Pitch synchronous residual MFCC for language identification.

Team-17

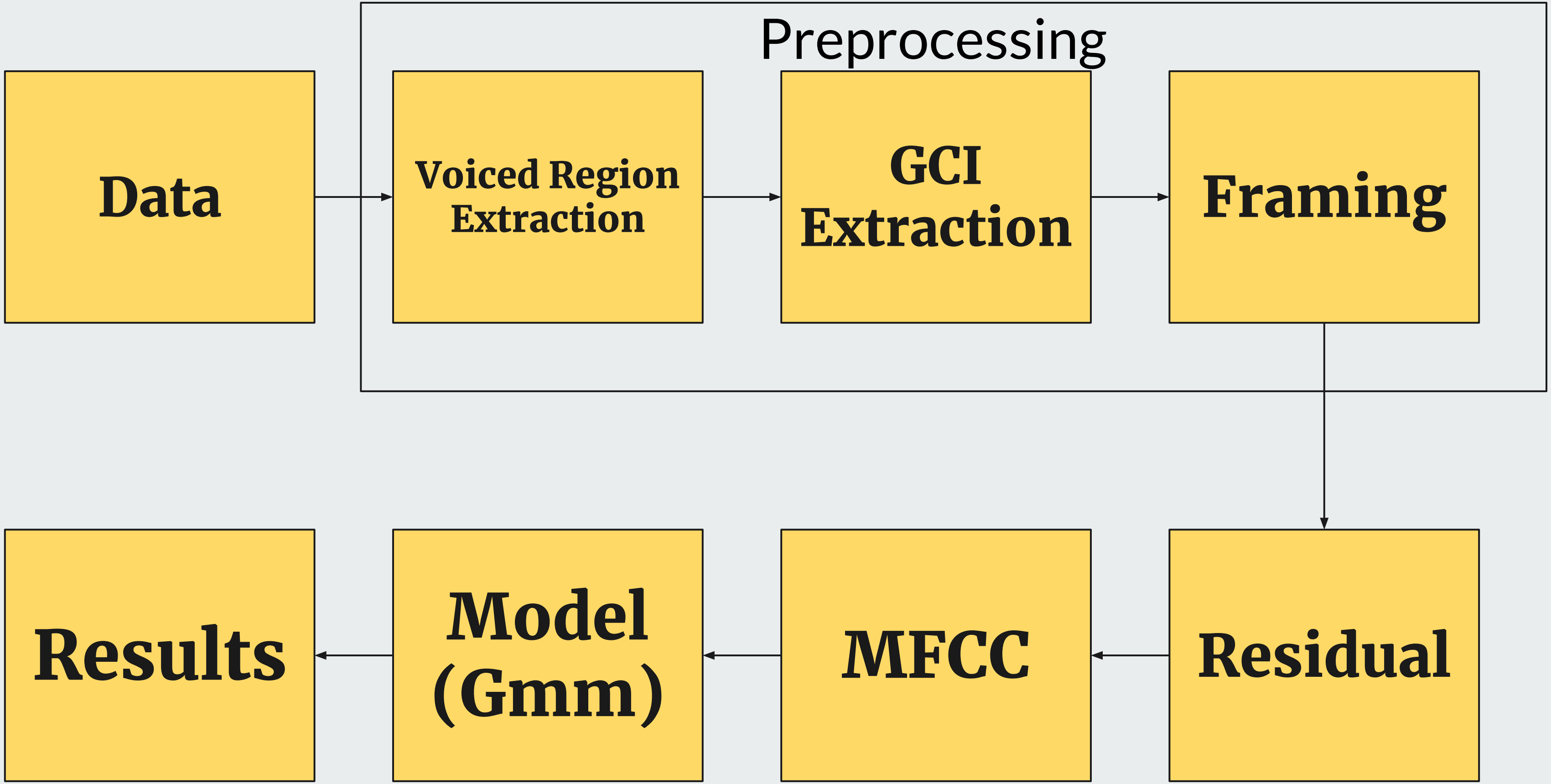
Rhuthik - 2019112013
Vaibhav - 2019112019
Sanjai Kumaran - 2019112012

