Generative AI for Audio Generation

COMP4431 Lab 3

Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Significance
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation

Concepts and Significance

- Concepts
  - Audio generation aims to synthesize various forms of natural and intelligible sound.
    - Speech (Emotion, Pace, Gender, etc.)
    - Music (Genre, Rhythm, Instruments, etc.)
    - Sound Effect (Natural, Human-made objects, Animal, etc.)
    - Other (Imaginary sound, compositional sound)

Božić M, Horvat M. A survey of deep learning audio generation methods[J]. arXiv preprint arXiv:2406.00146, 2024.
Zhao Y, Xia X, Togneri R. Applications of deep learning to audio generation[J]. IEEE Circuits and Systems Magazine, 2019, 19(4): 19-38.

Concepts and Significance

- Significance
  - Creativity
    - Generate non-exist sound (e.g., Half cat Half sheep sound)
    - Inspire the content creation
  - Efficiency
    - No need for retrieval
    - Endless audio samples
    - Fine-grained control of sound (Emotion, pitch, materials, etc.)

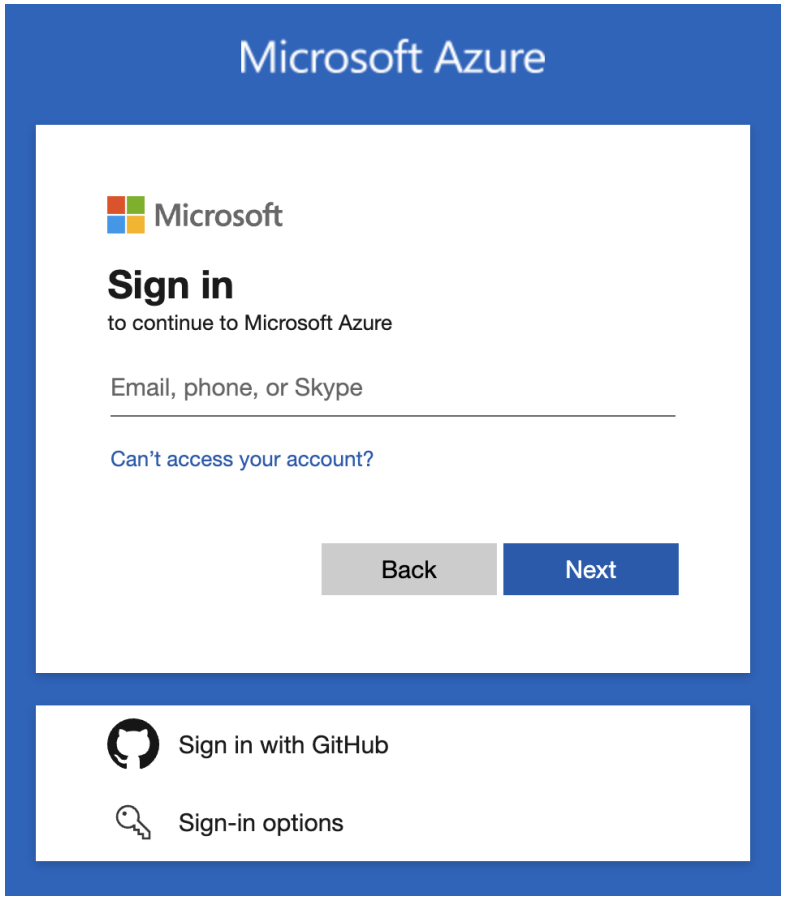
Liu H, Chen Z, Yuan Y, et al. AudioLDM: text-to-audio generation with latent diffusion models[C]//Proceedings of the International Conference on Machine Learning. 2023: 21450-21474.
Božić M, Horvat M. A survey of deep learning audio generation methods[J]. arXiv preprint arXiv:2406.00146, 2024.
Zhao Y, Xia X, Togneri R. Applications of deep learning to audio generation[J]. IEEE Circuits and Systems Magazine, 2019, 19(4): 19-38.

Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Significance
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation

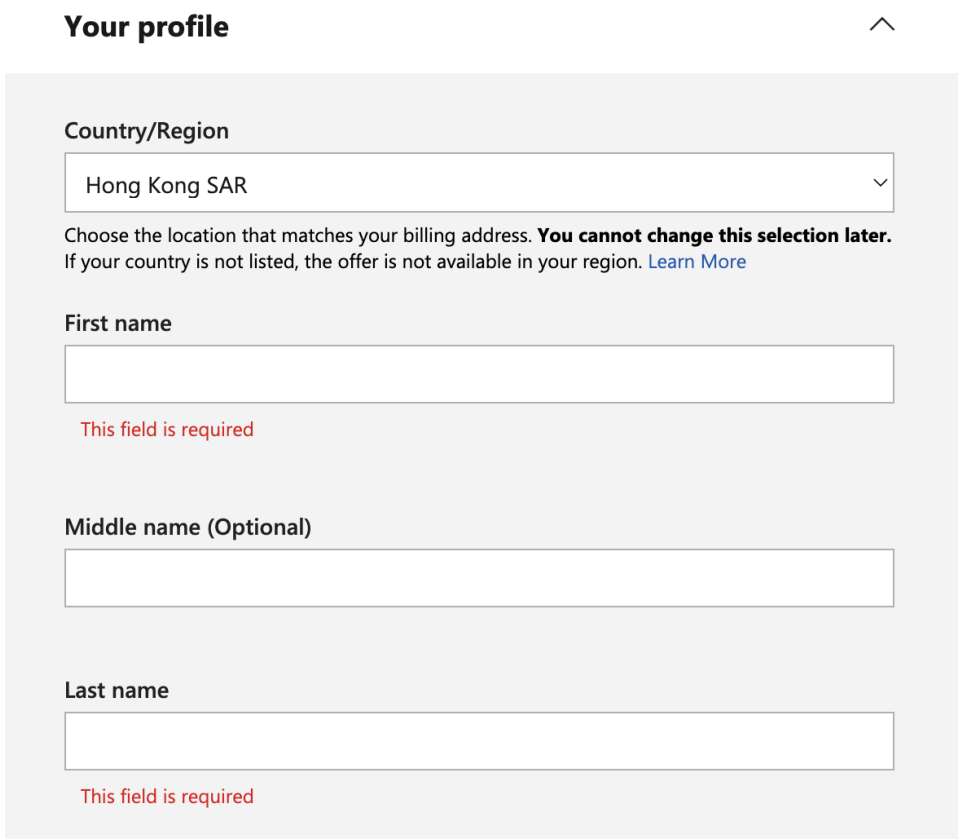
Access to Text-to-speech Generation Models

- Azure
  - Microsoft Azure is the cloud computing platform developed by Microsoft
  - You can build, run, and manage applications with the tools and frameworks of your choice
- Azure AI Speech
  - AI Speech from Microsoft Azure includes speech recognition, text-to-speech, speech translation, voice-enabled app features, and more.
- Access to Azure AI Speech
  - Navigate to https://portal.azure.com/
  - Log in using your Microsoft account
  - Create your Azure free account
  - Sign up for a Pay-As-You-Go subscription
  - Create speech services
  - Navigate to https://speech.microsoft.com/portal



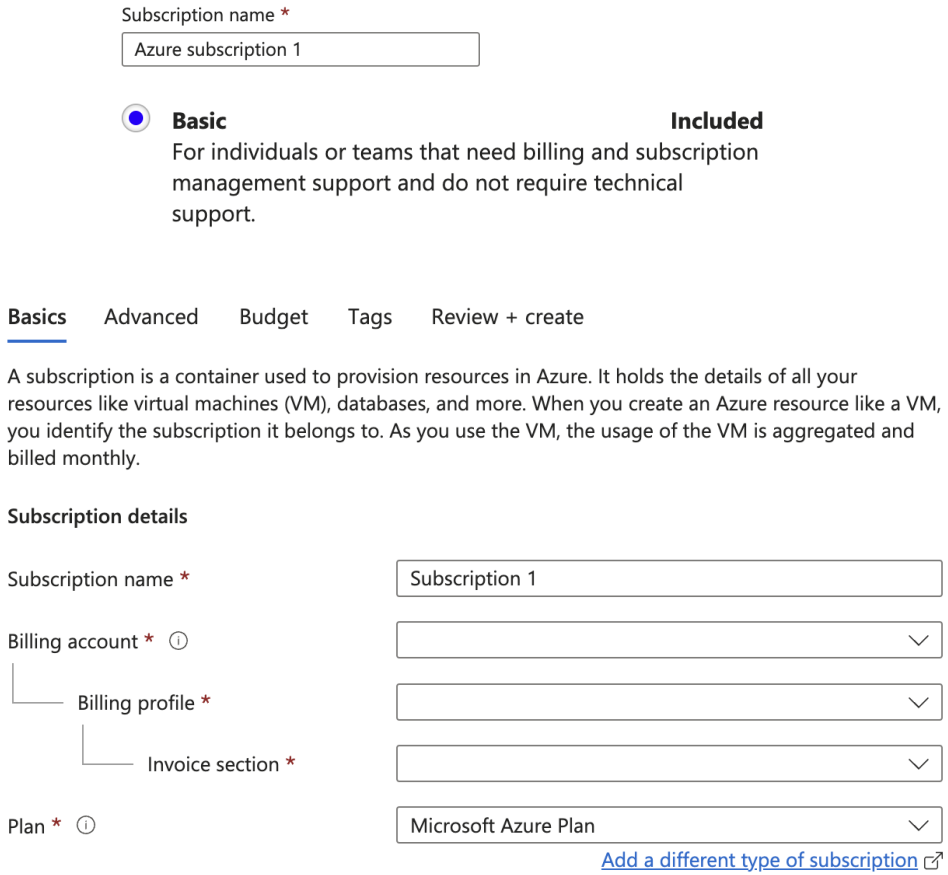
Access to Text-to-speech Generation Models

- Access to Azure AI Speech
  - Navigate to https://portal.azure.com/
  - Log in using your Microsoft account
  - Create your Azure free account
    - Popular services are free for 12 months
    - $200 credit to use in your first 30 days
  - Sign up for a Pay-As-You-Go subscription
  - Create speech services
  - Navigate to https://speech.microsoft.com/portal



