Speech Signal Processing Project Report
Excitation features using GVV for language identification

Pallav Subrahmanyam Koppisetti (2020102070)
Anubhav Pal (2020112012)
Pratyush Mohanty (2020101005)

Abstract

Speech signal processing's automatic language recognition has long been a difficult problem and a key field for research. It involves determining a language from a seemingly random spoken
speech. A speech input is passed to the Language Identification module, which produces the speech's language identity.
The process of detecting language from an audio clip by an unknown speaker, regardless of gender, manner of speaking, and distinct age speaker, is defined as spoken language identification (SLID). The considerable task is to recognize the features that can distinguish between languages clearly and efficiently.

Introduction

Glottal Volume Velocity (GVV) stands for glottal volume velocity and is an excitation signal used to model our vocal track during speech production. Since, we know that speech production involves the quasi-periodic vibration of the vocal folds. Thus studying this difference in air pressure created helps us in studying speech production in a more detailed manner and would help us replicate this process artificially in a more accurate fashion. This project implements one such approach for Language Identification using the GVV excitation features.The Time and Frequency Domain excitation features are extracted from the GVV waveform , which is obtained by passing the corresponding LP residual signal through a low pass filter.

Gaussian Mixture Models (GMM) is a machine learning algorithm used to classify data into different categories based on the probability distribution. It used to represent any data set that can be clustered into multiple Gaussian distribution).
Our model uses **Maximum Likelihood Estimation** and maximizes the posterior probabilities of all the training utterances that are correctly classified. They help in capturing the data distribution in the feature space and results in minimal error rate.

As mentioned, developing such LID systems can be used many upcoming speech applications such in ASRs and speech to speech translation(one language speech to another).

Motivation:

The motivation behind this project is to detect the language of an unknown speech signal based on it's excitation features which are specific to the pronunciation of phones and is independent of the vocal track properties of the speaker.

Flow of the LID system

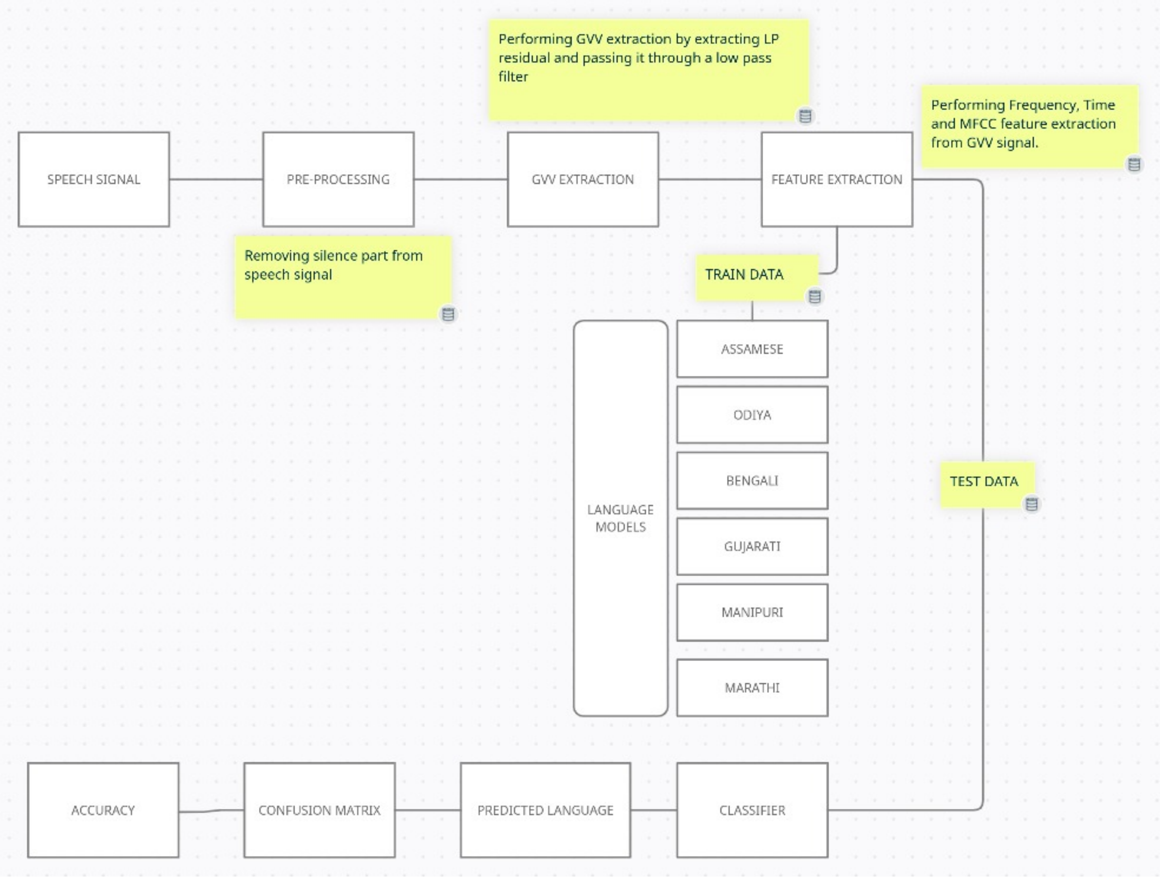

Performing GVV extraction by extracting LP residual and passing it through a low pass filter
Performing Frequency, Time and MFCC feature extraction from GVV signal.
SPEECH SIGNAL
PRE-PROCESSING
GVV EXTRACTION
FEATURE EXTRACTION
Removing silence part from speech signal
TRAIN DATA
ASSAMESE
ODIYA
BENGALI
GUJARATI
MANIPURI
MARATHI
LANGUAGE MODELS
TEST DATA
ACCURACY
CONFUSION MATRIX
PREDICTED LANGUAGE
CLASSIFIER

Figure 1: Flowchart of the method

Procedure:

We split our total LID dataset in the ratio of 80:20 for training and testing respectively and then send it for feature extraction.

Before feature extraction, we first pre-process the signal by removing the silenced part as they carry no useful information.

We then combine all the audio files of a particular language into one audio file and compute the GVV waveform for the signal and use it for our training feature extraction process. The following features are extracted from the glottal waveform :

- Harmonic richness factor
- H1-H2
- MFCC features
- Time domain features such as OQ,ClQ,NAQ,AQ and SQ

We then use these feature vectors to train a GMM model for each langauge where the probability of occurance of each of the feature vectors is expressed as a weighted combination of gaussian distribution.

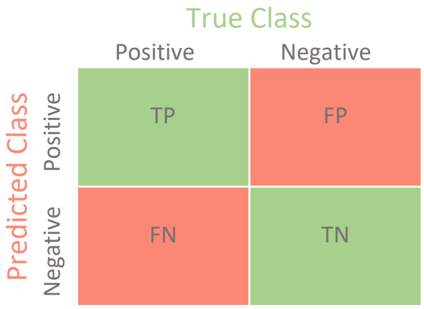
This is followed by extracting the GVV waveform of each and every testing audio file and extracting the excitation features from them and storing these feature vectors , which would be fed to our GMM model for prediction.

During the testing phase, we get the avergae likelihood of the testing file from each of the GMM model and the predicted class would be the corresponding model with the maximum likelihood.

We construct a confusion matrix based on the outputs from each of the testing data in order to assess the accuracy of our LID system.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} ,$$

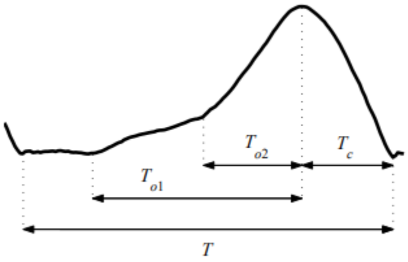
where TP= True Positive, TN= True Negative, FP= false positive and FN= False Negative.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Time domain features :

- The instant of closure is followed by a relatively short closed phase with nearly constant flow. Then, the flow starts to increase gently. This phase ends abruptly at a knee that begins a segment with more rapid flow increase. Thus, there are two instants that could be considered instants of glottal opening.