Introduction

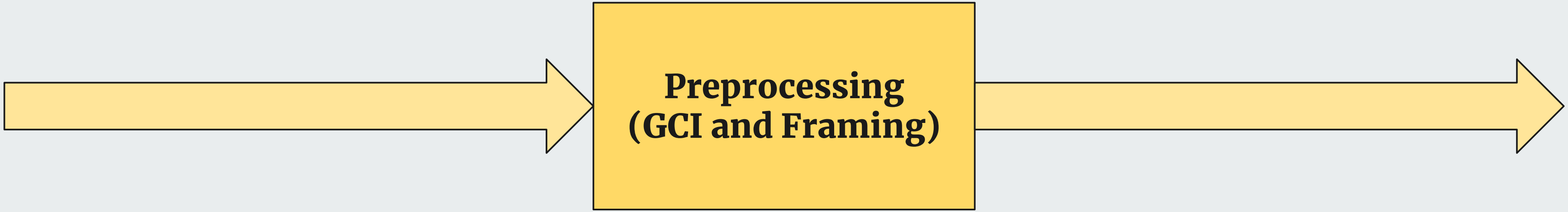
- Language Identification is the problem of identifying the language given an audio clip of a speaker irrespective of their Physical attributes(gender,accent etc).
- The primary Task is to prepare Individual GMM models for each given language and fine tune it to improve the accuracy score.
- Generally we use fixed window length during framing but in Pitch synchronous analysis we extract the locations of pitch and get frames between the pitch cycles.
- Residual is the error obtained between the predicted and the actual speech signal using LP analysis.
- MFCC is used as the feature extraction step which is done by using mel frequency filters and cepstral analysis.

Motivation

- Spoken language has a wide variety of variation like accent, dialects, etc, hence language identification is one of the primary challenges in this domain.
- Pitch synchronous frames defined by the glottal closure instants are used to extract speech parameters.
- Mel-frequency analysis of speech os based on human perception experiments. MFCC features highlight vocal tract information.
- Even with limited dataset GMM assured to give results with high accuracy as compared to Neural Networks. Hence GMM’s are employed in this task.

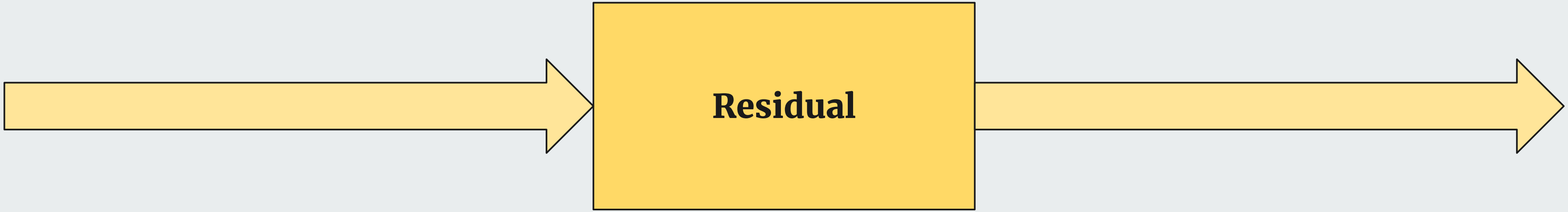


Preprocessing
Data
Voiced Region Extraction
GCI Extraction
Framing
Residual
MFCC
Model (Gmm)
Results



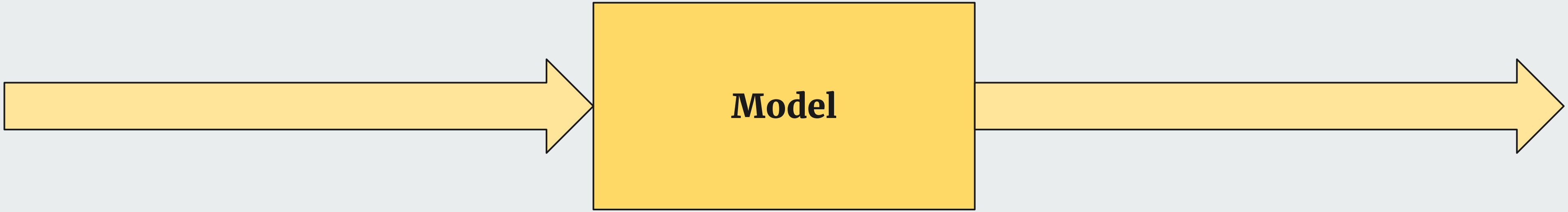
Preprocessing (GCI and Framing)

- There are 7 languages(Odia, Assamese, manipuri, etc..) which are divided randomly into 80:20 Train-Test ratio.
- Voiced region carries most information in a speech signal.
- We extract voiced regions from the data and apply preprocessing on it.
- We find the Glottal closure instants(GCI) using ZFF.
- Pitch cycles are identified and are extracted in the form of frames for each audio file.



- Residual is nothing but error obtained from the LP analysis equation
- Residual = Original Speech Signal - Signal reconstructed from LP analysis
- Residual is applied on each frame of the audio signal(pitch cycles)

- To all the frames after applying Residual we apply MFCC.
- We get 13 features for each frame
- We concatenate all these frames of MFCC’s into a single file for training the models.
- This final file is the Feature representation of the audio file.