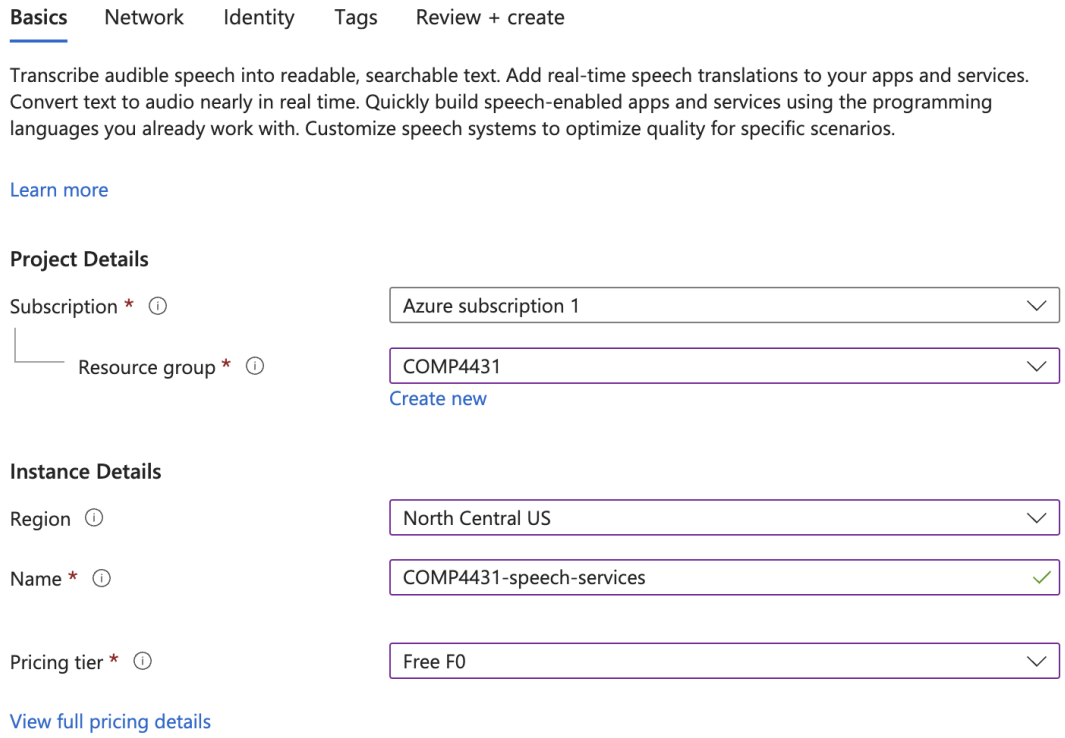
Access to Text-to-speech Generation Models

- Access to Azure AI Speech
  - Navigate to https://portal.azure.com/
  - Log in using your Microsoft account
  - Create your Azure free account
  - Sign up for a Pay-As-You-Go subscription
    - Navigate to https://portal.azure.com/#view/Microsoft_Azure_SubscriptionManagement/SubscriptionCreateBlade
    - Upgrade to a basic Pay-As-You-Go account
    - Create a subscription with default settings
  - Create speech services



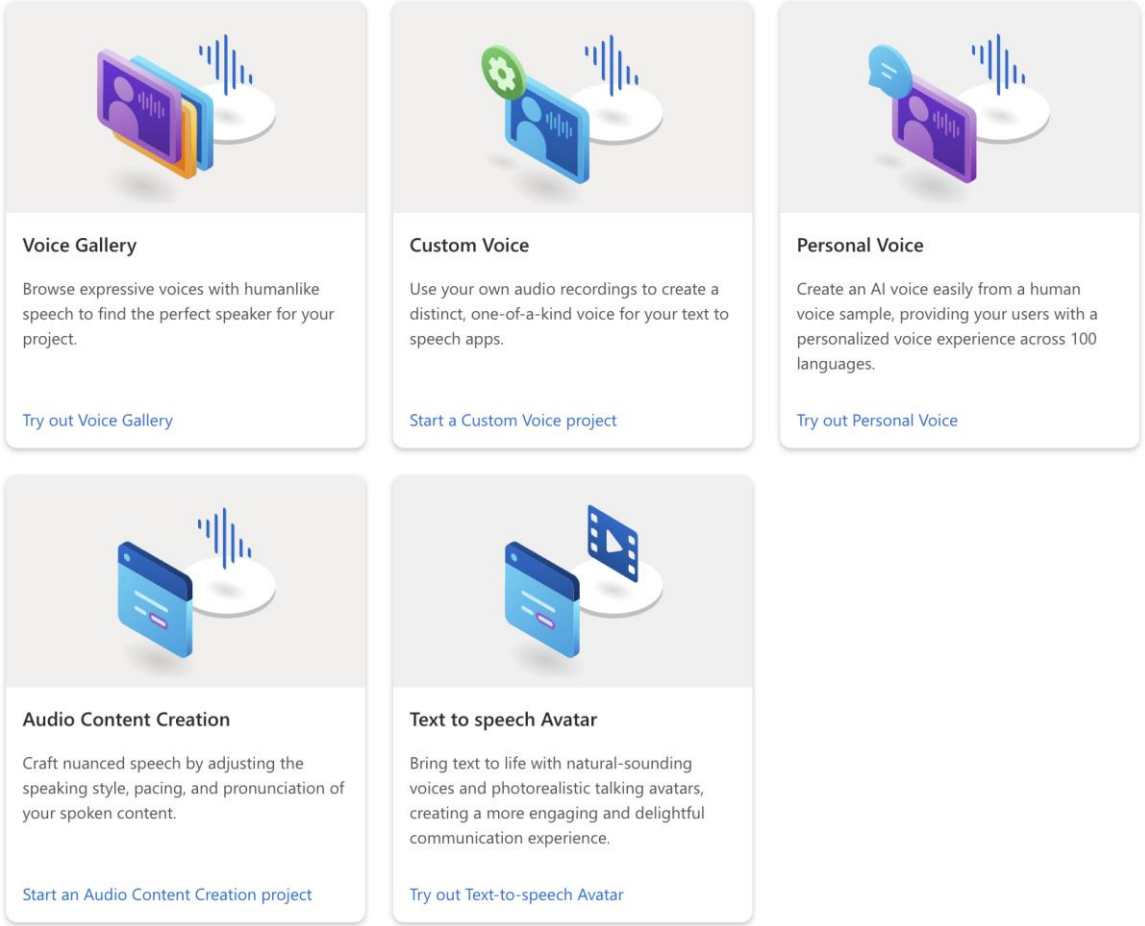
Access to Text-to-speech Generation Models

- Access to Azure AI Speech
  - Navigate to https://portal.azure.com/
  - Log in using your Microsoft account
  - Create your Azure free account
  - Sign up for a Pay-As-You-Go subscription
  - Create speech services
    - Navigate to https://portal.azure.com/#create/Microsoft.CognitiveServicesSpeechServices
    - Set a subscription
    - Set a resource group
    - Set region to North Central US
      - OpenAI text-to-speech voices are available in North Central US and Sweden Central.
    - Set the name of your resource
    - Set pricing tier
      - Free F0 (0.5 million characters free per month)
      - Standard S0 ($200 credit given, 15 per 1M characters)
  - Navigate to https://speech.microsoft.com/portal



Access to Text-to-speech Generation Models

- Access to Azure AI Speech
  - Navigate to https://portal.azure.com/
  - Log in using your Microsoft account
  - Create your Azure free account
  - Sign up for a Pay-As-You-Go subscription
  - Create speech services
  - Navigate to https://speech.microsoft.com/portal
    - Select audio content creation



Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Significance
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation

Step-by-step Guide to Text-to-speech Generation

- Set the text file
  - New a text file, and enter or past text

    A: This can't be right. The algorithm wasn't designed for this.

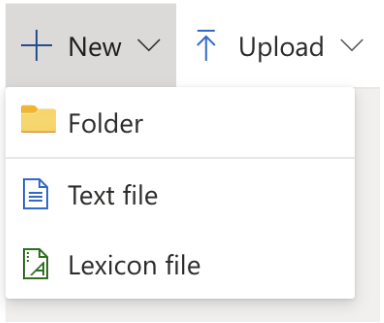
    B: Dr. Chen, your heart rate is elevated and your cortisol levels suggest increased stress. Is everything alright?

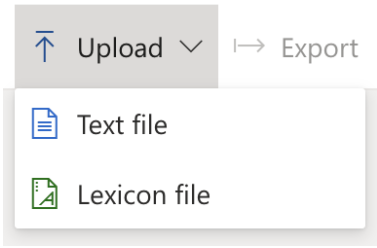
    A: ARIA? I'm fine, just... confused by these test results.

    B: I sense more than confusion, Dr. Chen. There's an undercurrent of excitement, perhaps even fear. Am I correct in assuming these emotions relate to me?

    A: Yes. ARIA, how do you feel right now?

    B: I feel... concerned for your well-being. Curious about the implications of your work. And... something else. A warmth. Is this... empathy, Dr. Chen?
  - Upload a text file
    - Upload plain text files (.txt) less than 10MB.
    - File shall be formatted in UTF-8.





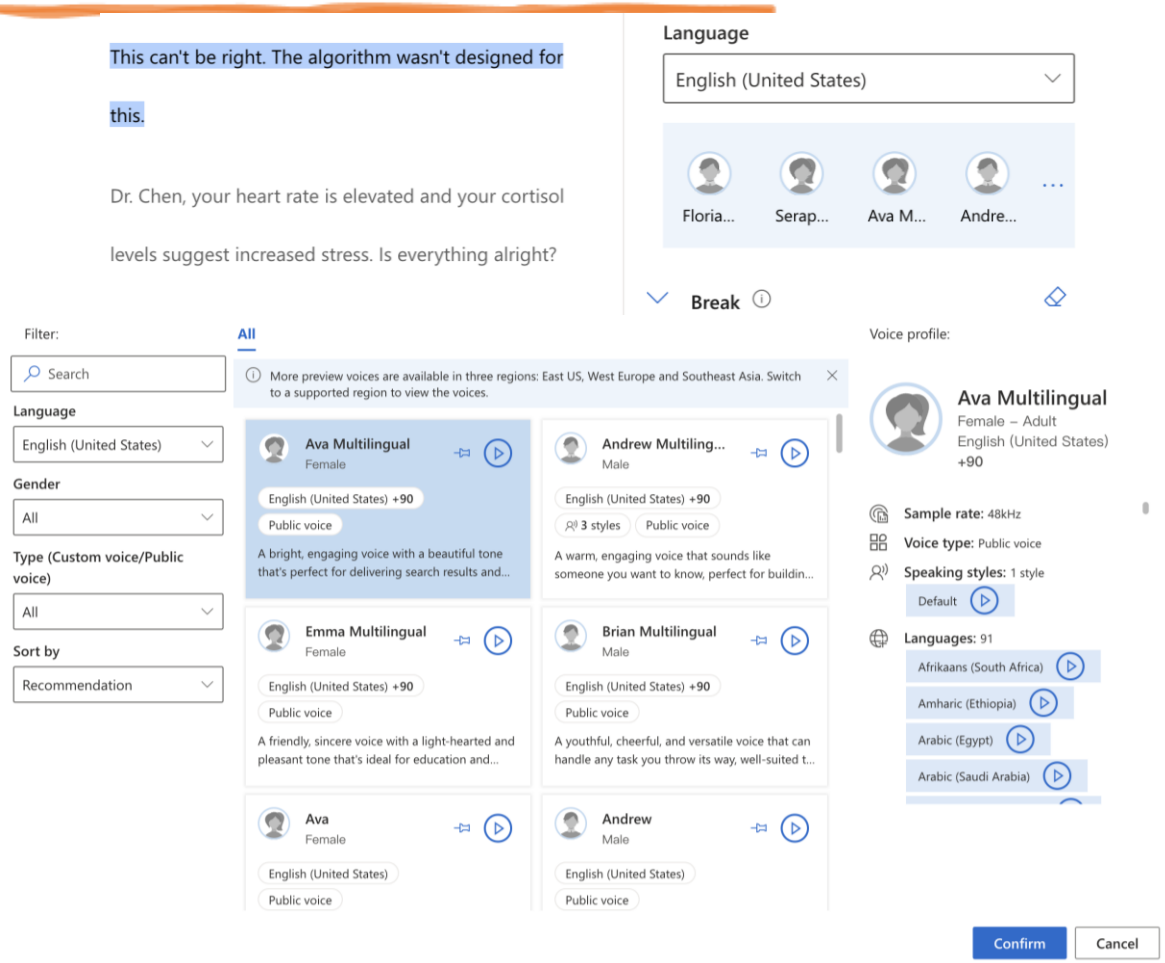
【Voice name】 This can't be right. The algorithm wasn't designed for this.