The length of the glottal cycle is denoted by T. The interval from the primary opening to the instant of maximum flow is indicated by T_{o_1} and the interval from the secondary opening to the instant of maximum flow by T_{o_2}. The closing phase length is denoted by T_{cl}.

1) Open Quotient: The Open Quotient (OQ) was defined as the duration of the cycle during which the vocal folds remain open (opening plus closing phase) divided by the duration of the entire cycle.

$$OQ_1 = \frac{T_{O_1} + T_{cl}}{T} \tag{1}$$

$$OQ_2 = \frac{T_{O_2} + T_{cl}}{T} \tag{2}$$

2) Close Quotient: Closing quotient (ClQ) was defined as the duration of the closing phase (not including the closed phased) divided by the duration of the entire cycle.

$$ClQ = \frac{T_{cl}}{T} \tag{3}$$

3) Speed Quotient: Speed Quotient (SQ) is defined as the ratio of rise time (increased contact) to fall time (decreased contact).

$$SQ_1 = \frac{T_{o_1}}{T_{cl}} \tag{4}$$

$$SQ_2 = \frac{T_{o_2}}{T_{cl}} \tag{5}$$

4) NAQ: The absolute value of amplitude quotient can be viewed as an indirect measure reflecting the viscoelastic property/stiffness of the vocal folds that is determined by the shape and amplitude of the GVV.

$$NAQ = \frac{f_{ac}}{d_{min}.T} \tag{6}$$

5) QQQ: The quasi-open quotient (QOQ) is a frequently used correlate of OQ which involves derivation of the quasi-open phase based on amplitude measures of the glottal pulse.

6) OQ1, OQ2: OQ1 and OQ2 are the open quotients calculated from the so-called primary and secondary openings of the glottal flow, respectively.

7) SQ1, SQ2: Speed Quotients calculated from the primary and the secondary openings, respectively.

If a flow pulse did not show two opening instants, the two opening marks were positioned at the same instant. In this case, $OQ1 = OQ2$ and $SQ1 = SQ2$.

Note: These time domain approaches, however, can be seriously impaired if there is negative signal polarity or phase distortion in the signal which can often occur in less than ideal recording conditions.

Frequency domain features :

1. **Harmonic Richness Factor:** The ratio between the amplitudes at the harmonics in the glottal waveform and the component's amplitude at the fundamental frequency is known as the harmonic richness factor (HRF). This parameter represents the spectral tilt of the glottal flow and has been used to identify different phonation types.

$$HRF = \frac{\sum_{i\geq2} H_i}{H_1}, \; where\ H_i\ represents\ the\ amplitude\ of\ the\ i^{th}\ harmonic.$$

1. **H1-H2:** The amplitute difference between 1st Harmonic and 2nd Harmonics. This measurement has been used as a glottal parameter as it is reported that changes in the openquotient of the glottal cycle produce a corresponding change in H1-H2.

$$H1 - H2 = H_1 - H_2 \ , where\ H_1\ and\ H_2\ are\ the\ amplitudes\ of the\ 1^{st}\ and\ 2^{nd}\ harmonic.$$

This is achieved by applying FFT on our GVV signal and computing the fundamental frequency from the FFT spectrum, followed by computing its harmonics that are present and their corresponding amplitudes, which are fed to our frequency domain feature extraction functions.

MFCC of GVV :

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) 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").

Observation plots :

Time Domain:

Following are the GVV waveform with its corresponding Time-Domain parameter for the 7 languages: Telugu, Marathi, Odiya, Bengali, Manipuri, Assamese, Gujarati.

