Utterance to Words Splitting

(Project overview)

...Suman Samui

General steps to be followed:

- Run Google cloud speech-to-text or any standard ASR API on the speech corpus
- Calculate the utterance level WER by comparing the obtained hypothesis with the ground-truth transcript.
- If WER==0:
  - Perform utterance-to-word trimming based on time-tags

Possible Challenges:

- For most spoken languages, the boundaries between lexical units are difficult to identify. There is hardly any pauses in between two words.
- Inter-word spaces used by many written languages, would rarely correspond to pauses in their spoken version,
- Only in case of very slow speech, the speaker deliberately inserts those pauses.
- In normal speech, one typically finds many consecutive words being said with no pauses between them, and often the final sounds of one word blend smoothly or fuse with the initial sounds of the next word. (Co-articulation effect)

Possible solutions:

Employ a VAD based voting criteria at the word-boundaries to select a valid word based on hypothesis:

1. Compute the RMS level of the wave.Set the Threshold as less than ~10-20 dB.
2. Run the Vad on the wave and get sample wise status whether the sample is speech or silence
3. Make a rectangular window of duration (10 ms) around the start and end point of the word
4. If the total number of silence samples at the word-boundaries (within the specified window) is greater than 50% of the total number of samples in the window, we can say the word-boundary is in silence.

Problem with VAD based voting hypothesis:

- There is always a fair chance that VAD-based hypothesis is being fooled by the unvoiced (low-energy) region of speech because VAD is highly sensitive to the threshold level that we set. In other words, VAD based hypothesis end up with by selecting a unvoiced region which has also very small short-time energy (resemblance to silence)
- To mitigate this problem, we can combine other criterion along with VAD-based voting criteria.

Word boundary selection criterion:

- We must select a word-boundary which has low STE (short-time energy) and low ZCR (Zero crossing rate)
- **Low STE =>** STE within the window @ word-boundary is less than the average of STEs over the word-interval
- **Low ZCR =>** ZCR within the window @ word-boundary is less than the average of ZCRs over the word-interval