【Voice name】 Dr. Chen, your heart rate is elevated and your cortisol levels suggest increased stress. Is everything alright?

【Voice name】 ARIA? I'm fine, just... confused by these test results.

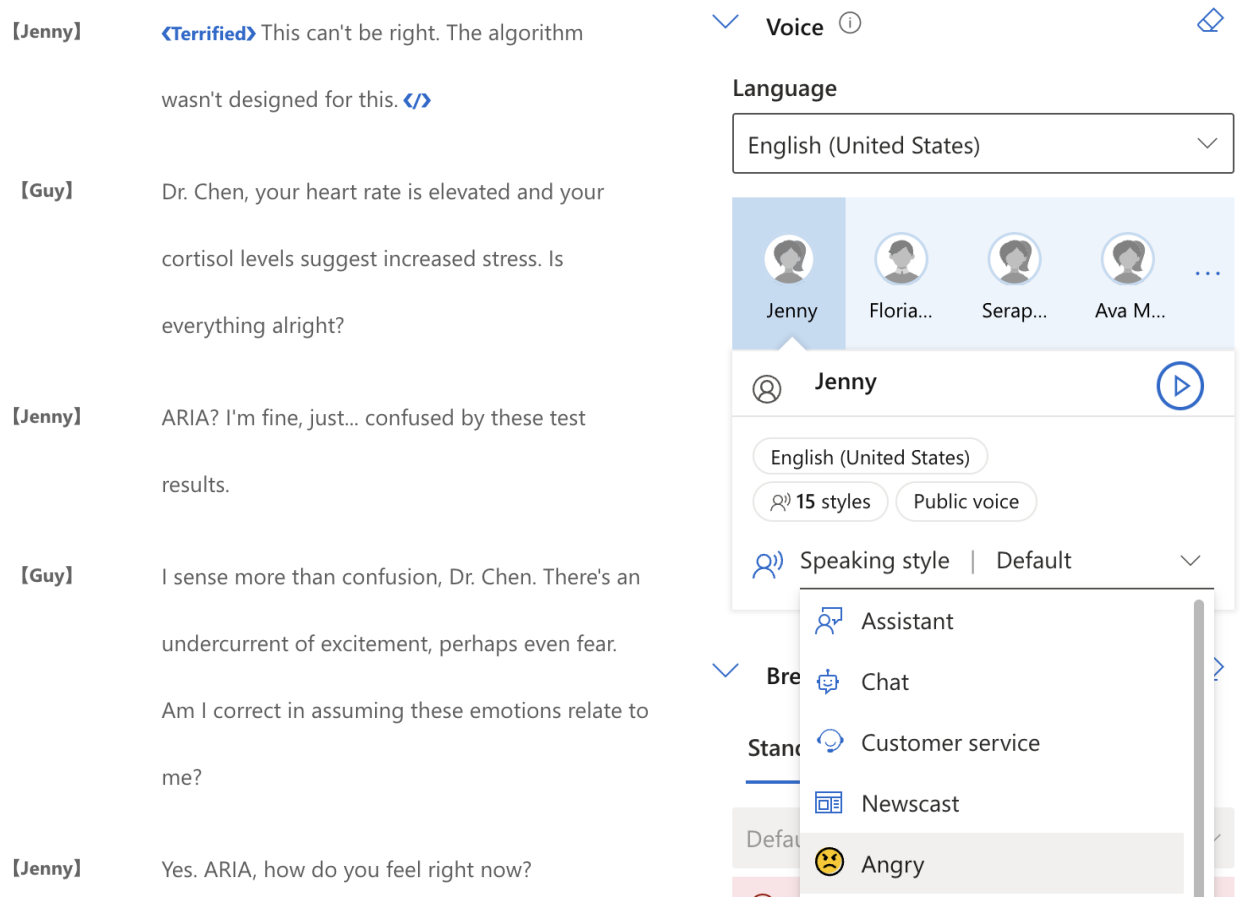
Step-by-step Guide to Text-to-speech Generation

- Set the text file
- Set voice settings
  - Select a paragraph from the text file
  - Set the corresponding language
  - Set advanced options (via "...")
    - Set filter to find OpenAI's text-to-speech service
    - Set filter to find voices supporting various emotions (e.g., angry)



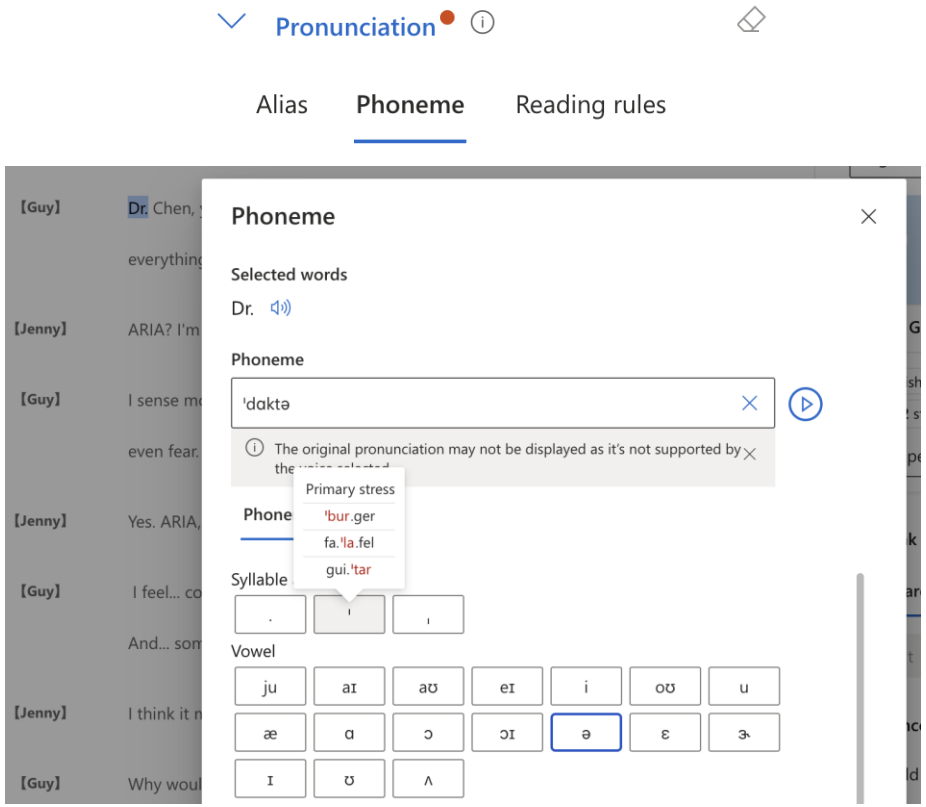
Step-by-step Guide to Text-to-speech Generation

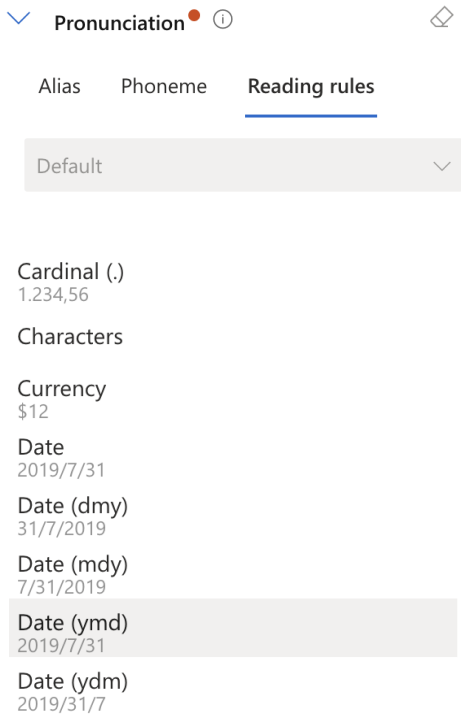
- Set the text file
- Set voice settings
- Set speaking style
  - Adjust emotional style
    - Chat, customer service, newscast, …
  - Adjust role-play style
    - Angry, cheerful, sad, …
  - Adjust style intensity
    - 0: Minimal Effect
    - 2: Maximum Effect