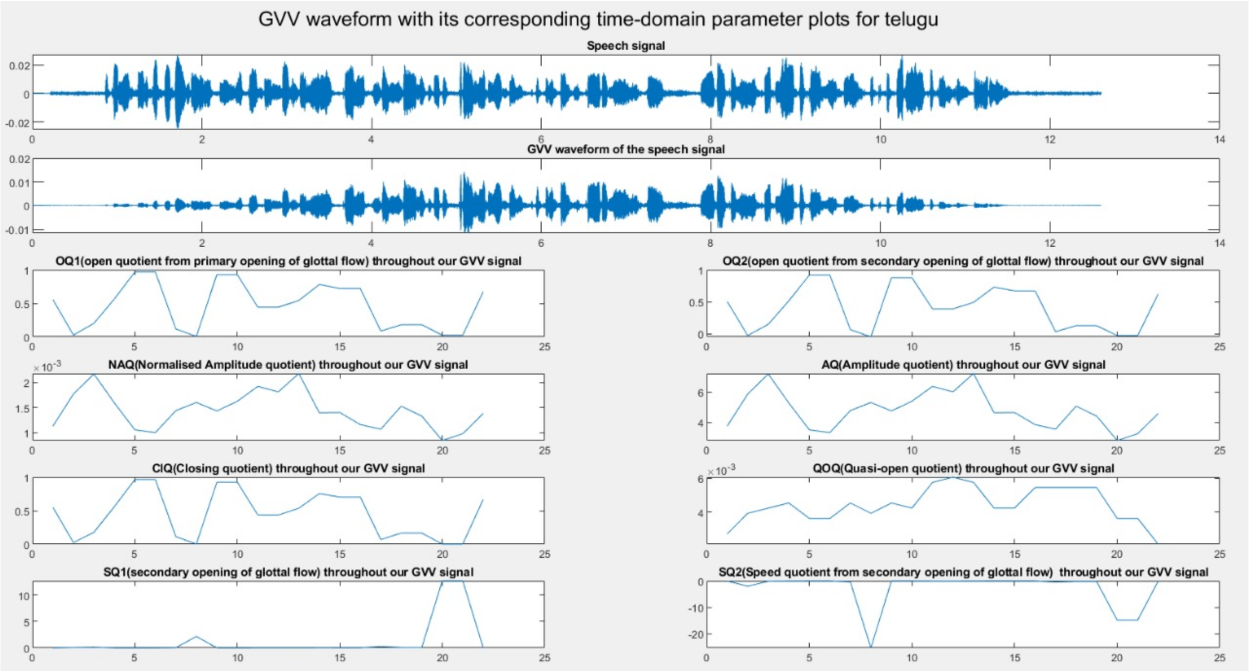


Figure 2: Telugu

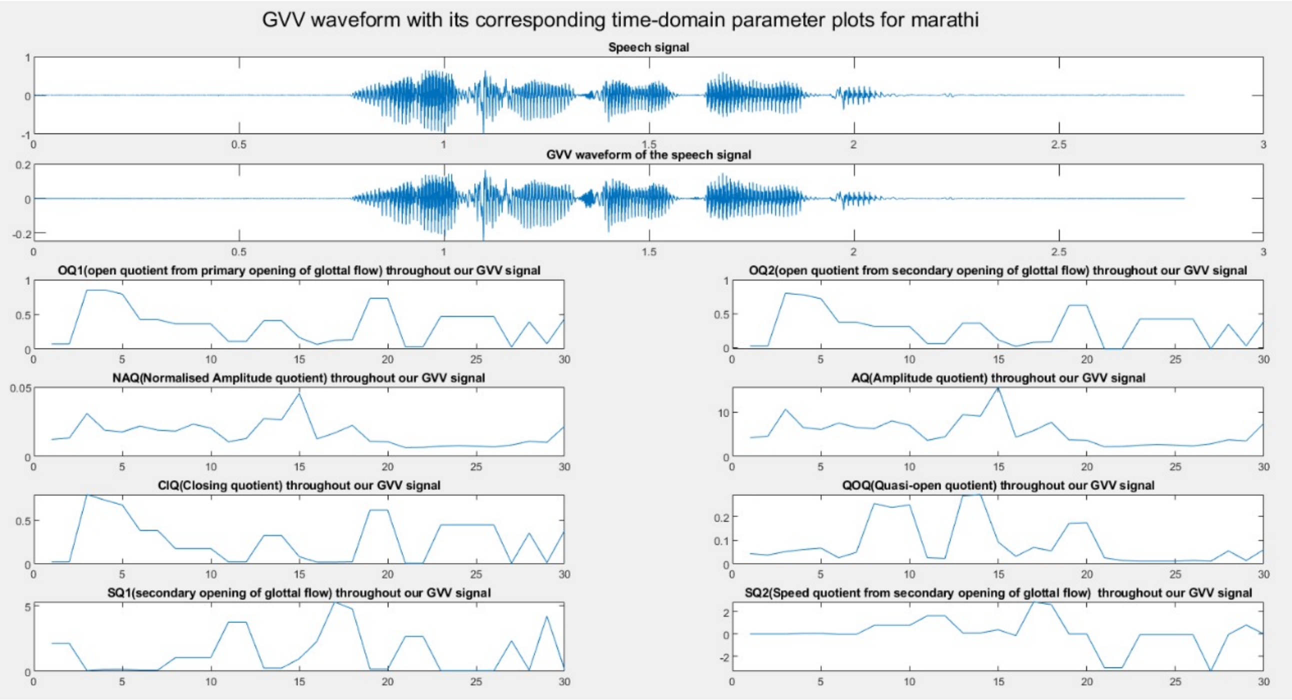


Figure 3: Marathi

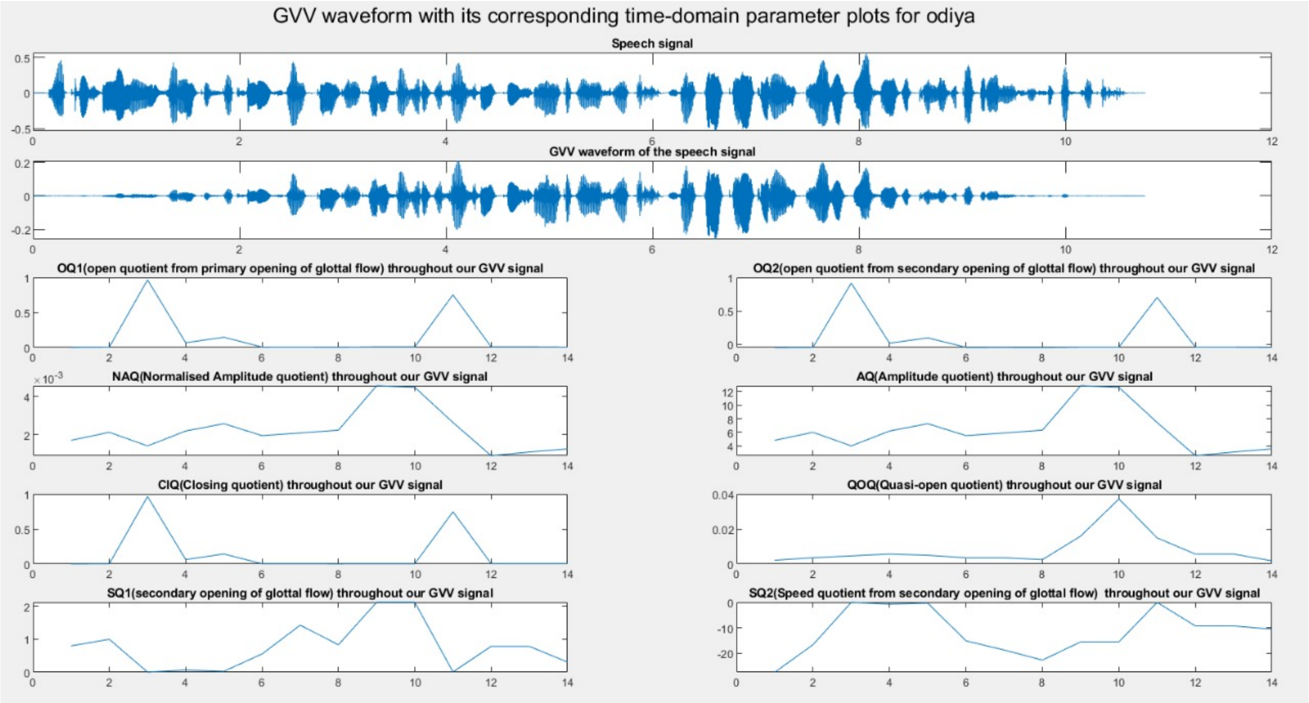


Figure 4: Odiya

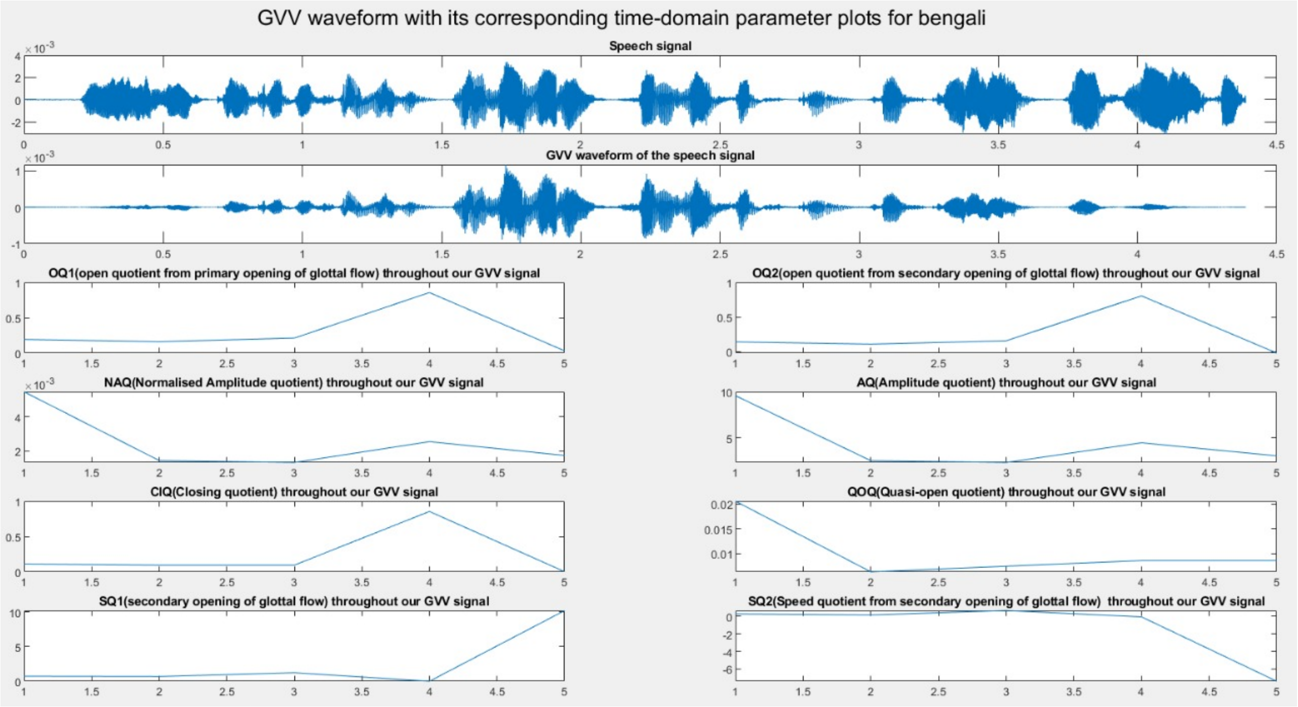


