U1.3 Load and explore an audio dataset

```
pip install datasets[audio]
```

```
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests>=2.32.2->datasets[audio]) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests>=2.32.2->datasets[audio]) (2024.7.4)
Requirement already satisfied: cffi>=1.0 in /usr/local/lib/python3.10/dist-packages (from soundfile>=0.12.1->datasets[audio]) (1.16.0)
Requirement already satisfied: audioread>=2.1.9 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (3.0.1)
Requirement already satisfied: scipy>=1.2.0 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (1.13.1)
Requirement already satisfied: scikit-learn>=0.20.0 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (1.3.2)
Requirement already satisfied: joblib>=0.14 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (1.4.2)
Requirement already satisfied: decorator>=4.3.0 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (4.4.2)
Requirement already satisfied: numba>=0.51.0 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (0.60.0)
Requirement already satisfied: pooch>=1.1 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (1.8.2)
Requirement already satisfied: soxr>=0.3.2 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (0.4.0)
Requirement already satisfied: lazy-loader>=0.1 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (0.4)
Requirement already satisfied: msgpack>=1.0 in /usr/local/lib/python3.10/dist-packages (from librosa->datasets[audio]) (1.0.8)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets[audio]) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets[audio]) (2024.1)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets[audio]) (2024.1)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-packages (from cffi>=1.0->soundfile>=0.12.1->datasets[audio]) (2.22)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.51.0->librosa->datasets[audio]) (0.43.0)
Requirement already satisfied: platformdirs>=2.5.0 in /usr/local/lib/python3.10/dist-packages (from pooch>=1.1->librosa->datasets[audio]) (4.2.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->datasets[audio]) (1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.20.0->librosa->datasets[audio]) (3.5.0)
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
```

```
ibis-framework 8.0.0 requires pyarrow<16,>=2, but you have pyarrow 17.0.0 which is incompatible.
Successfully installed datasets-2.20.0 dill-0.3.8 fsspec-2024.5.0 multiprocess-0.70.16 pyarrow-17.0.0 xxhash-3.4.1
```

Let's load and explore and audio dataset called MINDS-14, which contains recordings of people asking an e-banking system questions in several languages and dialects.

To load the MINDS-14 dataset, we need to copy the dataset's identifier on the Hub (PolyAI/minds14) and pass it to the load_dataset function. We'll also specify that we're only interested in the Australian subset (en-AU) of the data, and limit it to the training split:

```
from datasets import load_dataset

minds = load_dataset("PolyAI/minds14", name="en-AU", split="train")
minds
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
Downloading builder script: 100%  5.90k/5.90k [00:00<00:00, 307kB/s]
Downloading readme: 100%  5.29k/5.29k [00:00<00:00, 330kB/s]
The repository for PolyAI/minds14 contains custom code which must be executed to correctly load the dataset. You can inspect the repository content at https://hf.co/datasets/PolyAI/minds14.
You can avoid this prompt in future by passing the argument `trust_remote_code=True`.

Do you wish to run the custom code? [y/N] y
Downloading data: 100%  471M/471M [00:08<00:00, 57.1MB/s]
Generating train split:  654/0 [00:00<00:00, 5396.17 examples/s]
Dataset({
    features: ['path', 'audio', 'transcription', 'english_transcription', 'intent_class', 'lang_id'],
    num_rows: 654
})
```

```
example = minds[0]
example
```

```
{'path': '/root/.cache/huggingface/datasets/downloads/extracted/28aa727f91fee90575c34956bab09d1716cfaf460c6afcba86a10f04a7d58b83/en-AU~PAY_BILL/response_4.wav',
 'audio': {'path': '/root/.cache/huggingface/datasets/downloads/extracted/28aa727f91fee90575c34956bab09d1716cfaf460c6afcba86a10f04a7d58b83/en-AU~PAY_BILL/response_4.wav',
  'array': array([ 0.        ,  0.00024414, -0.00024414, ..., -0.00024414,
         0.00024414,  0.0012207 ]),
  'sampling_rate': 8000},
 'transcription': 'I would like to pay my electricity bill using my card can you please assist',
 'english_transcription': 'I would like to pay my electricity bill using my card can you please assist',
 'intent_class': 13,
 'lang_id': 2}
```

You may notice that the audio column contains several features. Here's what they are:

path: the path to the audio file (*.wav in this case). array: The decoded audio data, represented as a 1-dimensional NumPy array. sampling_rate. The sampling rate of the audio file (8,000 Hz in this example). The intent_class is a classification category of the audio recording. To convert this number into a meaningful string, we can use the int2str() method:

```
id2label = minds.features["intent_class"].int2str
id2label(example["intent_class"])
```

```
'nav_bill'
```

f you plan to train an audio classifier on this subset of data, you may not necessarily need all of the features. For example, the lang_id is going to have the same value for all examples, and won't be useful. The english_transcription will likely duplicate the transcription in this subset, so we can safely remove them.

