Lab Assignment 2: Speech Sound Classification

Phoneme Extraction from Speech Signals

Objective: The objective of this experiment is to process a speech signal, extract specific phonemes, and visualize their waveforms while labeling them.
In this experiment, we will:

1. Load a speech signal from the LJ Speech dataset.
2. Preprocess the audio (convert to mono, resample to 16kHz).
3. Use a pre-trained deep learning model (Wav2Vec2) to recognize phonemes.
4. Estimate phoneme time intervals.
5. Extract a phoneme segment from the speech signal based on time intervals.
6. Save extracted phoneme and visualize each selected phoneme segment from the speech waveform
7. Infer about the nature of source of sound for each phoneme.

This experiment will help understand how deep learning-based speech models process spoken language and how phonemes can be visualized from continuous speech.

Expected Outcome: By the end of this experiment, students should be able to:

1. Successfully load and preprocess a speech signal.
2. Run the Wav2Vec2 model to recognize phonemes in the speech signal.
3. Extract a specific phoneme segment from the waveform using time indexing.
4. Label the extracted phoneme by aligning it with recognized phonemes.
5. Visualize the phoneme waveform with its corresponding label in a plot.

Example Output

1. Recognized Phonemes: T EH S T IH NG W AH N T UW (This represents "TESTING ONE TWO" in phonetic format.)
2. Extracted Phoneme between specific time interval and Waveform Plot.

Tools & Libraries to be used

- **Python** for implementation
- **torch, torchaudio** for loading and processing speech signals
- **Librosa** for visualization
- **Wav2Vec2** (Pre-trained model) for phoneme recognition
- from **transformers** import **Wav2Vec2Processor, Wav2Vec2ForCTC**
- **Matplotlib** for waveform plotting

Brief notes:

Wav2Vec2Processor

Wav2Vec2Processor is a utility class that prepares raw speech signals for the Wav2Vec2 model and converts model outputs into readable labels. It normalizes raw audio waveforms, pads or truncates audio sequences, converts audio into model-ready tensors and decodes model predictions (token IDs) into phonemes or characters.

Wav2Vec2ForCTC is a pre-trained Wav2Vec2 neural network designed for sequence labeling tasks such as phoneme or speech recognition. There are three key components: (1) Wav2Vec2 encoder that learns rich speech representations from raw audio, (2) linear classification layer that maps representations to phoneme or character probabilities, (3) Connectionist Temporal Classification (CTC) loss that enables alignment-free training between speech frames and output labels. Speech and phoneme sequences have different lengths, and CTC allows the model to learn alignments automatically. This is a loss function and decoding framework used to train sequence-to-sequence models when the input and output sequences have different lengths and are not explicitly aligned.

References:

1. 2006.11477

https://doi.org/10.48550/arXiv.2006.11477

2. icml_2006.pdf

A. Graves, S. Fernández, and F. Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proc. of ICML, 2006.