Figure 5: Bengali

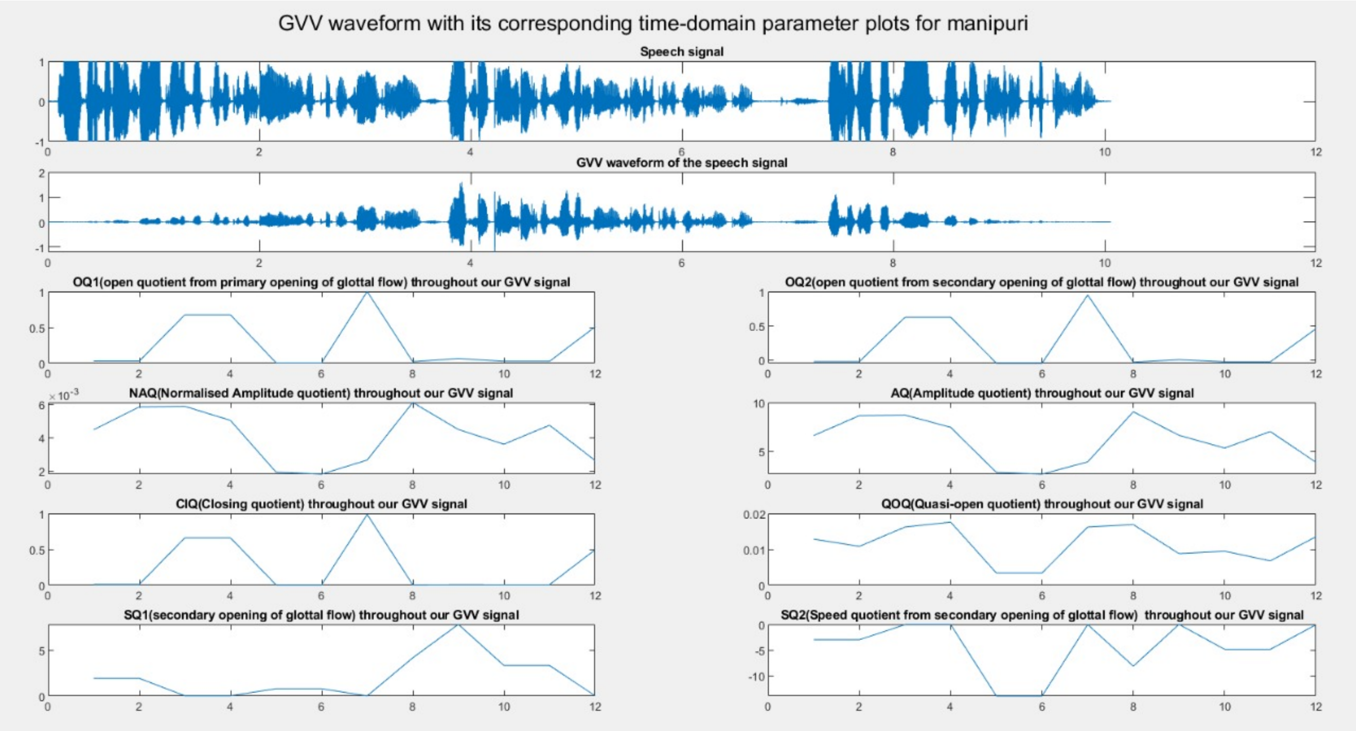


Figure 6: Manipuri

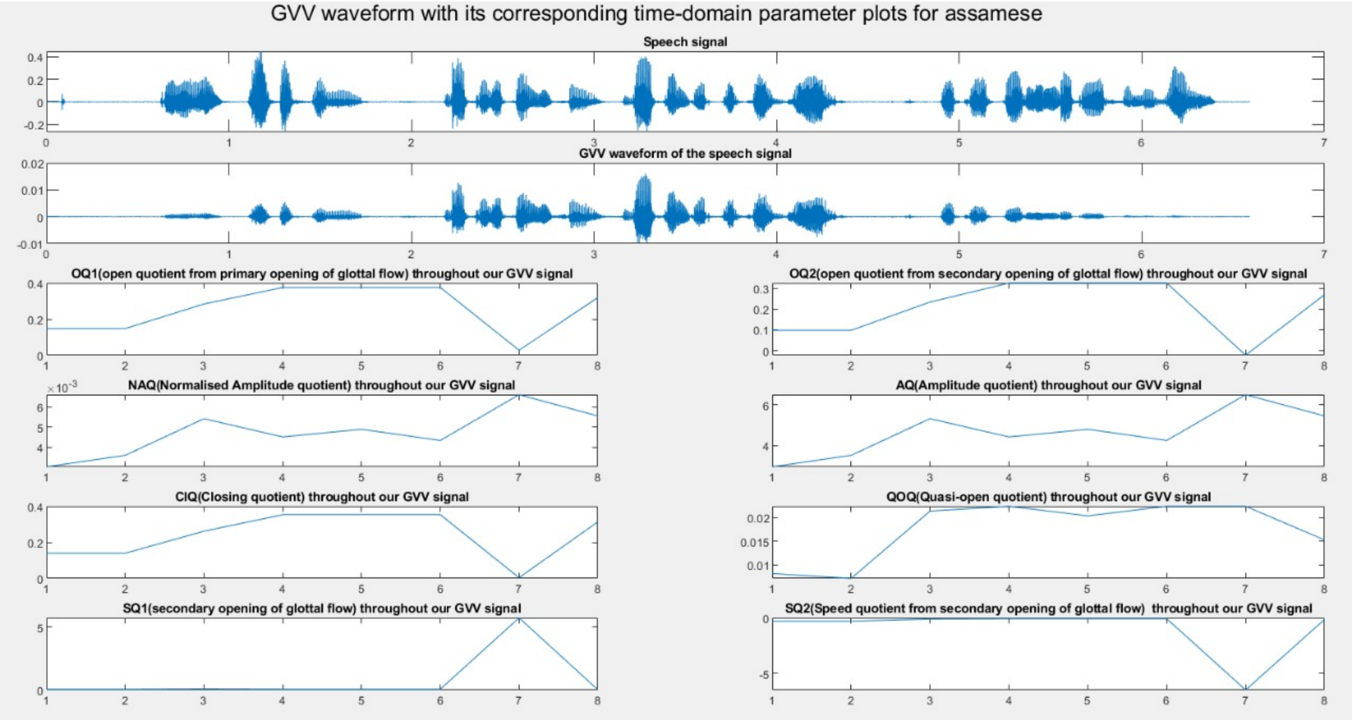


Figure 7: Assamese

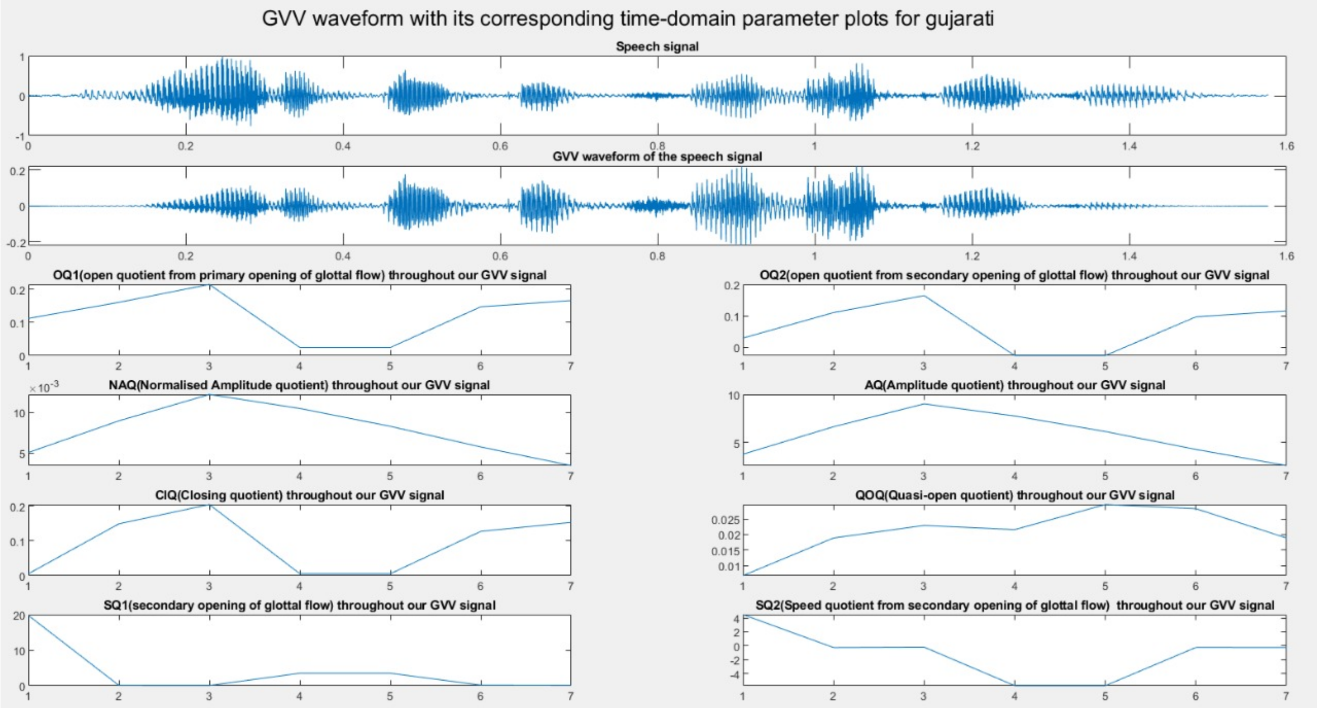


Figure 8: Gujarati

Frequency Domain:

