**Setting up the Python Environment:**

1. in venv:

python -m pip install --upgrade pip (upgrade python)

pip install torch transformers==4.32.1 datasets

pip install soundfile librosa numpy

2. to install kaldi:

# Clone Kaldi repository

git clone https://github.com/kaldi-asr/kaldi.git

# Build the tools (including sclite)

cd kaldi/tools

extras/check\_dependencies.sh # Install any missing dependencies it reports

make -j $(nproc) # This builds all tools including sclite

**Data format in OGI Kids:**

<https://github.com/OSU-slatelab/OGI-kids-phoneme-recognition/blob/main/ogi_prepare.py>

.wav files are in:

{data\_folder}/speech/scripted/

transcription files are in:

with open(train\_align\_file) as f:

train\_alignments = json.load(f)

with open(valid\_align\_file) as f:

valid\_alignments = json.load(f)

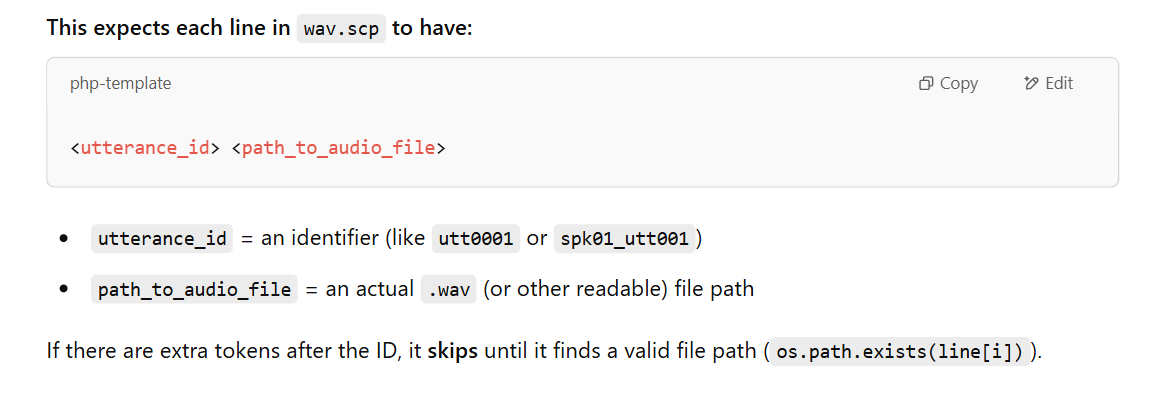
JSON files contain alignments —> probably mappings from sentence ID → transcribed words

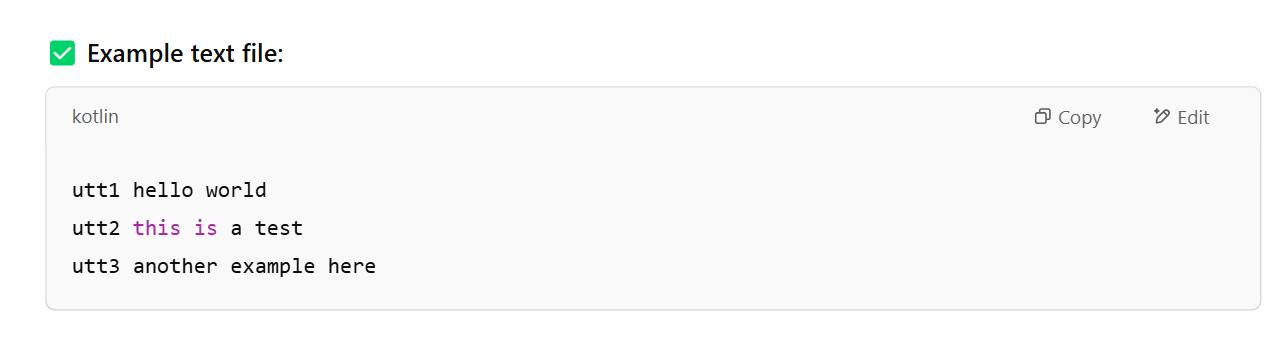
**Data accepted in code:**

wav.scp file: connects every utterance (sentence said by one person during particular recording session) with an audio file related to this utterance.

/src/data/whisper\_loader.py: loads dataset for Whisper, does data augmentation based on arguments in whisper\_small\_train.yaml

From whisper\_loader.py:





**For OGI dataset:**

Stage 1/3: Evaluation of Whisper (in run\_whisper.sh)

/egs/OGI/data/dev/wav.scp

/egs/OGI/data/dev/text

/egs/OGI/data/test/wav.scp

/egs/OGI/data/test/text

/egs/OGI/data/spont\_al/wav.scp

/egs/OGI/data/spont\_al/text

Stage 2: Finetuning of Whisper (in whisper\_small\_train.yaml)

Training data:

/egs/OGI/data/train/wav.scp

/egs/OGI/data/train/text

Validation data:

/egs/OGI/data/dev/wav.scp

/egs/OGI/data/dev/text

tasks, both OGI and MyST:

1. stage 1: evaluation of baseline whisper model

2. run stage 2 and 3: full-finetuning without data augmentation

3. run stage 2 and 3: full-finetuning with VTLP

4. run stage 2 and 3: full-finetuning with SP

5. run stage 2 and 3: full-finetuning with PP

6. run stage 2 and 3: full-finetuning with SA

try combinations:

1. SA + PP

2. SA + SP

3. SA + VTLP

try PIF with VLTP and PP.