Project I

Tang Quoc Thai

Supervisor: Assoc. Prof. Quan Thanh Tho

Bahnaric Phoneme Segmentation

Ho Chi Minh City
University of Technology

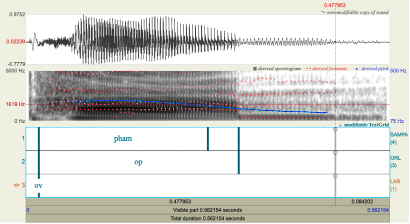
07 Dec 2023

Motivation

- **Objective**: Empower Bahnaric language speakers, fostering communication within their ethnic community and with other ethnic groups.
- **Significance of Phoneme Segmentation**:
  - Create a precise phoneme-level mapping for the Bahnaric language.
  - Enable the development of advanced Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) models.
- **Overall Goal**: Contribute to the empowerment and connectivity of Bahnaric ethnic communities through targeted advancements in speech processing.

Bahnaric phoneme

- The phoneme sample consists of single words, and each word is pronounced by a native speaker.
- The beginning and ending time of each phoneme marked by the ‘ov’ and ‘op’ label respectively.



Feature engineering

The following features are extracted from the audio clips:

- **MFCC** (Mel Frequency Cepstral Coefficients): These are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear ‘spectrum-of-a-spectrum’).
- **Zero Crossings**: This is the rate at which the signal changes from positive to negative or back.
- **Mel Spectrogram**: A Mel Spectrogram is a spectrogram where the frequencies are converted to the Mel scale.
- **Harmonics**: These are integer multiples of the base frequency in a sound. They contribute to the perceived timbre of a sound.

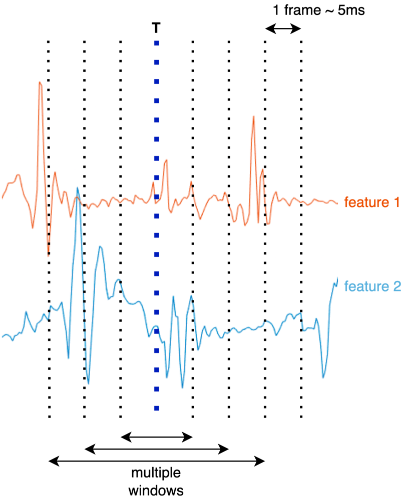
Feature engineering

- **Spectral Centroids**: It indicates where the 'center of mass' of the spectrum is located. It is used in digital signal processing to identify the brightness of a sound.
- **Chromagram**: A chromagram is a graphical representation of the chroma of a signal. In music, the chroma of a note is its position within the octave of the twelve-note chromatic scale.
- **Tempo BPM** (Beats Per Minute): This is a measure of tempo in music, indicating the number of beats occurring in one minute.
- **Spectral Bandwidth**: This is the difference between the highest and lowest frequencies in a continuous band of frequencies. It can be used to identify the smoothness of a sound.

Pseudocode of Feature engineering

```
for audio_clip in audio_clips:
  audio_clip_features = []
  for frame in audio_clip:
    frame_features = []
    for feature in acoustic_features:
      for window_length in range(85, 126, 10):
        windowed_audio = get_window(frame, window_length)
        mean_value = calculate_mean(window, feature)
        frame_features.append(
          {"feature_name_window_length": mean_value})
  audio_clip_features.append(
    {"audio_clip_name": frame_features})
```

Pseudocode of Feature engineering

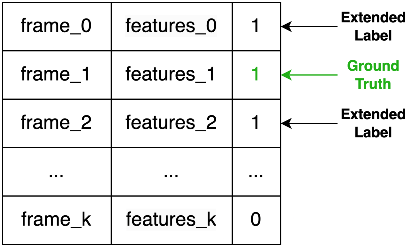


Labels

- The ‘ov’ and ‘op’ labels are extracted from the TextGrid files.
- The information obtained reveals the timestamps of these markers in milliseconds. Consequently, it is necessary to convert these timestamps into frame indices, with each frame corresponding to a 5ms interval.
- **A strong assumption** has been made: the neighboring frames of the ‘ov’ and ‘op’ labels are also labeled as ‘ov’ and ‘op’ respectively.

Training

- Each frame is treated as a data point, and the label is either 0 or 1.
- The extended labels are the the neighboring frames of the ‘ov’ or ‘op’ labels.
- LGBMClassifier is used to train two separate models for the ‘ov’ and ‘op’ labels.



frame_0	features_0	1
frame_1	features_1	1
frame_2	features_2	1
...	...	...
frame_k	features_k	0

Prediction

- The trained model is used to predict the probability of each frame being labeled as ‘ov’ or ‘op’.
- The only frame with the highest probability and larger than 0.5 is selected as the ‘ov’ or ‘op’ label.

frame_0	features_0	0.03
frame_1	features_1	0.52
frame_2	features_2	0.43
...	...	...
frame_k	features_k	0.14

Max of probabilities &
Larger than threshold

Result

- The model is trained on 80% of the data and tested on the remaininng 20%.
- The model achieves an accuracy of 0.67 and 0.79 for the 'ov' and 'op' labels respectively.

	Precision	Recall	Accuracy
op	0.78	0.38	0.67
ov	0.76	0.60	0.79

Thank You