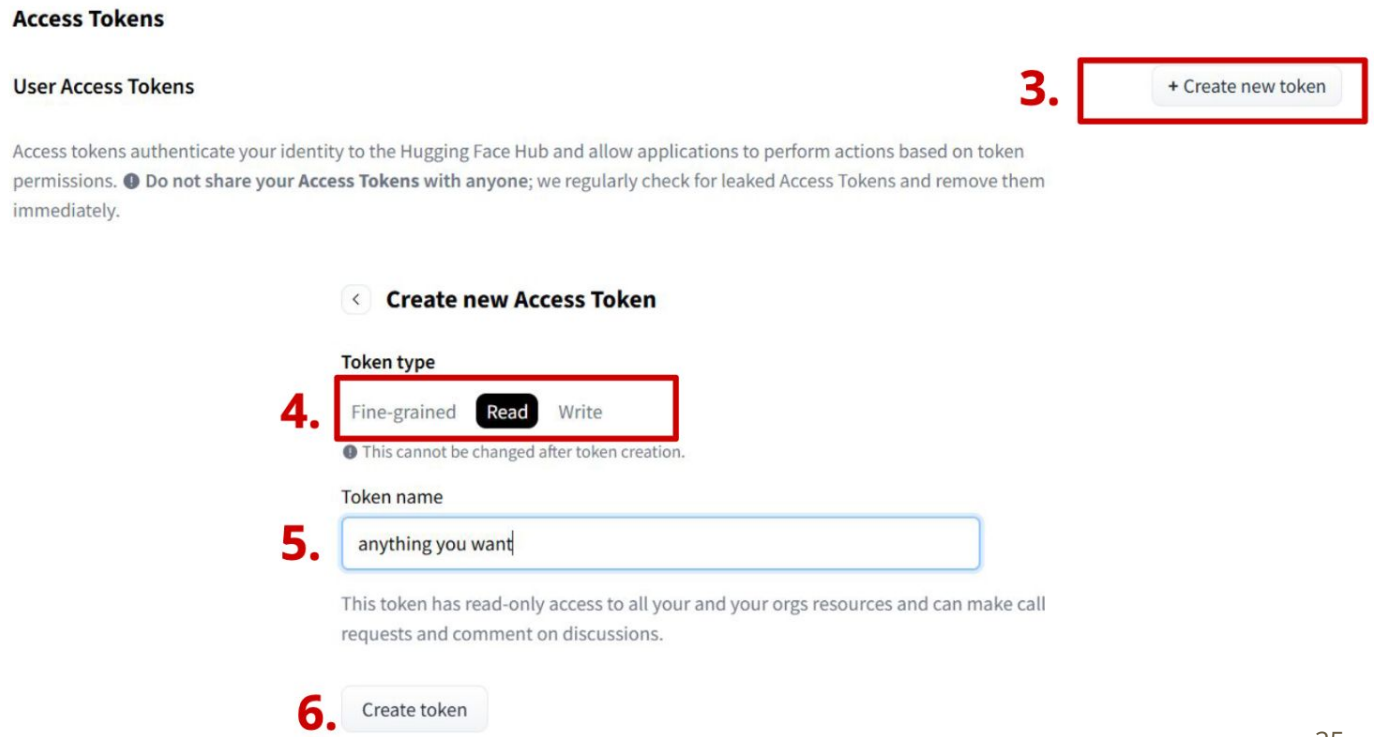
Colab - Llama-3.2-3B license

meta-llama/Llama-3.2-3B-Instruct



Colab - HF token





Submission & Deadline

- Submit your homework to **NTU Cool**
- 2026/**06/18** 23:59:59 (UTC+8)
- No late submission is allowed

Grading Release Date

- The grading of the homework will be released by 2026/**06/19** 23:59:59 (UTC+8)
- The grade revision will end before 2026/**06/21** 23:59:59 (UTC+8)

Final grade release

- The final grading of this course will be released before 2026/**06/22** 23:59:59 (UTC+8)

Grading - Regulations

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- Do NOT search for or use additional data for training or the answers for the testing data.
- Do NOT use closed-source LLM APIs like GPT-4, Gemini, etc.
- You should NOT modify your input file or prediction files manually.
- Make sure that TAs can reproduce the predictions using the code you submit. (Fix the random
- seed)
- Your final grade x 0.9 and get a score 0 for that homework if you violate any of the above rules
- first time (within a semester).
- Your will get F for the final grade if you violate any of the above rules multiple (> 1) times.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

If You Have Any Questions

- NTU Cool **HW10** 作業討論區
  - 如果同學的問題不涉及作業答案或隱私，請**一律使用** NTU Cool 討論區
  - 助教們會優先回答NTU Cool討論區上的問題
- Email: ntu-ml-2026-spring-ta@googlegroups.com
  - Title should start with [GenAI-ML 2026 Fall **HW10**]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Each Friday (5/29, 6/5) before / after class:
    - (Fri.) 13.20 ~ 14.10 / 17:30~18:00
    - Location: 博理112
  - Google meet (6/12) : Link
    - (Fri.) 13.20 ~ 14.10 / 17:30~18:00

