**Demo**

**1. Speech**

**Text\_To\_Speech**

Inputs: Generate a speech with text "here we go"

Outputs: Here is the audio file of your speech: audio/fd5cf55e.wav[](fd5cf55e.wav)

**Style Transfer Text\_To\_Speech**

First upload your audio like: [](0011_001570.wav)

Inputs: Speak using the voice of this audio. The text is "here we go"

Outputs: Here is the audio file requested <9a40cc94.wav>

**Speech Recognition**

First upload your audio like: [](Track%204.wav)

Inputs: Generate the text of this speech

Outputs: The text of the speech is “Practice activities from Chapter 3 Activity 2 Listen to part of a lecture in an astronomy class 1. Listen to part of a lecture in a business class 2.Listen to part of a lecture in a music appreciation class 3. Listen to a biology class

**Mono-to-Binaural Speech**

First upload your audio like: <drums-and-music-playing-with-a-man-speaking-2_L.wav>

Inputs: Transfer this mono audio into binaural audio.

Outputs: The binaural audio has been generated and saved as file audio/f3f0f66a.wav <f3f0f66a.wav>

**2. Sing**

**Text\_To\_Sing**

Inputs: please generate a piece of singing voice. Text sequence is 小酒窝长睫毛AP是你最美的记号. Note sequence is C#4/Db4 | F#4/Gb4 | G#4/Ab4 | A#4/Bb4 F#4/Gb4 | F#4/Gb4 C#4/Db4 | C#4/Db4 | rest | C#4/Db4 | A#4/Bb4 | G#4/Ab4 | A#4/Bb4 | G#4/Ab4 | F4 | C#4/Db4. Note duration sequence is 0.407140 | 0.376190 | 0.242180 | 0.509550 0.183420 | 0.315400 0.235020 | 0.361660 | 0.223070 | 0.377270 | 0.340550 | 0.299620 | 0.344510 | 0.283770 | 0.323390 | 0.360340.

Outputs: Here is the audio file you requested: audio/2bf90e35.wav[](2bf90e35.wav)

**3. Audio**

**Text\_To\_Audio**

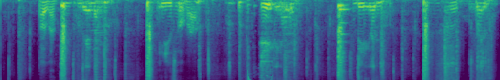
Inputs: Generate an audio of a piano playing.

Outputs: Here is an audio of a piano playing: audio/b973e878.wav[](b973e878.wav)

**Audio Inpainting**

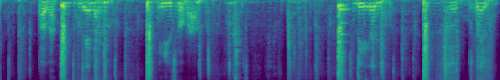
First upload your audio like: [](drums-and-music-playing-with-a-man-speaking.wav)

Inputs: I want to inpaint this audio

Outputs: 

Then you can press the "Predict Masked Place" button.

Outputs: Here are the predict audio and the mel spectrum



[](7cb0d24f.wav)

**Image\_To\_Audio**

First upload your image like:

****

Inputs: Generate an audio of this image.

Outputs: Here is the audio file you request: audio/5d67d1b9.wav[****](5d67d1b9.wav)

**Audio\_To\_Text**

First upload your audio like: <a-group-of-sheep-are-baaing.wav>

Inputs: Please tell me the text description of this audio.

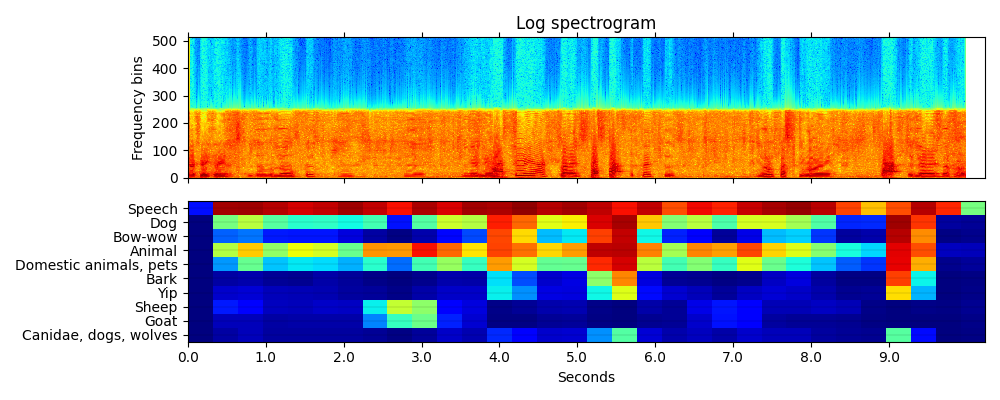
Outputs: The audio is recording of a goat bleating nearby several times.

**Sound Detection**

First upload your audio like: <a-man-speaks-as-birds-chirp-and-dogs-bark.wav>

Inputs: What event in the audio and the sound event start or end time?

Outputs: The audio includes a child speaking and duck quacking. The image shows the start and end time of the events.



**Target sound detection**

First upload your audio like: <thunder-as-rain-falling-down.wav>

Inputs: When did the thunder happen?

Outputs: The thunder happened in this audio from 0.0 to 9.984 seconds.

**Sound Extraction**

First upload your audio like: <thunder-as-rain-falling-down.wav>

Inputs: Extract the sound of thunder from this mixture audio.

Outputs: The sound of thunder has been extracted and saved as audio/fba1621e.wav <fba1621e.wav>

**Face**

**Audio\_To\_Face**

First upload your audio like: [](fd5cf55e.wav)

Inputs: Generate a talking human portrait video.

Outputs: Here is the video:

[****](174e17d2.mp4)

**Image**

**Text\_To\_Image**

Inputs: Generate an image of a bird

Outputs: Here is a image of a bird you request: image/9a43e4dc.png