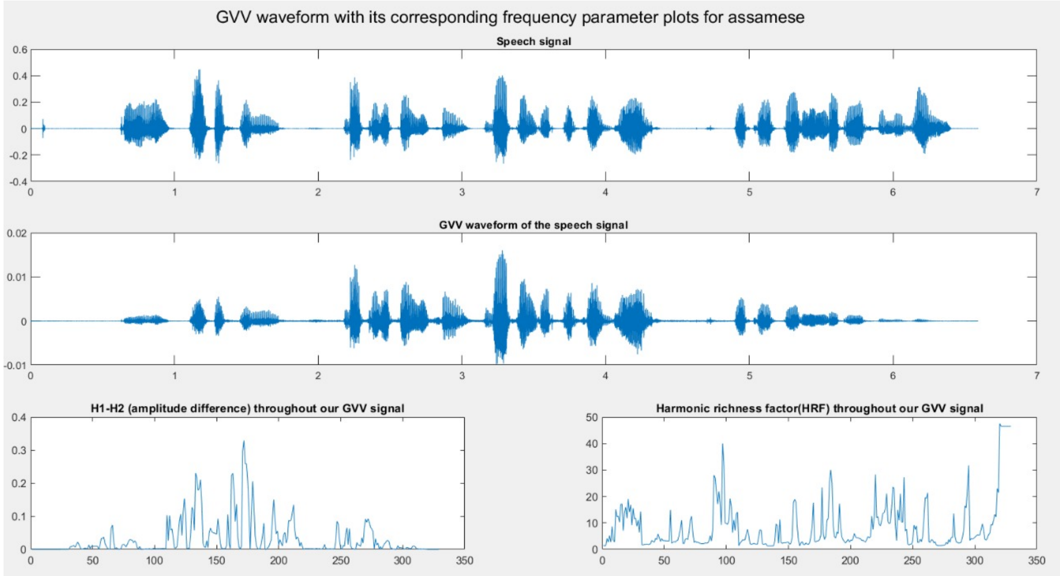


Figure 9: Assamese

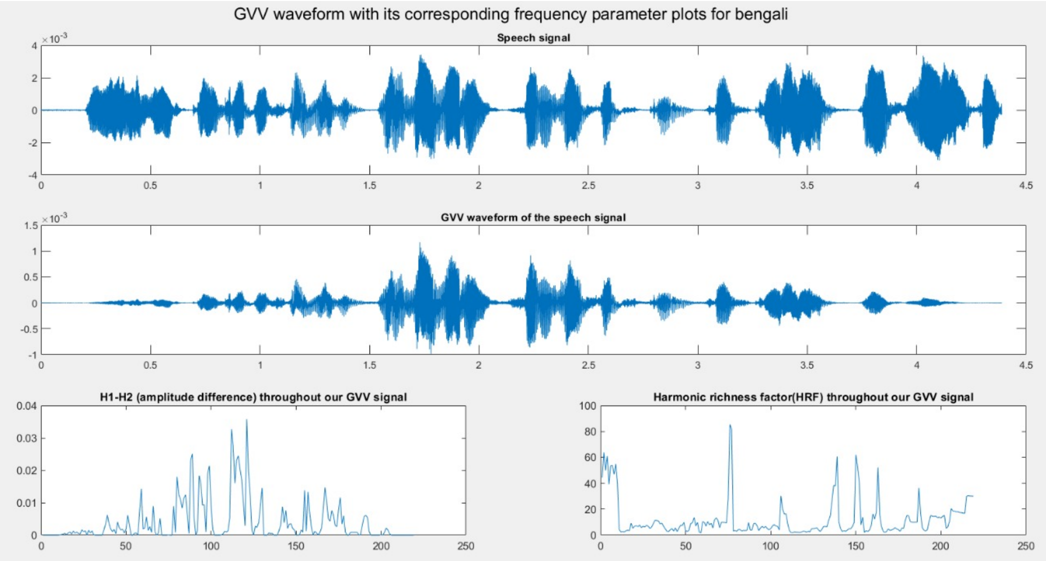


Figure 10: Bengali

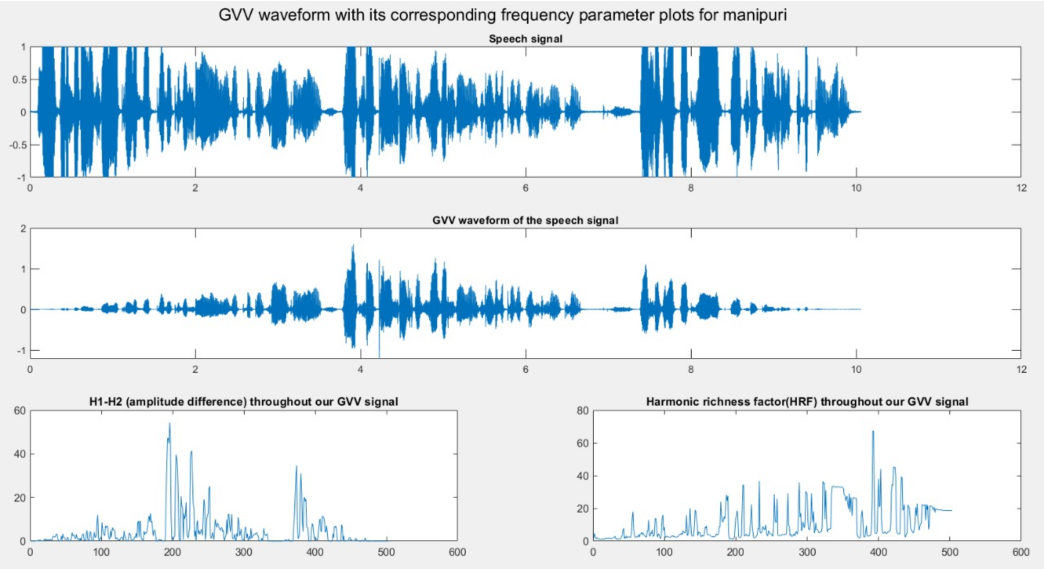


Figure 11: Manipuri

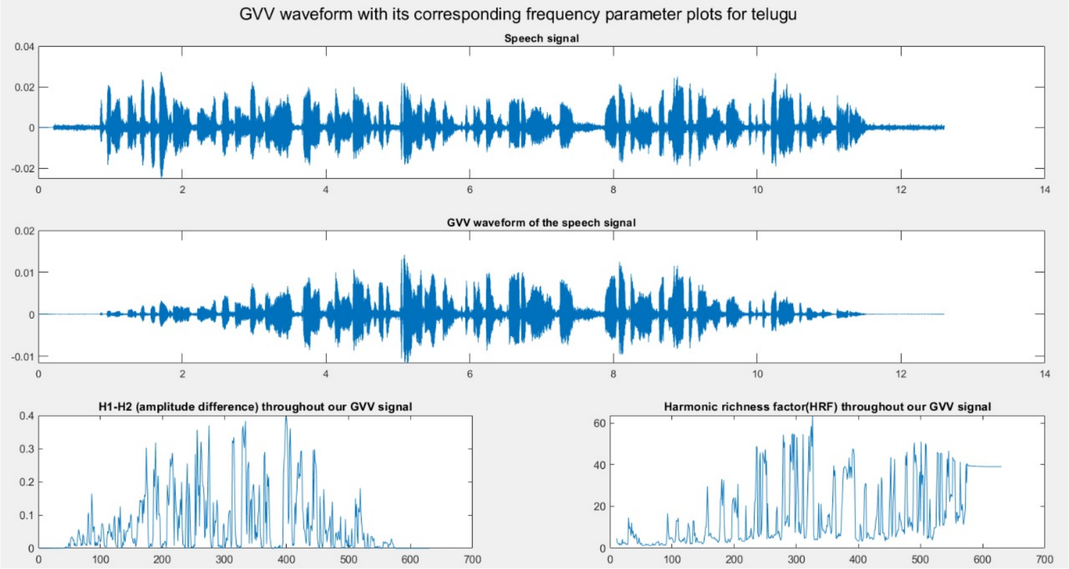