You can easily remove irrelevant features using 🤗 Datasets' remove_columns method:

```
columns_to_remove = ["lang_id", "english_transcription"]
minds = minds.remove_columns(columns_to_remove)
minds
```

```
Dataset({
    features: ['path', 'audio', 'transcription', 'intent_class'],
    num_rows: 654
})
```

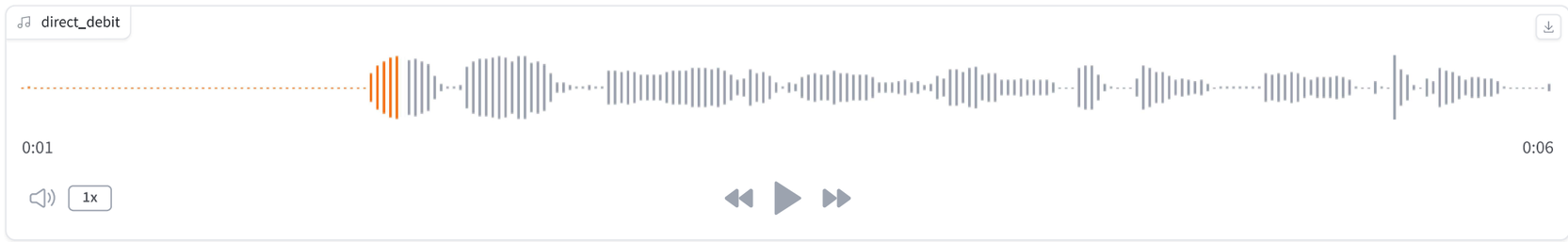
Now that we've loaded and inspected the raw contents of the dataset, let's listen to a few examples! We'll use the Blocks and Audio features from Gradio to decode a few random samples from the dataset:

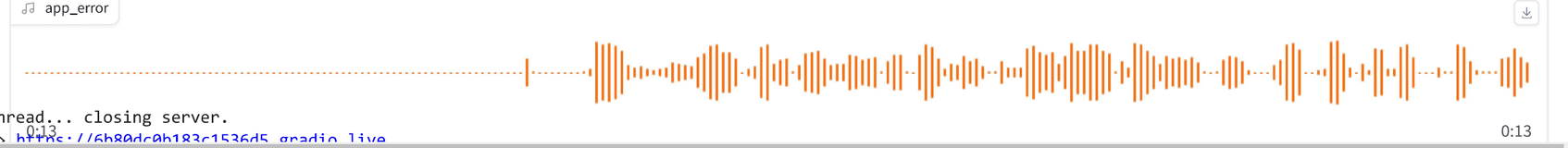
```
!pip install gradio
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 141.1/141.1 kB 9.9 MB/s eta 0:00:00
  Downloading python_multipart-0.0.9-py3-none-any.whl (22 kB)
  Downloading ruff-0.5.6-py3-none-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (10.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.2/10.2 MB 90.4 MB/s eta 0:00:00
  Downloading semantic_version-2.10.0-py2.py3-none-any.whl (15 kB)
  Downloading uvicorn-0.30.5-py3-none-any.whl (62 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.8/62.8 kB 4.4 MB/s eta 0:00:00
  Downloading fastapi-0.112.0-py3-none-any.whl (93 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 93.1/93.1 kB 6.5 MB/s eta 0:00:00
  Downloading ffmpy-0.4.0-py3-none-any.whl (5.8 kB)
  Downloading pydub-0.25.1-py2.py3-none-any.whl (32 kB)
  Downloading h11-0.14.0-py3-none-any.whl (58 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 58.3/58.3 kB 4.0 MB/s eta 0:00:00
  Downloading starlette-0.37.2-py3-none-any.whl (71 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.9/71.9 kB 4.2 MB/s eta 0:00:00
  Downloading websockets-12.0-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (130 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 130.2/130.2 kB 9.8 MB/s eta 0:00:00
  Installing collected packages: pydub, websockets, tomlkit, semantic-version, ruff, python-multipart, orjson, h11, ffmpy, aiofiles, uvicorn, starlette, httpcore, httpx, fastapi, gradio-client, gradio
    Attempting uninstall: tomlkit
      Found existing installation: tomlkit 0.13.0
      Uninstalling tomlkit-0.13.0:
        Successfully uninstalled tomlkit-0.13.0
```

```
import gradio as gr


def generate_audio():
    example = minds.shuffle()[0]
    audio = example["audio"]
    return (
        audio["sampling_rate"],
        audio["array"],
    ), id2label(example["intent_class"])


with gr.Blocks() as demo:
    with gr.Column():
        for _ in range(4):
            audio, label = generate_audio()
            output = gr.Audio(audio, label=label)

demo.launch(debug=True)
```

```
/usr/local/lib/python3.10/dist-packages/gradio/processing_utils.py:574: UserWarning: Trying to convert audio automatically from float64 to 16-bit int format.
  warnings.warn(warning.format(data.dtype))
Setting queue=True in a Colab notebook requires sharing enabled. Setting `share=True` (you can turn this off by setting `share=False` in `launch()` explicitly).

Colab notebook detected. This cell will run indefinitely so that you can see errors and logs. To turn off, set debug=False in launch().
Running on public URL: https://6b80dc0b183c1536d5.gradio.live

This share link expires in 72 hours. For free permanent hosting and GU upgrades, run `gradio deploy` from Terminal to deploy to Spaces (https://huggingface.co/spaces)
```






```
Keyboard interruption in main thread... closing server.
Killing tunnel 127.0.0.1:7860 <> https://6b80dc0b183c1536d5.gradio.live
```