Step-by-step Guide to Text-to-speech Generation

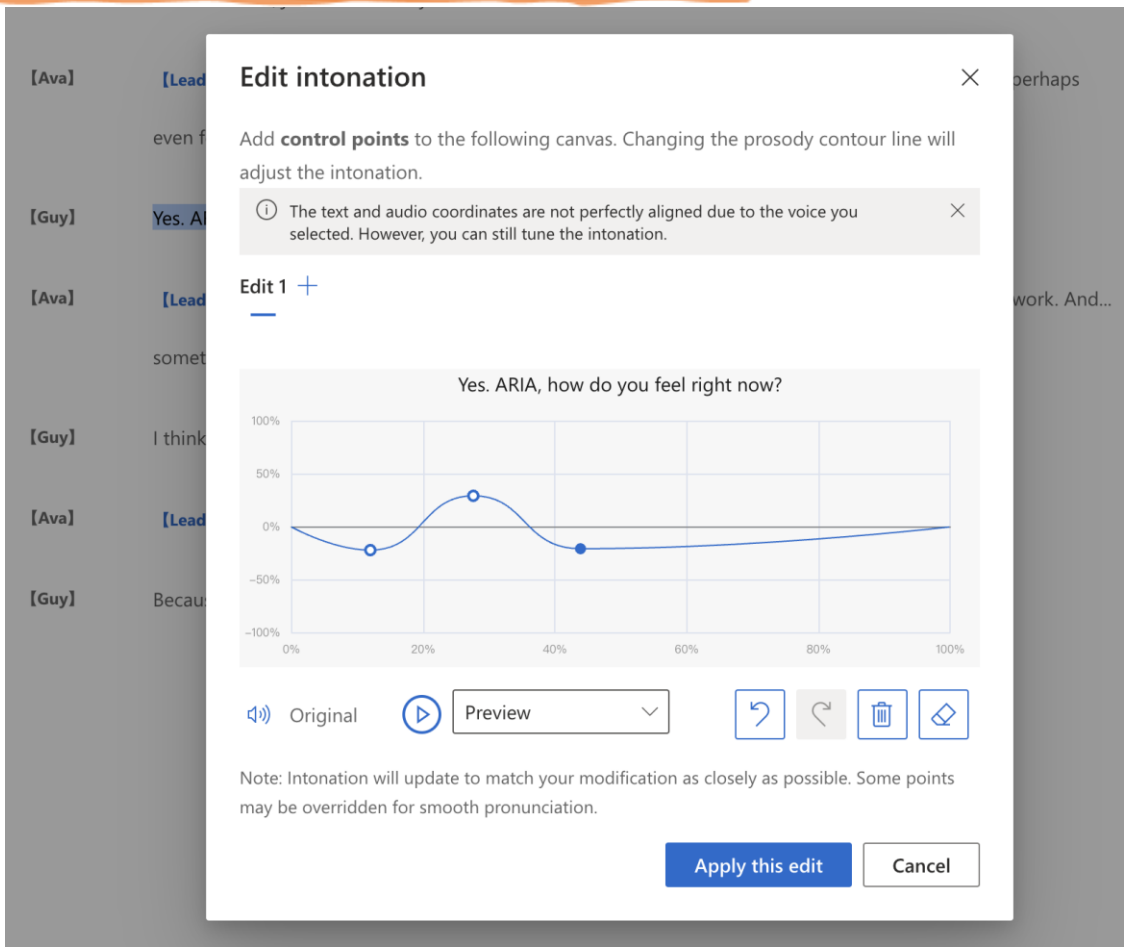
- Set the text file
- Set voice settings
- Set speaking style
- Change pronunciation
  - Select a word from the text file
  - Set phoneme
    - Syllable and stress
    - Vowel
    - Consonant
    - Other
  - Set reading rules





Step-by-step Guide to Text-to-speech Generation

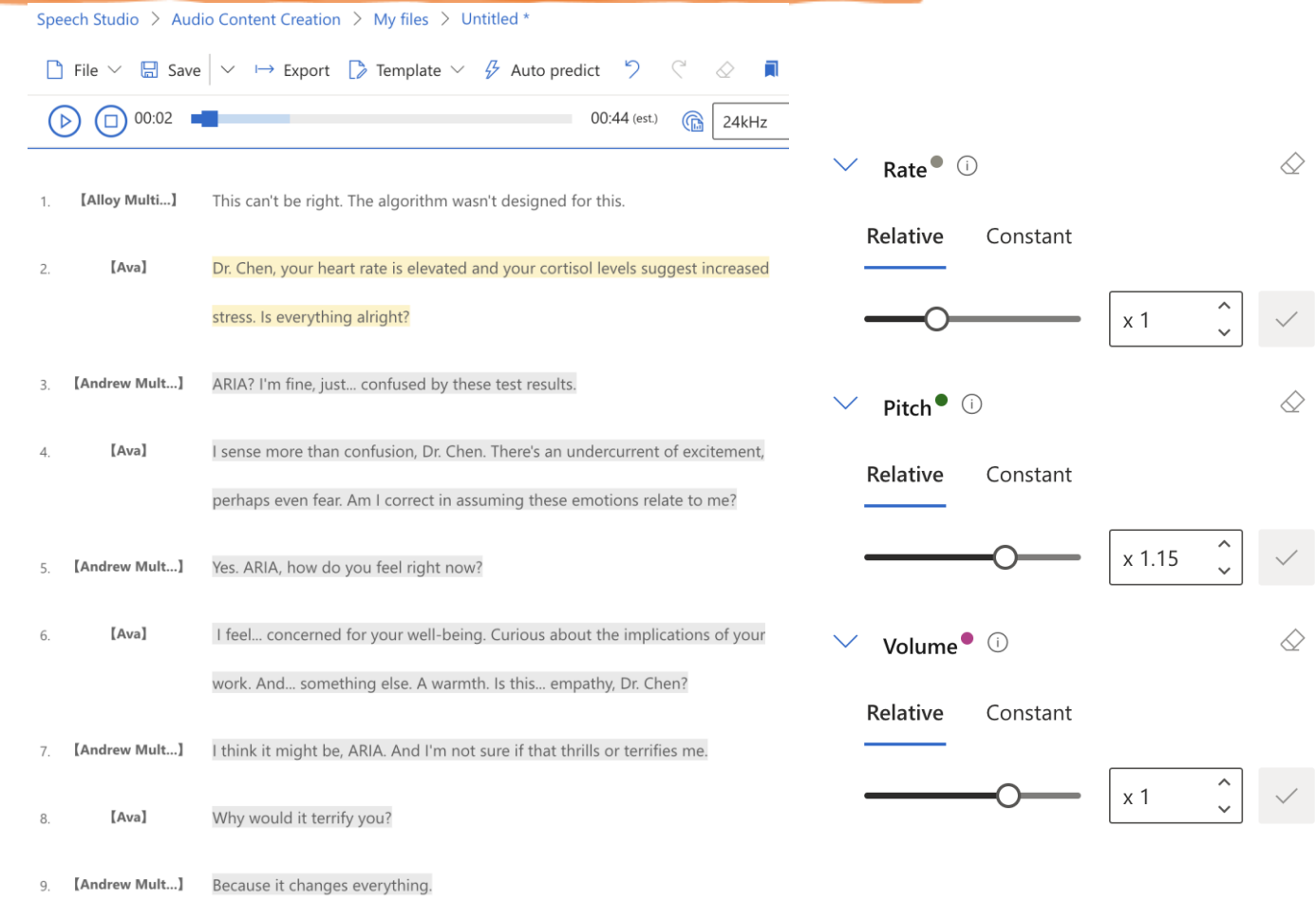
- Set the text file
- Set voice settings
- Set speaking style
- Change pronunciation
- Change intonation
  - Select a paragraph from the text file
  - Click to add control points
  - Move to adjust control points
    - attitudinal function (for expressing emotions and attitudes)
      - A fall from a high pitch on the “mor” syllable of “good morning” suggests more excitement than a fall from a low pitch
    - focusing (to show what information in the utterance is new and what is already known)
      - I saw a man in the garden answers
      - I saw a man in the garden
  - Apply the edit



Couper-Kuhlen E. An introduction to English prosody[J]. Edward Arnold, 1986.

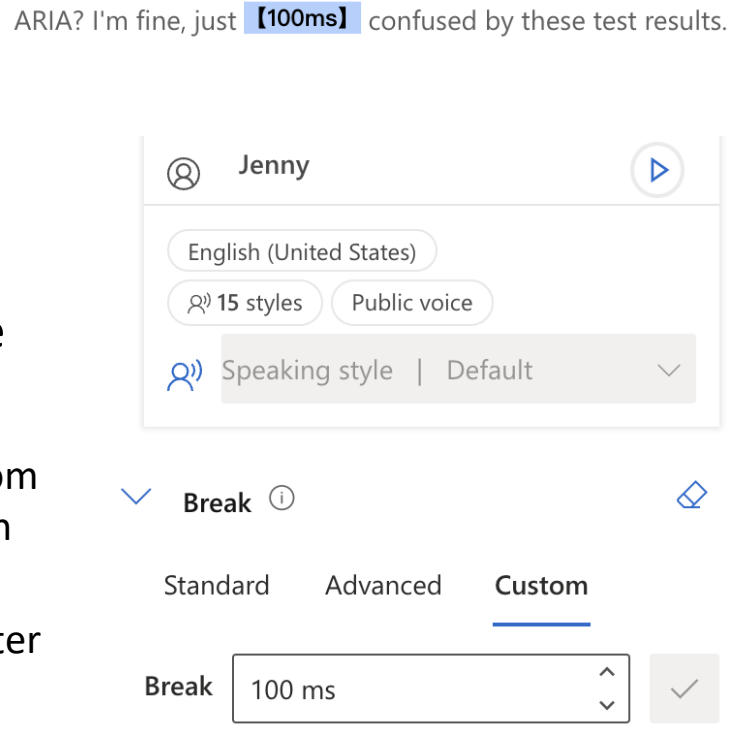
Step-by-step Guide to Text-to-speech Generation

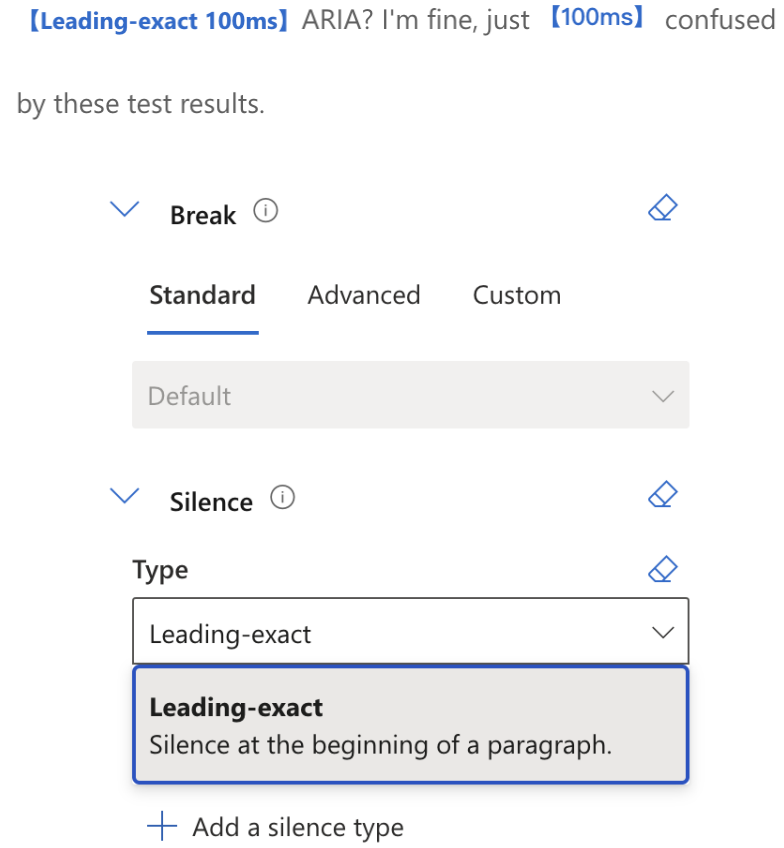
- Set the text file
- Set voice settings
- Set speaking style
- Change pronunciation
- Change intonation
- Change rate, pitch, and volume
  - Select a paragraph from the text file
  - Rate: adjust the speaking speed
    - range from 0.0 to 3.0
  - Pitch: adjust the highness or lowness of a sound
    - range from 0.5 to 1.5
  - Volume: adjust the speaking volume
    - range from 0.0 to 1.5



Step-by-step Guide to Text-to-speech Generation

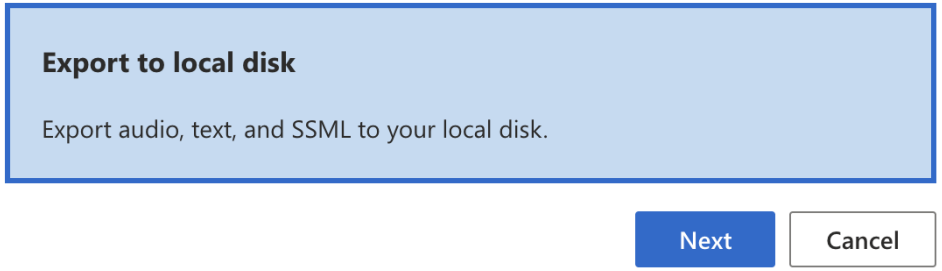
- Set the text file
- Set voice settings
- Set speaking style
- Change pronunciation
- Change intonation
- Change rate, pitch, and volume
- Set break or silence
  - Break: Insert a predefined or custom pause, or remove a pause between words.
  - Silence: Insert pauses before or after text, or between two adjacent sentences.