Figure 12: Telugu

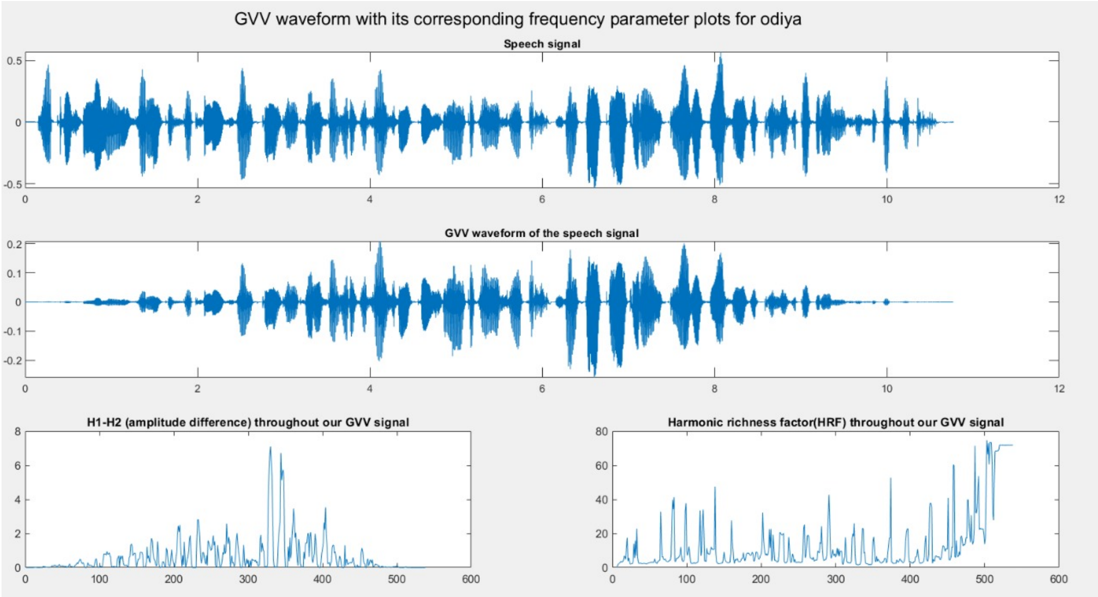


Figure 13: Odiya

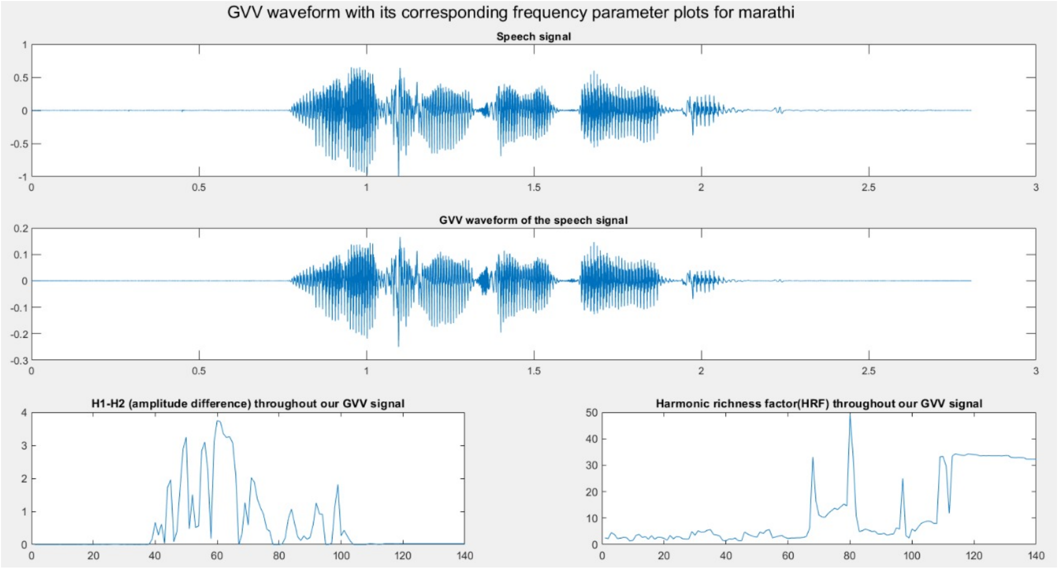


Figure 14: Marathi

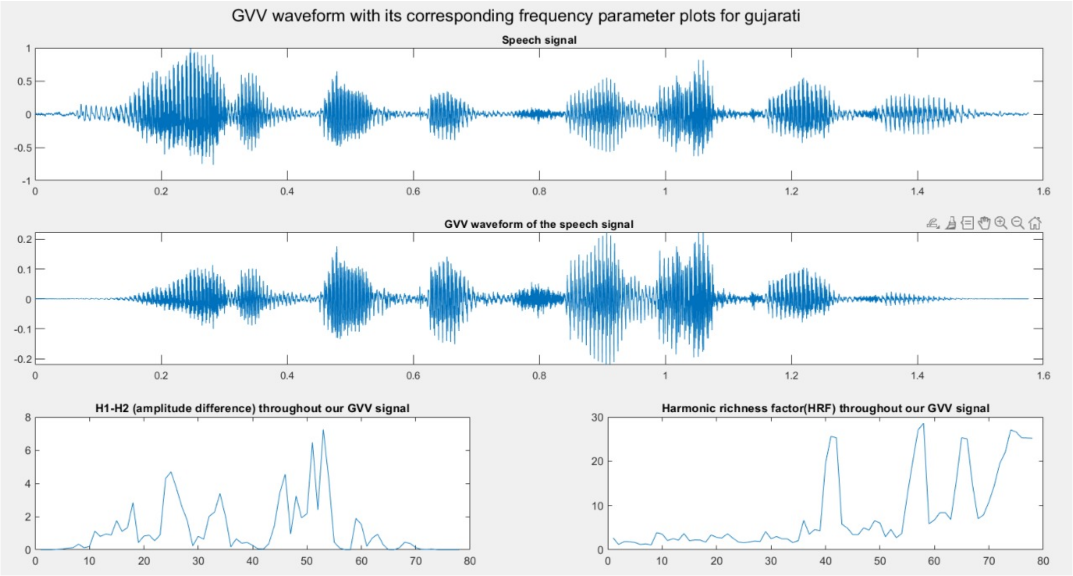


Figure 15: Gujarati

MFCC:

