ASR - Automatic Speech Recognition

**Converting audio files and performing automatic speech recognition (ASR)**

MP3 to WAV [Conversion](https://colab.research.google.com/drive/1Hq2d8rH018Aa_4L0XiuYJQac6OM1320l?usp=sharing) and Stereo to Mono Conversion

1. **MP3 to WAV Conversion:**

A portion of the script converts MP3 files to WAV format.

**Functions**:

convert\_mp3\_to\_wav(mp3\_path, output\_folder):   
 Converts a single MP3 file to WAV.

convert\_mp3\_to\_wav\_folder(input\_folder, output\_folder):   
 Converts all MP3 files in a folder to WAV format.

**Libraries used:** pydub, os.

Pydub library -> AudioSegment class

1. **Stereo to Mono Conversion:**

Converts stereo WAV files to mono.

**Functions**:

convert\_stereo\_to\_mono\_folder(input\_folder, output\_folder):

Converts all stereo WAV files in a folder to mono.

For each .wav file, the script loads the audio using **AudioSegment.from\_wav().**

If the audio has two channels (stereo), it converts the audio to mono with set\_channels(1) and exports the file to the output folder in .wav format.

**Automatic Speech Recognition with** [**AI4Bharath**](https://colab.research.google.com/drive/1DrtqJDwpZ7f3MCjWfD5Av8RZ_gS1Zkl2?usp=sharing)

**Model**: ai4bharat/indic-conformer-600m-multilingual.

**Libraries**: transformers, torch, torchaudio.

transformers: Library to load pre-trained models.

torch: PyTorch, deep learning library.

torchaudio: Audio handling library that works well with PyTorch for loading and processing audio data.

Resamples the audio to 16kHz if it's not already in that sample rate. Most ASR models expect audio input to be at this sample rate.

1. **Performing ASR with CTC Decoding:**

transcription\_ctc = model(wav, "te", "ctc")

This performs speech-to-text transcription using **CTC** (Connectionist Temporal Classification) decoding. te indicates the language to be transcribed (Telugu in this case), and "ctc" specifies the decoding algorithm.

1. **Performing ASR with RNNT Decoding:**

transcription\_rnnt = model(wav, "mr", "rnnt")

This performs transcription using RNNT (Recurrent Neural Network Transducer) decoding. mr indicates the language (Marathi in this case), with "rnnt" as the decoding algorithm.

**Automatic Speech Recognition with NVIDIA FastConformer**

**Model:** nvidia/stt\_en\_fastconformer\_transducer\_large

**Libraries**: nemo\_toolkit, torch, pydub

**nemo\_toolkit:** Provides access to NVIDIA’s pre-trained ASR (Automatic Speech Recognition) models and utilities for transcription.

**torch:** PyTorch, a popular deep learning framework used by NeMo.

**pydub:** Audio processing library for format conversion and sample rate adjustments.

**ffmpeg:** External tool used by pydub to handle audio file conversion (e.g., MP3 to WAV).

**Automatic Speech Recognition with NVIDIA Canary**

**Model:** nvidia/canary-180m-flash

**NeMo toolkit:** Provides access to the pre-trained model and utilities for data processing and transcription.

**PyTorch:** The deep learning framework used by NeMo for model training and inference.

**pydub:** Audio processing library for handling different audio formats and sample rate conversions.

**ffmpeg :** External tool used by pydub for audio file conversion

Other models

1. **ibm-granite/granite-speech-3.3-8b**
2. **speechbrain/asr-crdnn-commonvoice-indic**
3. **OpenAI’s Whisper**
4. [**UsefulSensors/moonshine-base**](https://colab.research.google.com/drive/1dHjQa3nvFh5c6-iFMHqHRB5Tw-16JPmn?usp=sharing)
5. [**facebook/wav2vec2-conformer-rope-large-960h-ft**](https://colab.research.google.com/drive/1C9Bj_rG39qAY4ITruIelvzfSg3eS1UGc?usp=sharing)