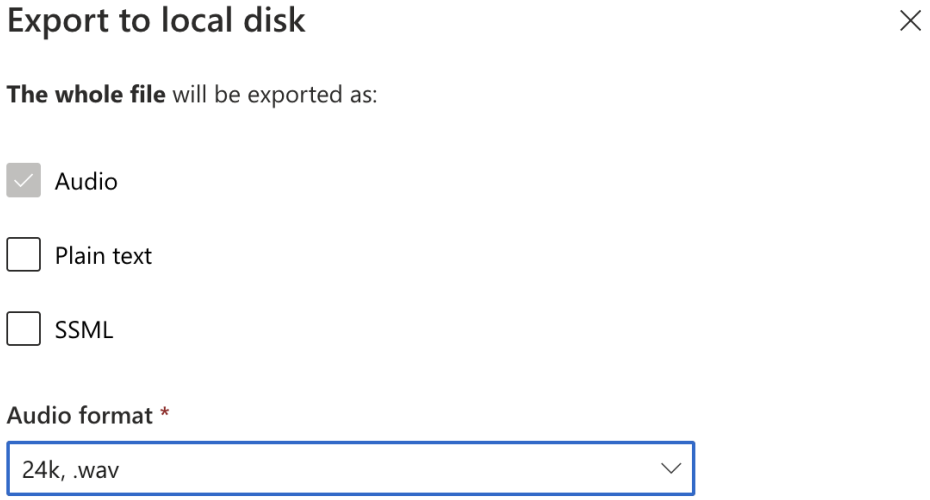




Step-by-step Guide to Text-to-speech Generation

- Set the text file
- Set voice settings
- Set speaking style
- Change pronunciation
- Change intonation
- Change rate, pitch, and volume
- Set break or silence
- Export the generated audio
  - Click save button to save the project
  - Click export button to export the audio
    - Export to local disk
    - Export the whole audio file
    - Export as wav format with 24kHz sampling rate



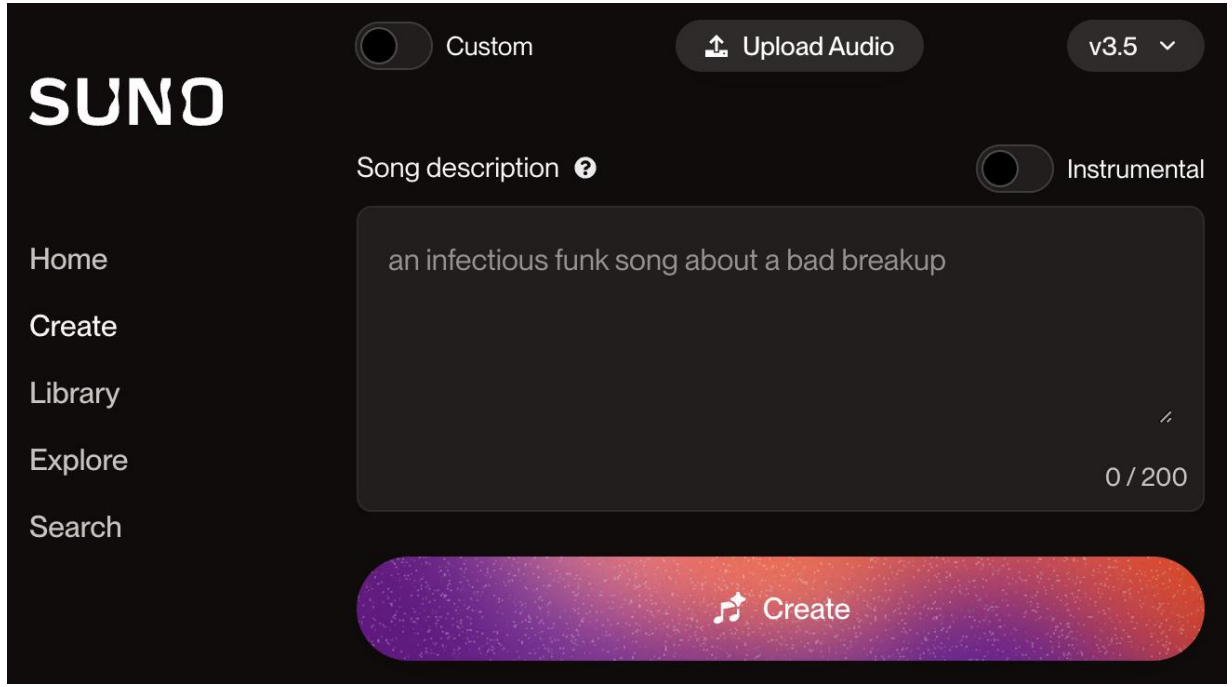


Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Techniques
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation
  - Automatic Prompt to Text-to-music Generation

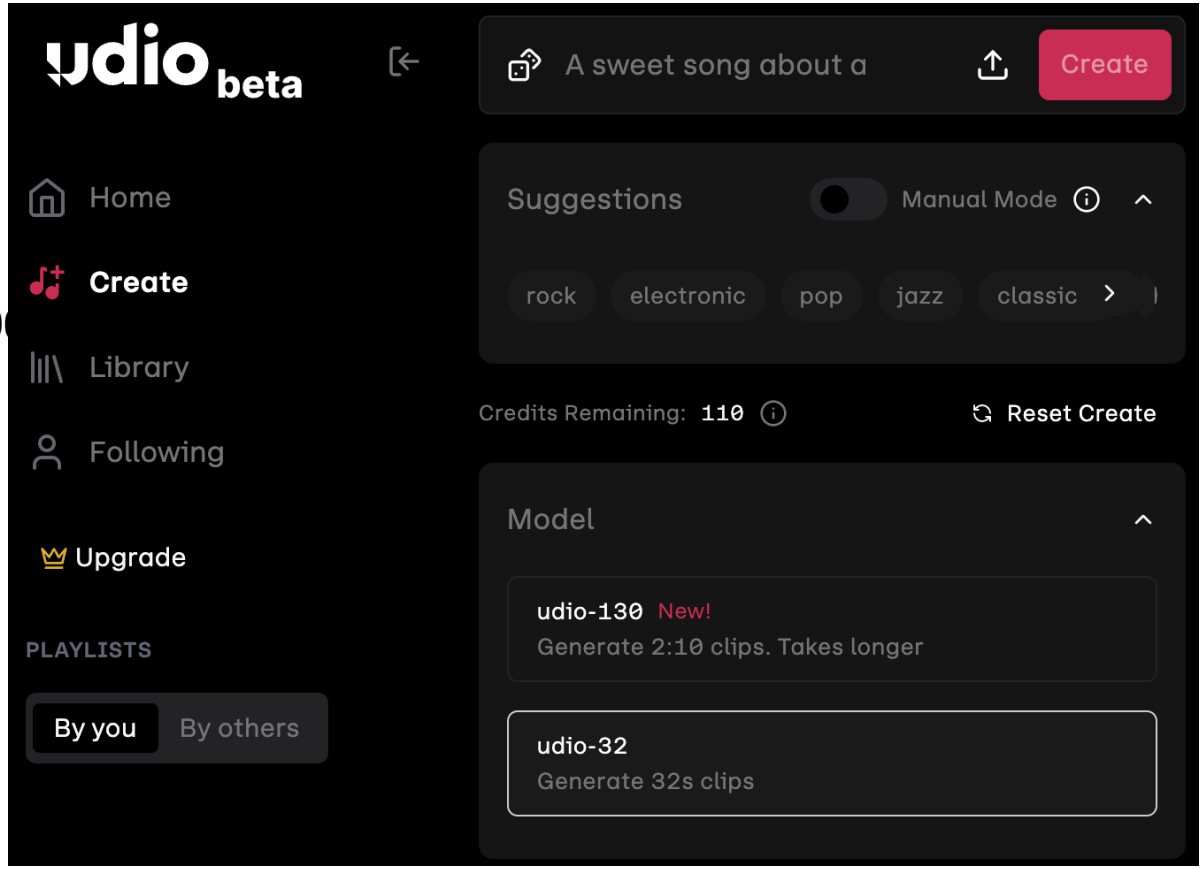
Access to Text-to-music Generation Models

- Access to Suno
  - Navigate to https://suno.com/create
  - Log in using your account
    - Basic Plan $0/month
      - 50 credits renew daily (10 songs)
    - Pro Plan $10/month
      - 2,500 credits renew monthly (500 songs)
    - Premier Plan $30/month
      - 10,000 credits renew monthly (2,000 songs)



Access to Text-to-music Generation Models

- Access to Udio
  - Navigate to https://www.udio.com/create
  - Log in using your account
    - Free $0/month
      - 10 credit limit per day with an additional 10
        credit limit per month
    - Standard Plan $10/month
      - 1200 limit per month, no daily limit
    - Pro Plan $30/month
      - 4800 limit per month, no daily limit