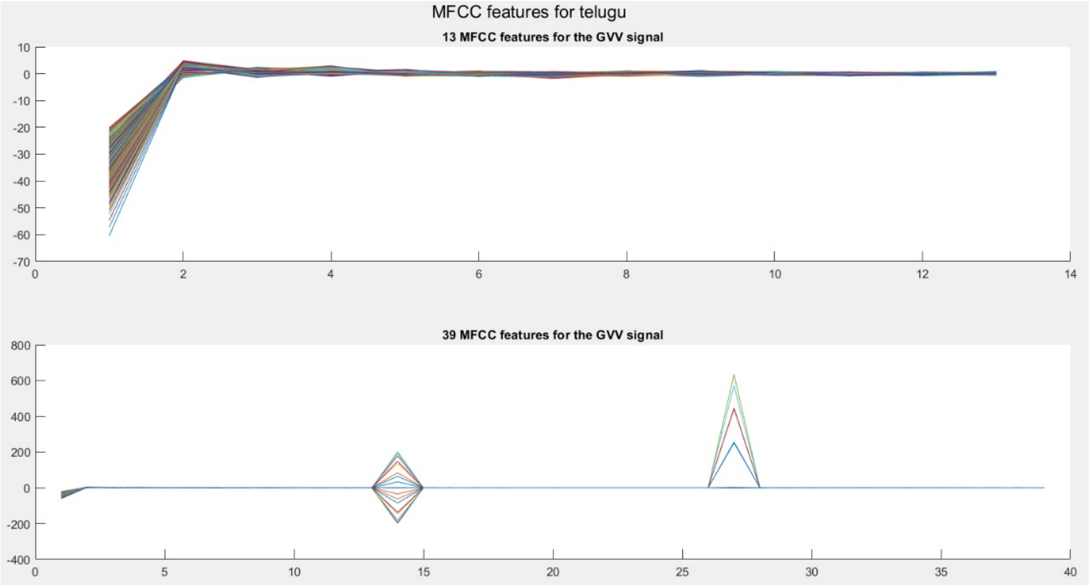


Figure 16: Telugu

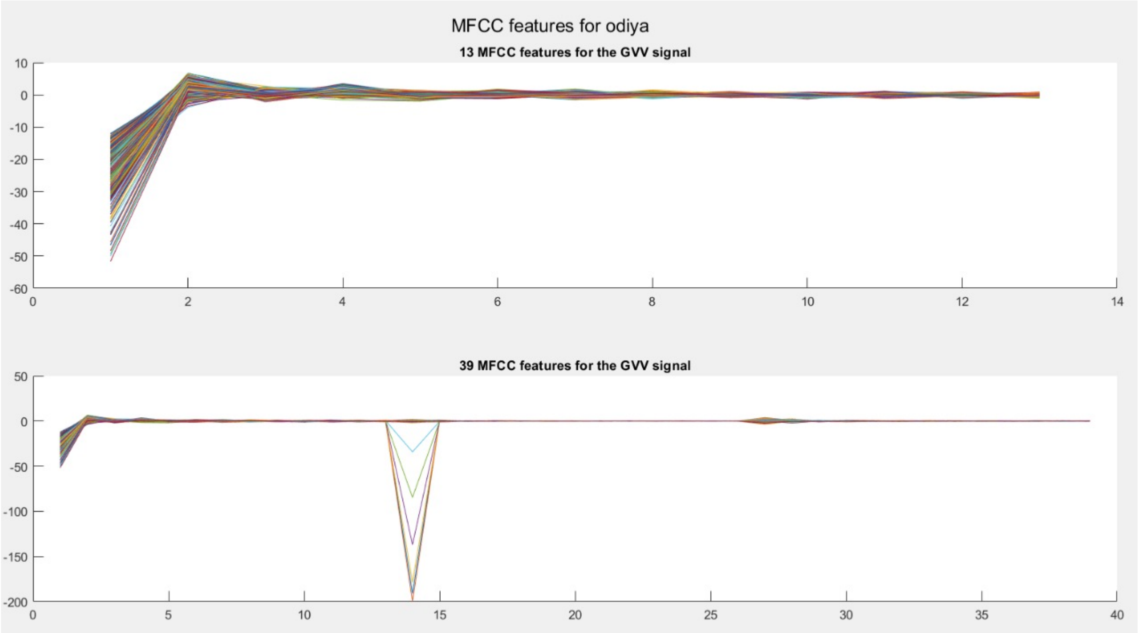


Figure 17: Odiya

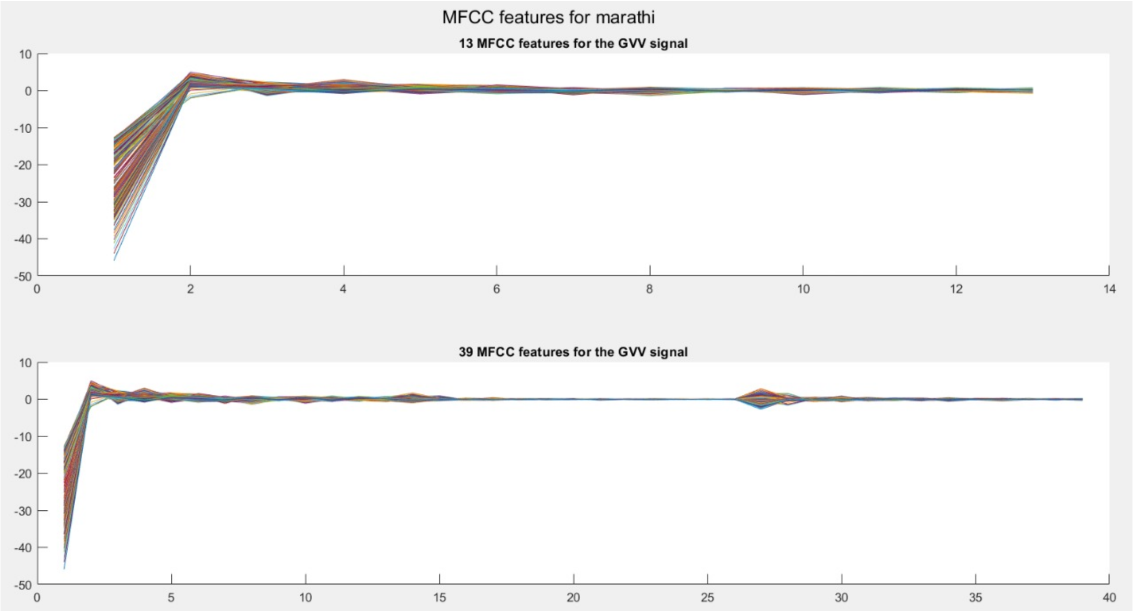


Figure 18: Marathi

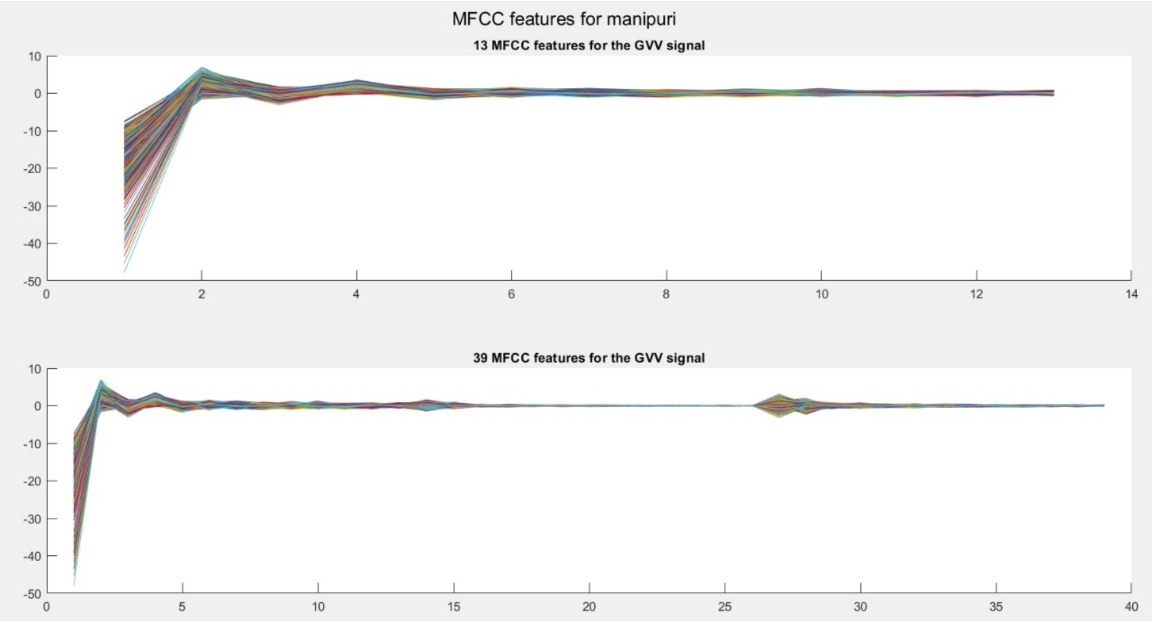


Figure 19: Manipuri

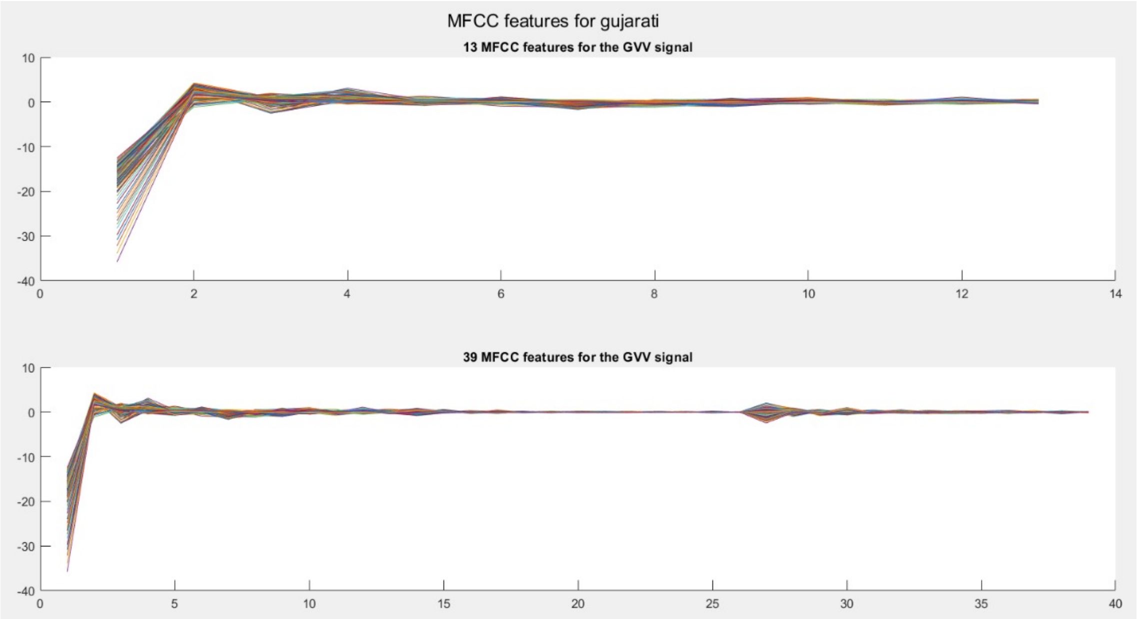


Figure 20: Gujarati

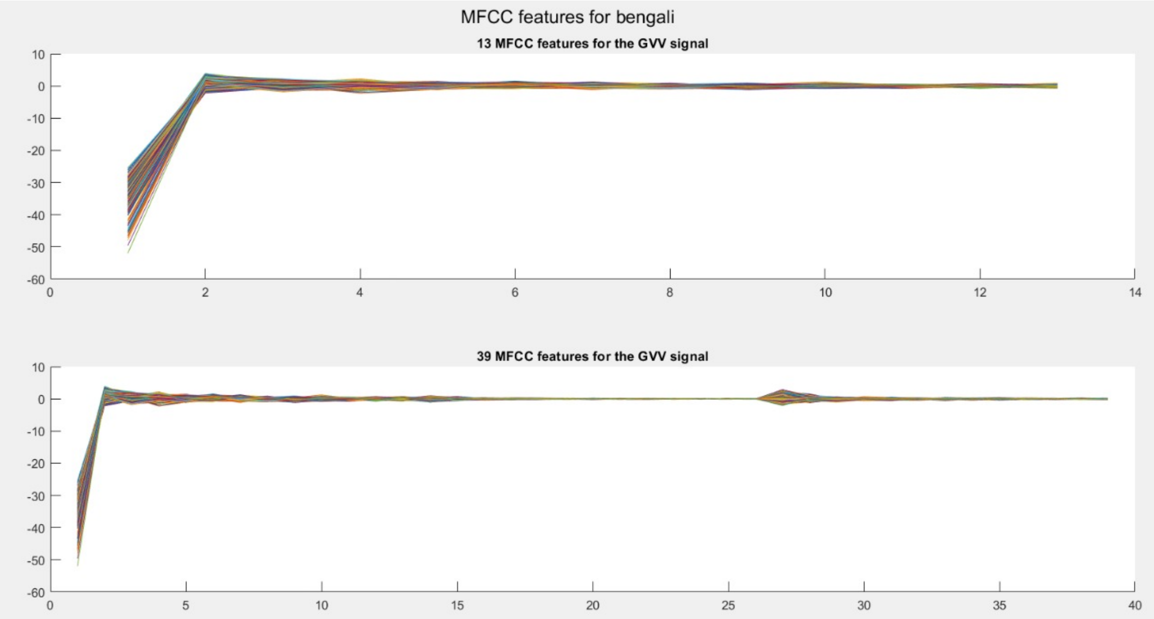


Figure 21: Bengali

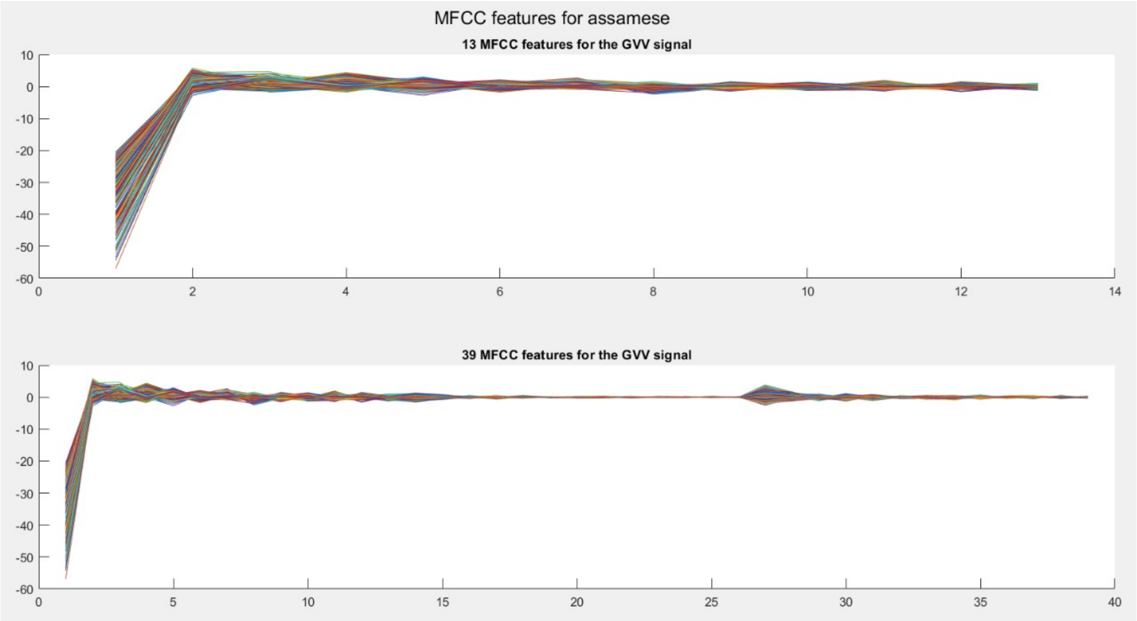


Figure 22: Assamese

Results :

Time Domain	Frequency Domain	Time + Frequency Domain	MFCC	MFCC + Time + Frequency Domain
24.6%	48.5%	64.3%	90.4%	96.2%

Observation:

We can primarily observe that MFCC gives a better accuracy when compared to our excitation features. The increase in number of clusters causes increases in accuracy but reaches eventually reaches saturation.

Accuracy magnitude ⟹
Time Domain < Frequency Domain < Frequency+ Time < MFCC < MFCC + Time+ Frequency.

References :

http://research.spa.aalto.fi/publications/theses/pulakka_mst.pdf
Language using excitation sources
Estimation of the glottal pulse from speech or singing voice