Appendix. (Task questions with its options)

(1)

Regarding the comparison between cascade speech models and non-cascade speech models, which of the following is correct? Moshi paper

- End to end model using ASR and TTS for processing the speech input and output.
- Unlike cascade models, where ASR relies on VAD (voice activity detection) for turn-taking, Moshi achieves more natural turn-taking by simultaneously maintaining a user audio stream and a model audio stream allowing the model to perceive speech states in real time.
- The inability of cascade models to understand paralinguistic information such as emotion and accent is due to the inherent limitations of the LLM, which can only process semantic content, rather than insufficient information being provided by ASR model.
- In cascade models, text is an intermediate output that user speech must be transcribed into before LLM processing.
- Non-cascade models and cascade models adopt a single-stage training strategy.

Appendix. (Task questions with its options)

(2)

The following notebook contains an implementation of a cascade speech model and non-cascade spoken language model, along with three sample audio files for the models to identify the gender of the speaker (2 females, 1 male). Please answer the following question based on the experimental results.

- Model A is the cascade model, because it justify itself as textbased AI
- Model A is the cascade model, because it can correctly identify the gender of the speaker from the audio.
- Model A is the cascade model, because it use acoustic evidence to classify the gender correctly.
- Model B is the cascade model, because it use text evidence to classify the gender correctly.
- Model B is the non-cascade model, because it can correctly identify the gender of the speaker from the audio.

Appendix. (Task questions with its options)

(3)

Which of the following statements accurately describe the Thinker-Talker architecture of LLaMA-Omni 2? LLaMA-Omni 2 paper

- LLaMA-Omni 2 keep a pretrained LLM with existing language abilities as the core, and extend it to speech interaction tasks by adding extra speech modules.
- When processing speech input, LLaMA-Omni 2 first use modules such as a speech encoder or adaptor to convert audio into representations that the LLM can use.
- The speech output from Talker is usually guided by the text content or representations produced by the Thinker.
- Since the model involves both text and speech, it must place text tokens and speech tokens into the same interleaved sequence and model them jointly.
- The main goal of this architecture is to replace the original LLM with an audio-only model, so that the system no longer relies on text or linguistic representations.

Appendix. (Task questions with its options)

(4)

Consider LLaMA-Omni 2 as a Thinker-Talker system and Moshi as a representative of end-to-end modeling. Which of the following statements regarding these two approaches are correct? LLaMA-Omni 2 paper, Moshi paper

- Thinker-Talker architectures usually separate “language understanding and reasoning” from “speech generation” into different modules: Thinker is responsible for reasoning, while Talker is responsible for producing speech tokens.
- In Thinker-Talker models, Talker can be regarded as a TTS-like module, but it is usually not limited to conventional text-to-speech and may also use LLM hidden states or other intermediate representations.
- In Moshi’s multi-stream architecture models, the user speech stream and the model speech stream separately, allowing it to handle overlapping speech, interruptions, and backchanneling.
- Because Moshi place text and speech tokens in the same sequence, they must finish generating the entire text response before starting speech generation.
- Compared with Thinker-Talker models, Moshi more naturally supports full-duplex dialogue because it models the user speech stream and the model speech stream separately, so it does not rely on clearly segmented speaker turns and can keep track of both sides of the conversation while the dialogue is ongoing.

Appendix. (Task questions with its options)

(5)

Following the procedure on Colab, plot the UMAP classification maps of 32 layers for different emotion audio at layers 0, 6, 16, and 31. Examine the results and answer the question.

- UMAP is a non-linear dimensionality reduction and visualization tool.
- In UMAP, points of the same color represent audio files labeled with the same emotion category.
- Comparing Layer 0 and Layer 31, the higher the layer number, the higher the cluster separation of the embeddings across emotion categories.
- Comparing Layer 0 and Layer 31, the lower the layer number, the higher the cluster separation of the embeddings across emotion categories.
- A close observation of the Layer 6 scatter plot reveals that this layer can roughly cluster "Sleepy" and "Angry" separately compared than Layer 16.

Appendix. (Task questions with its options)

(6)

Compare TTS_English_speech.wav and TTS_Chinese_speech.wav; which of the following statements is correct?

- Chinese TTS performance is significantly superior to English, indicating that Mimi has a greater advantage in processing Chinese speech.
- Neither English nor Chinese TTS can be effectively reconstructed, suggesting that Mimi is not suited for cross-lingual speech tasks.
- The reconstruction results for both are excellent, demonstrating that Mimi is capable of handling both Chinese and English audio information simultaneously.
- English TTS performance is significantly superior to Chinese, indicating that Mimi is primarily suitable for English speech.

Appendix. (Task questions with its options)

(7)

Compare laughter.wav and music.wav: which audio file yields the worst results after decoding, and what might this signify?