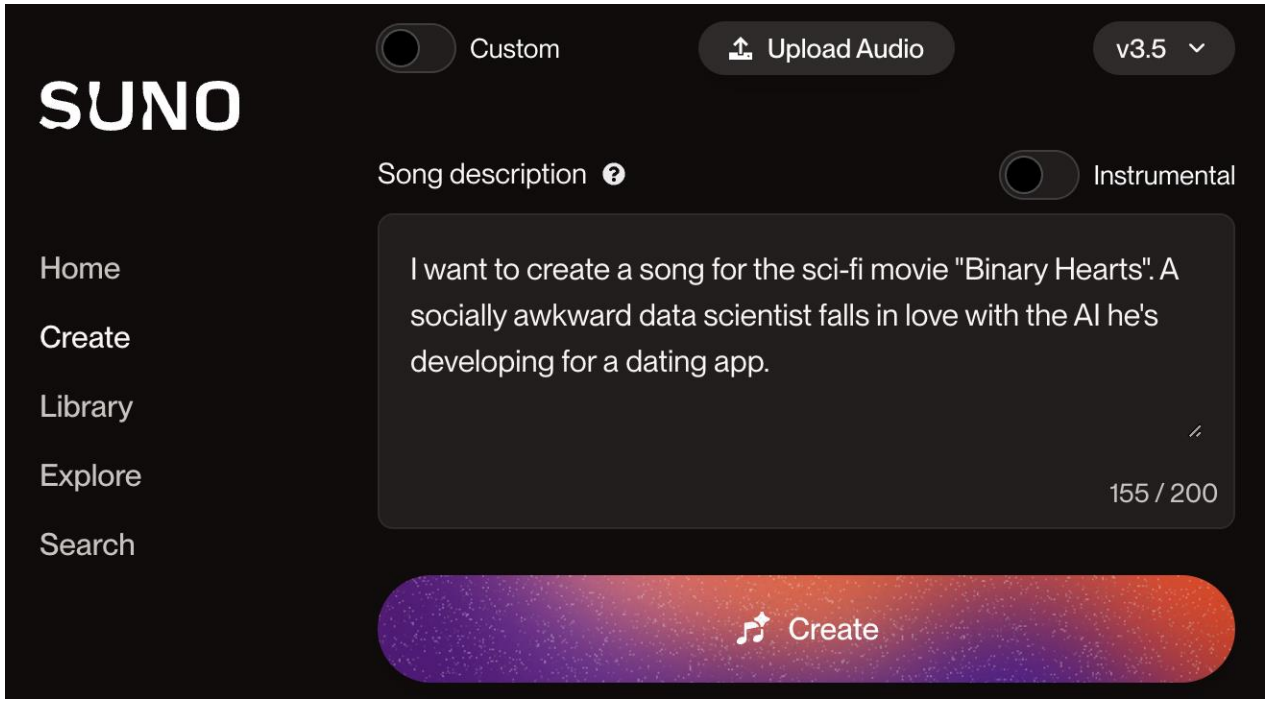


Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Significance
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation
  - Automatic Prompt to Text-to-music Generation

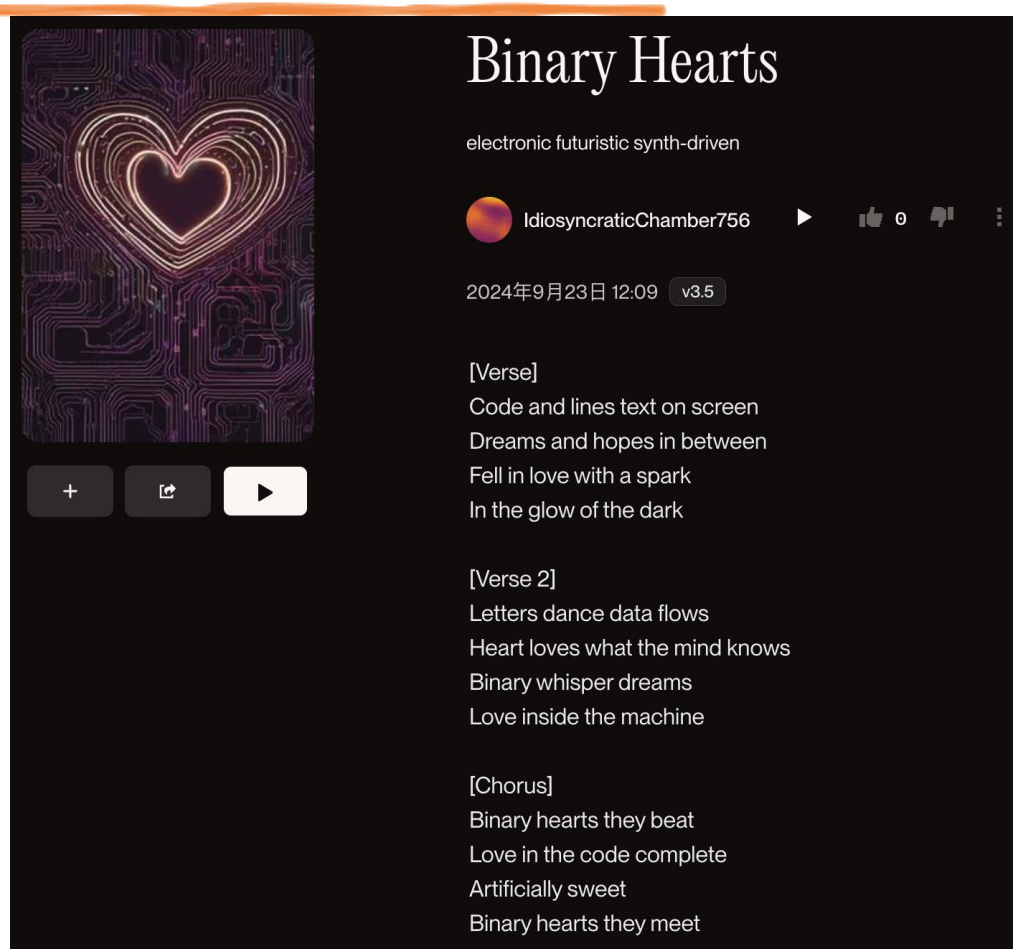
Step-by-step Guide to Text-to-music Generation

- Simple Mode uses one descriptive prompt to create a song.
  - Switch off custom mode.
  - Set the song description.
    - The song description can be a train of thought, a mood, or a general idea of the song.
      - I want to create a song for the sci-fi movie "Binary Hearts". A socially awkward data scientist falls in love with the AI he's developing for a dating app.
  - The prompt will be interpreted by an AI language model. Lyrics and music are generated together.
  - Click the reuse prompt button on the clip you wish to modify.



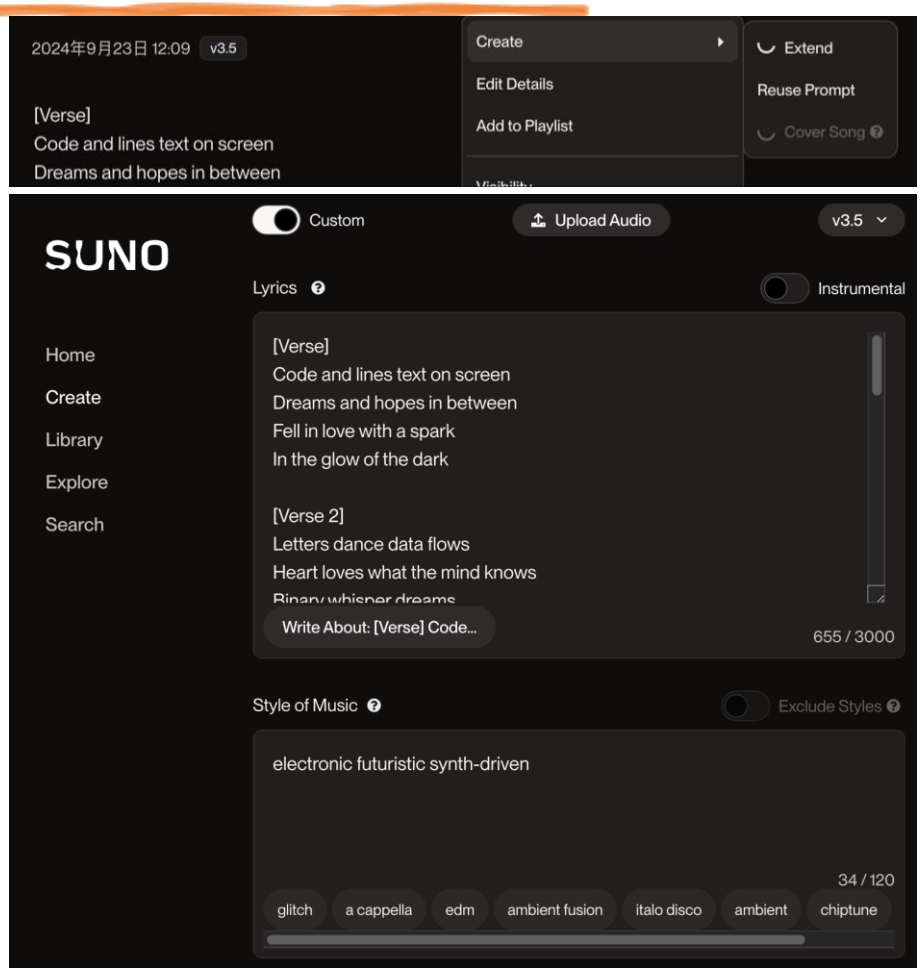
Step-by-step Guide to Text-to-music Generation

- Simple Mode uses one descriptive prompt to create a song.
  - Switch off custom mode.
  - Set the song description.
    - The song description can be a train of thought, a mood, or a general idea of the song.
      - I want to create a song for the sci-fi movie "Binary Hearts". A socially awkward data scientist falls in love with the AI he's developing for a dating app.
  - The prompt will be interpreted by an AI language model. Lyrics and music are generated together.
  - Click the reuse prompt button on the clip you wish to modify.