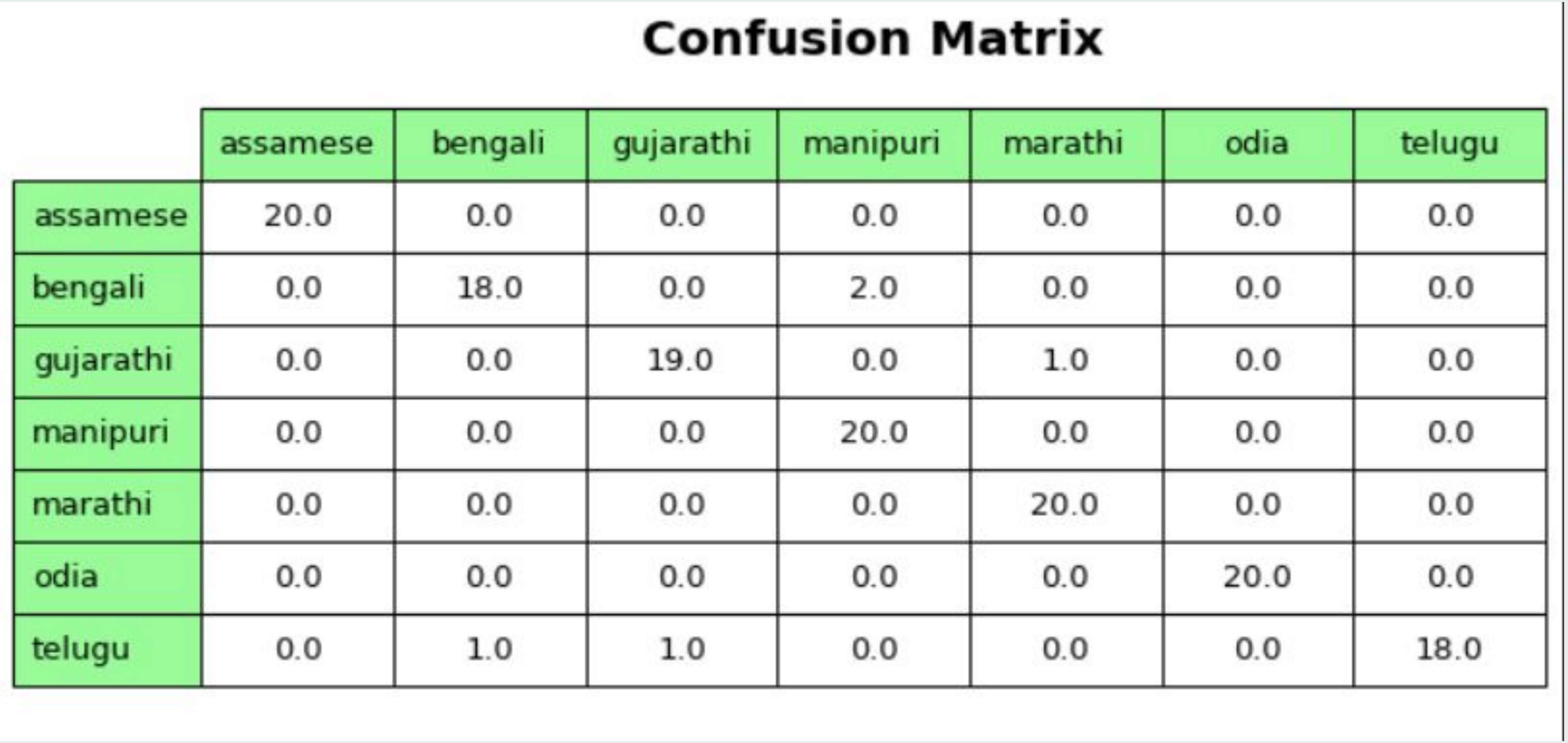


Model

- All the Features files are combined form a single Vector for each language.
- Gaussian Mixture Model for each language is trained using this feature representation of the corresponding language.
- We experimented with 8,16,32,64,128,256 Guassian components in the model.
- Testing was done on each of these models.

Results

Gaussians / epochs	8	16	32	64	128	256
50	92.14	92.85	94.28	93.57	95.0	96.42
100	92.85	93.57	95.0	95.0	95.0	96.42
200	92.85	94.57	94.28	94.28	94.28	95.71



Confusion Matrix

	assamese	bengali	gujarathi	manipuri	marathi	odia	telugu
assamese	20.0	0.0	0.0	0.0	0.0	0.0	0.0
bengali	0.0	18.0	0.0	2.0	0.0	0.0	0.0
gujarathi	0.0	0.0	19.0	0.0	1.0	0.0	0.0
manipuri	0.0	0.0	0.0	20.0	0.0	0.0	0.0
marathi	0.0	0.0	0.0	0.0	20.0	0.0	0.0
odia	0.0	0.0	0.0	0.0	0.0	20.0	0.0
telugu	0.0	1.0	1.0	0.0	0.0	0.0	18.0

Contributions

Sanjai:

- Wrote script to remove Silence and unvoiced regions of the data.
- Splitting the data.
- Preparing slides

Rhuthik:

- Finding LP residual.
- Epoch extraction.
- Preparing slides

Vaibhav:

- MFCC feature extraction.
- Configuring MFCC for different pitch periods.
- Preparing slides

Thank You