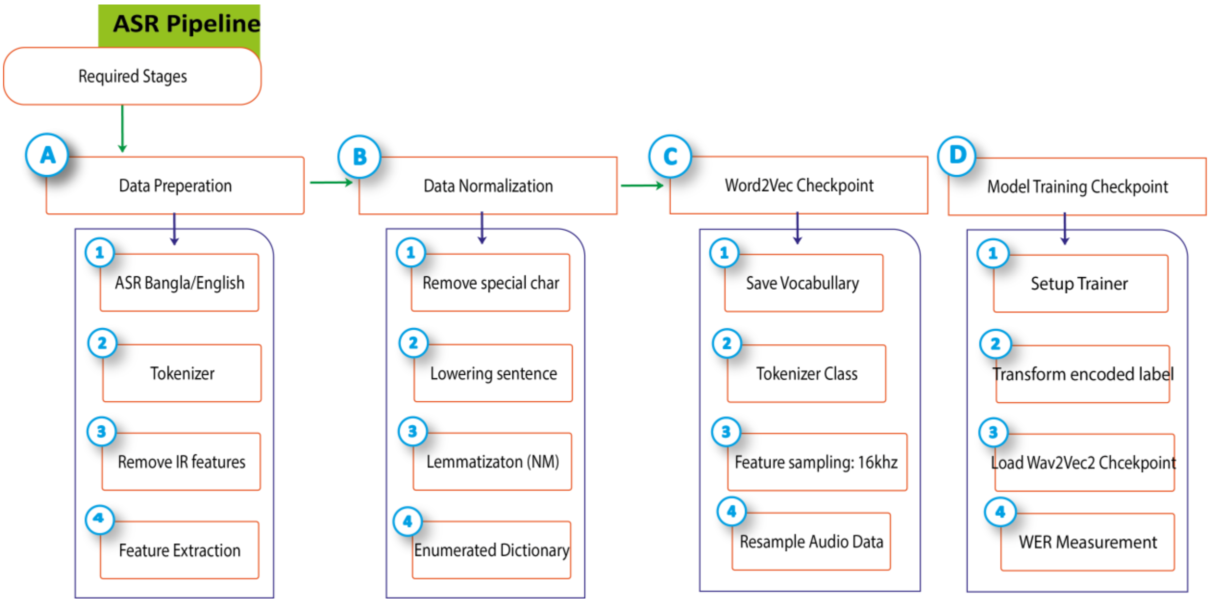
Automatic Speech Recognition (ASR) using Wav2Vec2

The following pipeline can be used while training through the Bangla speech dataset. At this stage, this diagram is designed based on the observation of English speech data. It might work well on top of the target audio data.



Phase 1: Load the target dataset to complete the downstream NLP task

Phase 2: Look over the dataset toward data preparation. Three essential steps have to be carried out here, e.g., tokenizer, removing irrelevant features, and feature extraction. We shall use the tokenizer and feature extraction module from the Wav2Vec model. On the other side, we may not be required additional features, so we need to keep the dataset simple.

Phase 3: The target dataset has to be normalized to speed up the model's performance, e.g., by removing special characters, lowering sentences, lemmatization (not mandatory), and enumerating the dictionary.

Phase 4: Save the vocabulary as a json file. Additionally, the audio data need to be resampled because the common voice is sampled at 48kHz; we need to resample the audio files to 16kHz.

Phase 5: Stage of the training:

- Define a data collator. In contrast to most NLP models, Wav2Vec2 has a much larger input length than output length. E.g., a sample of input length 50000 has an output length of no more than 100. Given the large input sizes, it is much more efficient to pad the training batches dynamically meaning that all training samples should only be padded to the longest sample in their batch and not the overall longest sample.

- Evaluation metric. During training, the model should be evaluated on the word error rate. We should define a compute_metrics function accordingly.
- Load a pretrained checkpoint. We need to load a pretrained checkpoint and configure it correctly for training.
- After fine-tuning the model, we will correctly evaluate it on the test data and verify that it has indeed learned to transcribe speech correctly.

Phase 6: Measuring Word Rate Error (WER)

Phase 7: Measuring model's performance