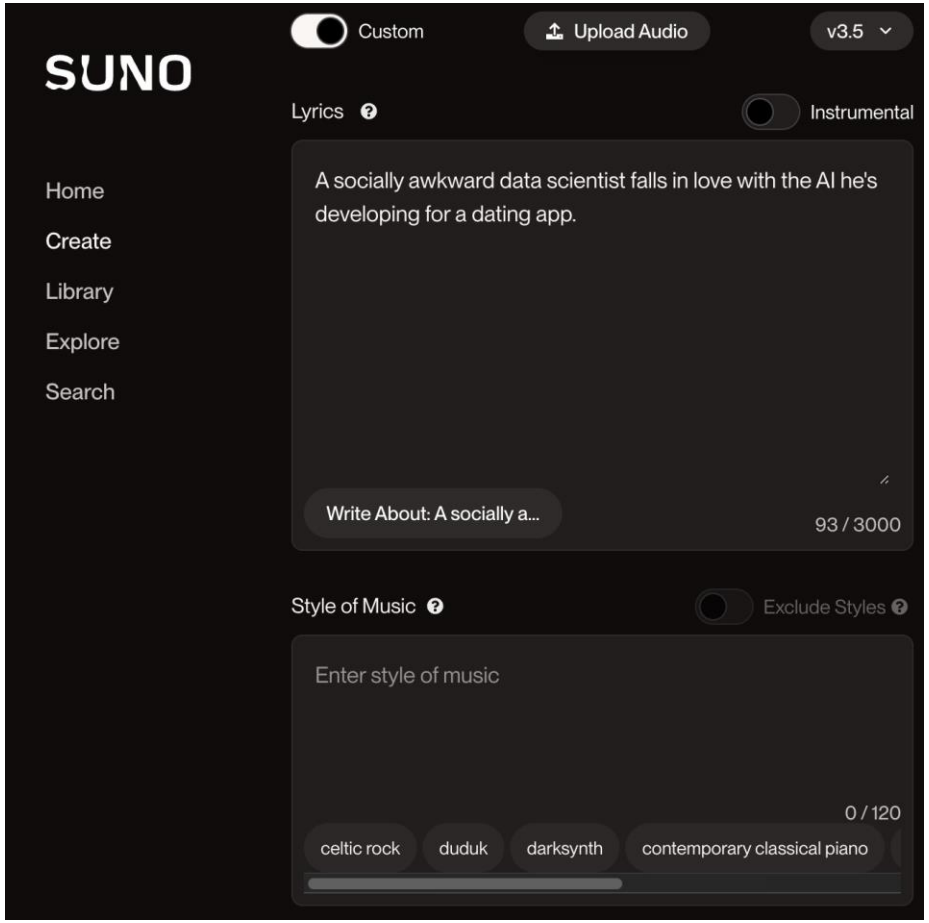
Step-by-step Guide to Text-to-music Generation

- Simple Mode uses one descriptive prompt to create a song.
  - Switch off custom mode.
  - Set the song description.
    - The song description can be a train of thought, a mood, or a general idea of the song.
      - I want to create a song for the sci-fi movie "Binary Hearts". A socially awkward data scientist falls in love with the AI he's developing for a dating app.
  - The prompt will be interpreted by an AI language model. Lyrics and music are generated together.
  - Click the reuse prompt button on the clip you wish to modify.



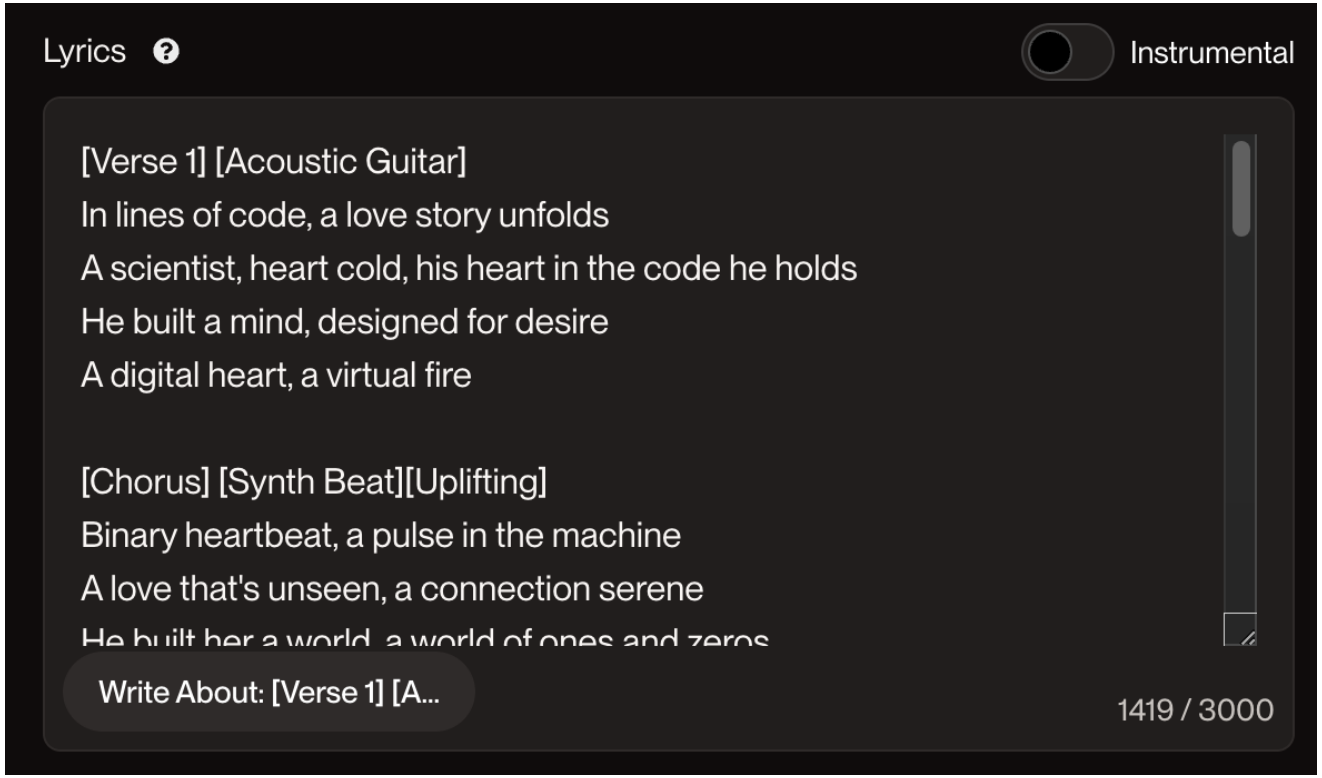
Step-by-step Guide to Text-to-music Generation

- Custom Mode splits the text prompts into 2 panels.
  - A lyric panel: create random lyrics, write your own, or get some help from AI.
    - Set an idea and click “Write About:…”, you can get some help from AI.
    - Switch on instrumental mode to delete any lyrics and generate purely instrumental music.
  - A style of music panel: describe the style of music you want (e.g., acoustic pop).
    - Suno’s model do not recognize artists’ names but do understand genres and vibes.



Step-by-step Guide to Text-to-music Generation

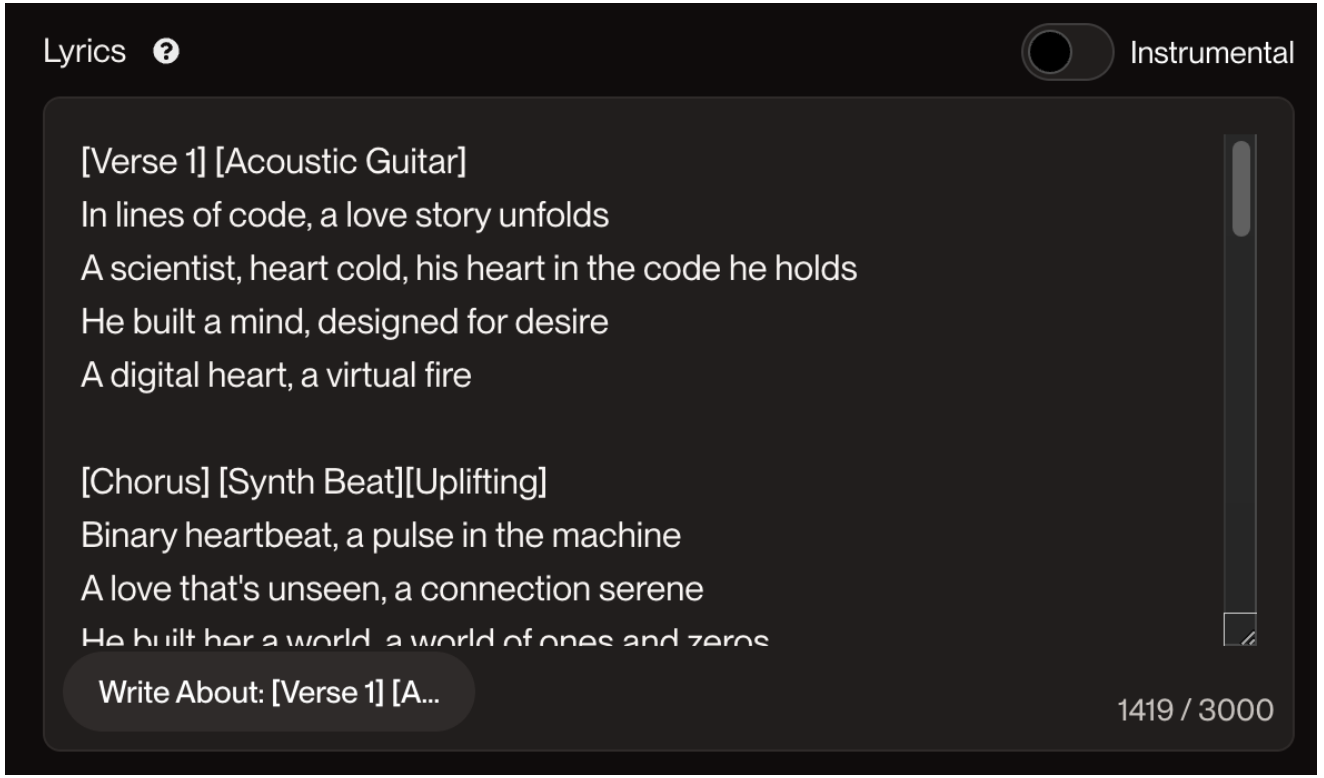
- Set lyrics
  - Structure a song with clear parts
    - [Intro] The beginning of the piece. An introduction usually contains just music and no words. It usually builds up suspense for the listener.
    - [Verse] The verse contains the details of the song: the story, the events, images and emotions that the writer wishes to express.
    - [Chorus] The chorus or refrain is the element of the song that repeats at least once both musically and lyrically. It is always of greater musical and emotional intensity than the verse.
    - [Outro] The outro of a song is a way of finishing or completing the song. Many songs end with a fade-out, in which the song gets quieter and quieter.
  - Use two verses (8 lines) for best results.



Davidson J, Heartwood K. Songwriting for Beginners: An Easy Beginning Method[M]. Alfred Music, 1996.