Try it out! Download another dialect or language of the MINDS-14 dataset, listen and visualize some examples to get a sense of the variation in the whole dataset. You can find the full list of available languages here.

```
example_minds = load_dataset("PolyAI/minds14", name="ru-RU", split="train")
example_minds
```

```
Generating train split:        539/0 [00:00<00:00, 6179.22 examples/s]
Dataset({
    features: ['path', 'audio', 'transcription', 'english_transcription', 'intent_class', 'lang_id'],
    num_rows: 539
})
```

```
example2 = example_minds[0]
example2
```

```
{'path': '/root/.cache/huggingface/datasets/downloads/extracted/28aa727f91fee90575c34956bab09d1716cfaf460c6afcba86a10f04a7d58b83/ru-RU~LATEST_TRANSACTIONS/6030093cbb1e6d0fbce93a74.wav',
 'audio': {'path': '/root/.cache/huggingface/datasets/downloads/extracted/28aa727f91fee90575c34956bab09d1716cfaf460c6afcba86a10f04a7d58b83/ru-RU~LATEST_TRANSACTIONS/6030093cbb1e6d0fbce93a74.wav',
  'array': array([ 0.        ,  0.        ,  0.        , ..., -0.00048828,
         -0.00073242, -0.00073242]),
  'sampling_rate': 8000},
```

```
        'transcription': 'Здравствуйте я бы хотела пересмотреть свои предыдущие последние операции которые проходили по моей карте прямым помимо ему счёту Покажите пожалуйста операции последних трёх месяцев',
        'english_transcription': 'Hello, I would like to review my previous last transactions that took place on my card directly in addition to his account. Please show the transactions of the last three months',
        'intent_class': 12,
        'lang_id': 12}
```

```
columns_to_remove = ["lang_id"]
example_minds = example_minds.remove_columns(columns_to_remove)
example_minds
```

```
Dataset({
    features: ['path', 'audio', 'transcription', 'english_transcription', 'intent_class'],
    num_rows: 539
})
```

```
def generate_audio():
    example2 = example_minds.shuffle()[0]
    audio = example2["audio"]
    return (
        audio["sampling_rate"],
        audio["array"],
    ), id2label(example2["intent_class"])


with gr.Blocks() as demo:
    with gr.Column():
        for _ in range(4):
            audio, label = generate_audio()
            output = gr.Audio(audio, label=label)

demo.launch(debug=True)
```

```
/usr/local/lib/python3.10/dist-packages/gradio/processing_utils.py:574: UserWarning: Trying to convert audio automatically from float64 to 16-bit int format.
  warnings.warn(warning.format(data.dtype))
Setting queue=True in a Colab notebook requires sharing enabled. Setting `share=True` (you can turn this off by setting `share=False` in `launch()` explicitly).

Colab notebook detected. This cell will run indefinitely so that you can see errors and logs. To turn off, set debug=False in launch().
Running on public URL: https://42a8b5bff43c0f9c6f.gradio.live

This share link expires in 72 hours. For free permanent hosting and GPU upgrades, run `gradio deploy` from Terminal to deploy to Spaces (https://huggingface.co/spaces)
```