- laughter.wav shows the poorest reconstruction performance, indicating that Mimi is unable to recognize paralinguistic signals related to emotion.
- laughter.wav shows the poorest reconstruction performance, suggesting that Mimi is only suitable for processing speech and is ill-suited for non-verbal sounds.
- music.wav shows the poorest reconstruction performance, indicating that if Mimi's tokens are utilized by downstream models, they may struggle to handle music.
- music.wav shows the poorest reconstruction performance, showing that the model is particularly sensitive to long-duration audio, making it difficult to reconstruct music.

Appendix. (Task questions with its options)

(8)

Regarding the specific methodology and experimental results of TWIST (Text-pretrained Weight Initialization for Speech Transformer), which of the following statements are correct? TWIST paper

- Freezing all hidden layers of the network and only training new speech embeddings. Experiments show that the performance of this approach is inferior to a cold-start model (Cold-Init) trained from scratch.
- Replacing text embeddings with randomly initialized speech embeddings while retaining the remaining backbone weights for continued training. Experiments show that it requires only 10% of the data to match the performance of a cold-start model using 100% of the data.
- This cross-modal transfer is not limited to text; experiments confirmed that using an image-pretrained model (such as ImageGPT) for initialization yields equivalent performance improvements.
- Completely retaining the original text embedding table for training. Experiments also indicated that the best performance is achieved when paired with speech tokens at a higher sampling frequency (e.g., 50Hz).
- In addition to improving final performance, text model initialization significantly accelerates training convergence. Experiments show that the number of training update steps TWIST needs to reach the same perplexity as a cold-start model is only about one-quarter of that required by the cold-start model.

Appendix. (Task questions with its options)

(9)

In Spoken Language Models, audio signals are typically compressed into tokens before being processed by LLM architectures. Which of the following statements accurately describe the common types of discrete speech tokens and their roles in modern SpokenLM pretraining? Audio LM paper

- Semantic tokens (derived from self-supervised masked language models like w2v-BERT) are primarily utilized to capture long-term structural coherence.
- Acoustic tokens (produced by neural audio codecs like SoundStream) are primarily used to capture the details of the audio waveform, including speaker identity and prosody.
- Modern SpokenLMs completely bypass the need for audio-derived tokens by converting all speech directly into character-level text tokens before pretraining.
- A generative framework can employ a hierarchical approach, where the model first uses semantic tokens to enable long-term structural coherence, and then models acoustic tokens conditioned on those semantic tokens to achieve high-quality audio synthesis.
- Semantic tokens are extracted using traditional, rule-based linguistic software that manually applies grammatical rules.

Appendix. (Task questions with its options)

(10)

In the architectures of Moshi and GLM-4-Voice, the model interleaves discrete text tokens with acoustic (audio) tokens. Which of the following statements correctly describe the purpose or implementation of this interleaving method as detailed in their respective papers? Moshi paper GLM4-Voice paper

- GLM-4-Voice utilizes a unified token sequence where text and speech tokens are interleaved, which alternate between generating 13 text tokens and 26 speech tokens
- Interleaving in both model is used primarily to reduce the sequence length of the audio, as text tokens require significantly less memory than any acoustic codebook.
- GLM-4-Voice uses synthetic interleaved data (replacing text spans with speech tokens) to bridge the data gap and align the high-intelligence text modality with the speech modality.
- In Moshi, text tokens are interleaved only at the end of a speaker's turn to provide a summary of the conversation, rather than being predicted alongside audio in real-time.
- Moshi inserts special PAD tokens into the text stream so the audio and text stay time-aligned.

Appendix. (Task questions with its options)

(11)

Which of the following technical implementations are responsible for Moshi's low-latency, real-time performance? Moshi paper

- Using a 12.5 Hz frame rate in the Mimi codec, which results in a 80 ms algorithmic delay.
- Moshi uses a small Depth Transformer to handle all RVQ codebook layers per timestep, offloading that work from the large Temporal Transformer.
- The elimination of the traditional ASR-LLM-TTS cascade, replacing it with a single, end-to-end causal model that processes audio tokens directly.
- Moshi process 10-second "chunks" of audio and then generate a response for each 10 seconds.
- Moshi process uncompressed 44.1 kHz waveform for accurate audio generation.

Appendix. (Task questions with its options)

(12)

In the Moshi architecture, which strategies specifically enable "Full-Duplex" behavior where the model and user can speak simultaneously? Moshi paper

- Use Voice Activity Detection to detect user interruptions and instantly clear the model's output buffer to reset the conversation state.
- The model treats the dialogue as a joint modeling task of three parallel streams: User audio, Assistant audio, and Assistant text.
- A Dedicated Turn-Taking Token (<TT>) inserted into the text stream to signal the model when it is mathematically safe to transition from listening to speaking.
- A volume-sensing interrupt trigger that uses traditional signal processing to measure the user’s decibel levels and force the AI to stop talking if the user gets too loud.
- The implementation of an autoregressive Temporal Transformer that processes interleaved tokens from both the user’s continuous audio stream and the assistant’s output streams frame-by-frame, allowing it to continuously react to the user without waiting for a turn.