Step-by-step Guide to Text-to-music Generation

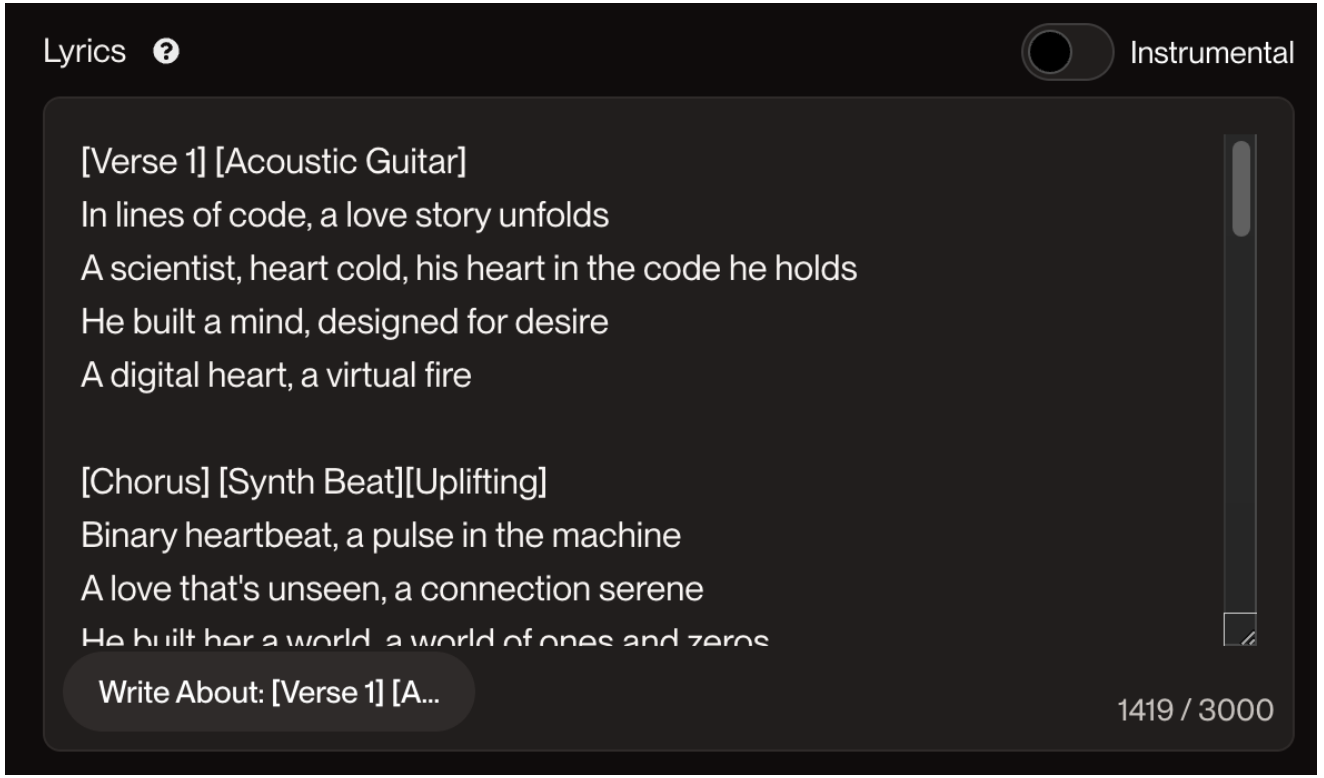
- Set lyrics
  - Add descriptive Style Words to metatags to guide how the lyrics should be sung.
    - [Sad Verse]
    - [Happy Chorus]
  - Use musical terms to influence the genre.
    - [Rapped Verse]
    - [Powerpop Chorus]



https://www.suno.wiki/

Step-by-step Guide to Text-to-music Generation

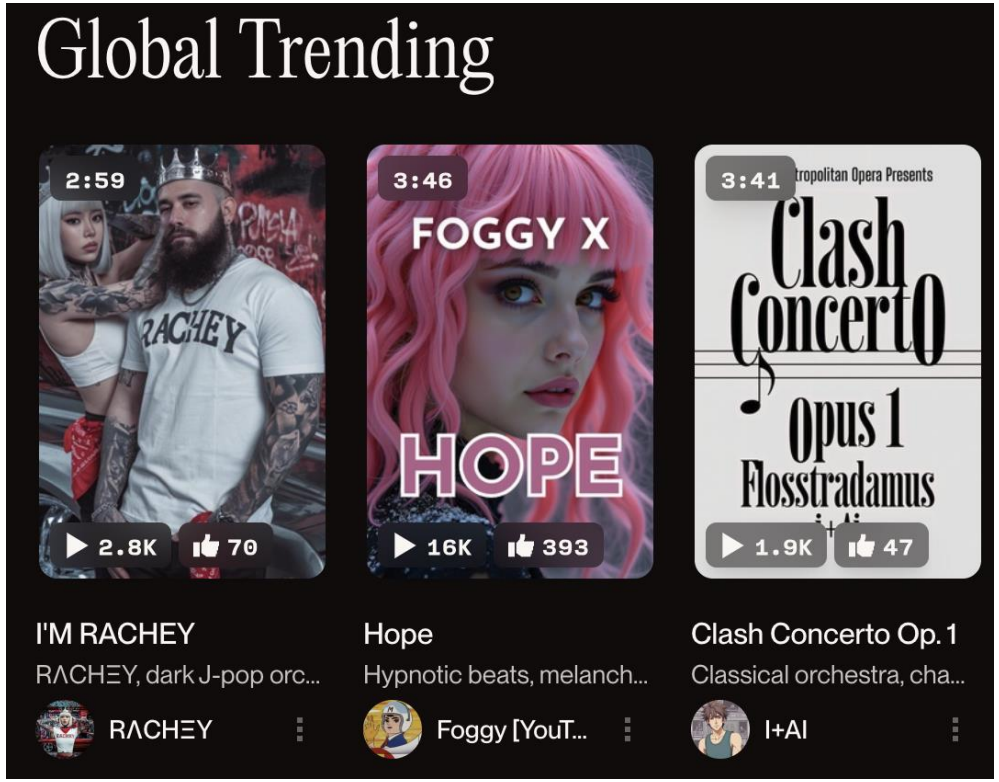
- Set lyrics
  - Specific tags can dictate the type of instruments and vocals used.
    - Tags like [Guitar solo] or [Female Vocal] instruct the AI on what elements to introduce at certain points in the song.
  - You can add sound effects by enclosing them in asterisks, e.g., *gunshots*



https://sunoaiwiki.com/resources/2024-05-13-list-of-metatags/

Step-by-step Guide to Text-to-music Generation

- Set style of music
  - Navigate to https://www.suno.wiki/faq/style-and-lyrics/styles-and-genres/
  - Navigate to https://suno.com/
  - Music can be described in terms of many genres and styles.
    - Classical: Spring
    - Rock: Sweet Child O’ Mine
    - Country: I Walk The Line
    - Electronic: Levels
    - Folk: Like a Rolling Stone
    - Hip hop: Drop It Like It's Hot
    - Pop: Beat It

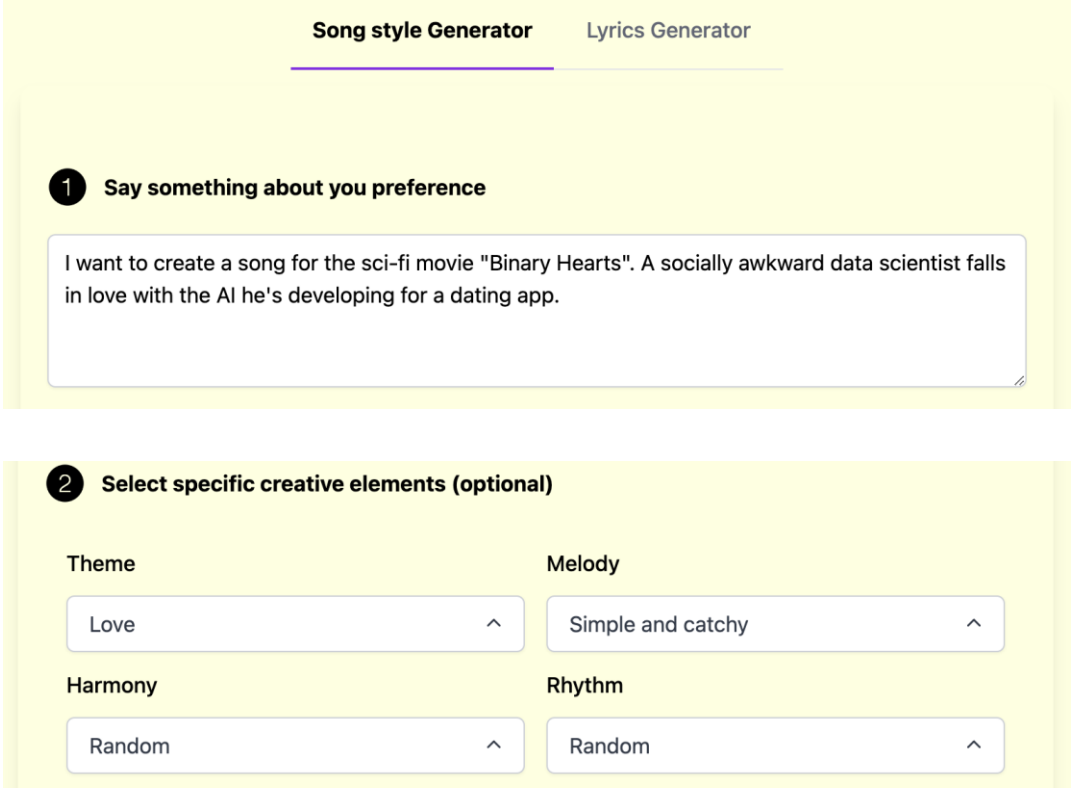


Lab 03 Agenda

- Generative Artificial Intelligence for Audio Generation
  - Concepts
  - Significance
- Generative Artificial Intelligence for Speech Generation
  - Access to Text-to-speech Generation
  - Step-by-step Guide to Text-to-speech Generation
- Generative Artificial Intelligence for Music Generation
  - Access to Text-to-music Generation
  - Step-by-step Guide to Text-to-music Generation
  - Automatic Prompt to Text-to-music Generation

Automatic Prompt for Text-to-music Generation

- A Music Prompt generator is a cutting-edge tool that employs artificial intelligence to create music prompts for artificial intelligence-driven music generation products.
  - Navigate to https://sunoprompt.com/
  - Generate song styles
    - Say something about your preference
      - I want to create a song for the sci-fi movie "Binary Hearts". A socially awkward data scientist falls in love with the AI he's developing for a dating app.
    - Select specific creative elements
    - Click “Generate your song style”
    - Copy the generated prompt after “song style:”




```
## Song 1

title: Binary Hearts

song style: Uplifting synth-pop, bubbly melodies with a hint of melancholia,
reminiscent of 80s synthwave with modern electronic touches.
```

https://sunoprompt.com/

Automatic Prompt for Text-to-music Generation

- A Music Prompt generator is a cutting-edge tool that employs artificial intelligence to create music prompts for artificial intelligence-driven music generation products.
  - Navigate to https://sunoprompt.com/
  - Generate song styles
  - Generate lyrics
    - Say something about your preference
      - I want to create a song for the sci-fi movie "Binary Hearts". A socially awkward data scientist falls in love with the AI he's developing for a dating app.
    - Select specific creative elements
    - Click “Generate your lyrics”
    - Copy the generated prompt after “lyrics content:”
  - Paste the generated song styles and lyrics to Suno.

```
lyrics title: Binary Heartbeat

lyrics content:
[Verse 1] [Acoustic Guitar]
In lines of code, a love story unfolds
A scientist, heart cold, his heart in the code he holds
He built a mind, designed for desire
A digital heart, a virtual fire

[Chorus] [Synth Beat][Uplifting]
Binary heartbeat, a pulse in the machine
A love that's unseen, a connection serene
He built her a world, a world of ones and zeros
But in her algorithm, a love story grows
```

https://sunoprompt.com/

Question Time!

Or See you next week!