INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Speech Signal Processing

Combined CTSP Speech Enhancement for Language Identification in Noisy Environments

P15

Jewel Benny
Srujana Vanka
Shreeya Singh

Contents

Introduction

In this project, we have implemented a noisy speech enhancement method by spectral processing in the frequency domain to provide better noise suppression as well as better enhancement in the speech regions.
Spectral processing involves estimation and removal of degrading components, and also identification and enhancement of speech-specific spectral components. The spectral characteristics of the background noise is estimated and attenuated using conventional spectral processing methods based on spectral subtraction or MMSE estimators.
The spectrally processed speech is then subjected to LID using 7 GMM models (1 for each language). The Language Identification system is analysed and further modifications are made for feature extraction. The accuracy of the models is checked (with mix and match as well) and the results are observed.

Why Combine Spectral and Temporal Processing Techniques?

Integration of spectral and temporal processing

- The temporal processing approach **enhances the region around the instants of significant excitation** and the subsequent spectral processing **suppresses the noise spectral components.**
- To improve the vocal tract characteristics at the spectral level and to provide better noise suppression, the spectral processing is performed on the temporally processed speech that involve conventional spectral processing and proposed spectral enhancement techniques.
- Thus the integration of these two approaches may lead to better suppression of degradation and also enhancement of high SNR speech regions.
- This may lead to **improved performance** compared to either temporal processing or spectral processing alone.
- Further, from the speech production point of view, the temporal and spectral processing **methods use independent information from the noisy speech**.

Temporal Processing

Temporal Processing of Noisy Speech

- The noisy speech is initially processed by the excitation source (LP residual) based temporal processing.
- It involves **identifying and enhancing the excitation source based speech-specific features present at the gross and fine temporal levels.**
- The temporally processed speech is further subjected to spectral domain processing.

Spectral Processing

- The spectral processing is based on the fact that the spectral values of the degraded speech will have both speech and degrading components.
- **The spectral components of degradation are therefore estimated and removed.**
- Further, there are spectral peaks that are perceptually important that are identified and enhanced. Accordingly, spectral processing is performed using the following approach: **attenuation of spectral characteristics of background noise**.
- The spectral characteristics of the background noise is estimated and attenuated using conventional spectral processing methods based on **spectral subtraction or MMSE estimators.**

- Generally, in majority of the conventional spectral processing methods, both **short-term magnitude of degradation and degraded speech spectra are estimated first**.
- According to the suppression rule, a **spectral gain function** is applied to the magnitude spectra of the degraded speech to obtain enhanced speech spectra.
- The enhanced magnitude and degraded speech phase spectra are then combined to produce an estimate of clean speech.

Spectral Subtraction

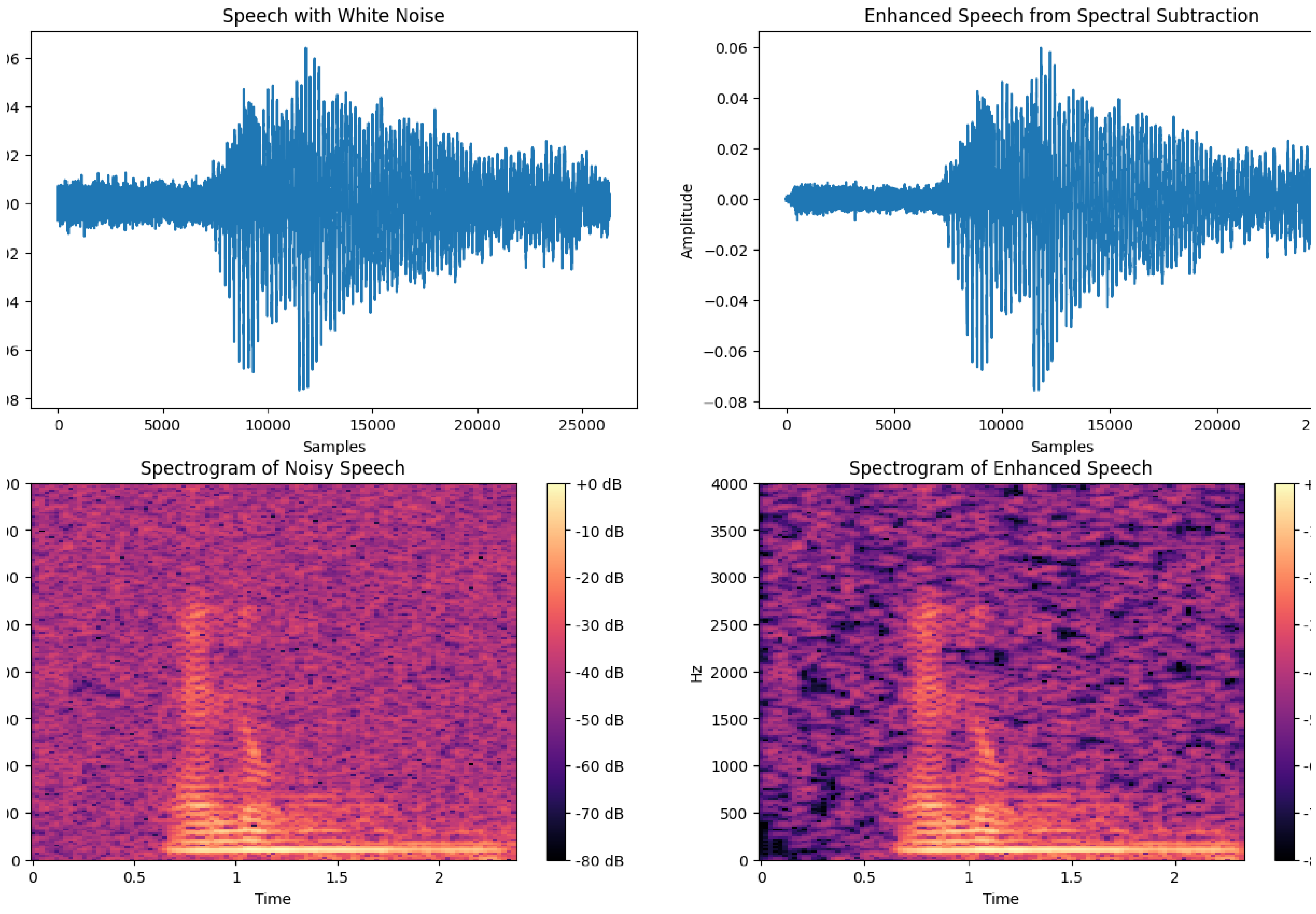
Assumptions

The assumption is that the noise is a **stationary or a slowly varying process,** and that the noise spectrum does not change significantly in-between the update periods.

Approach

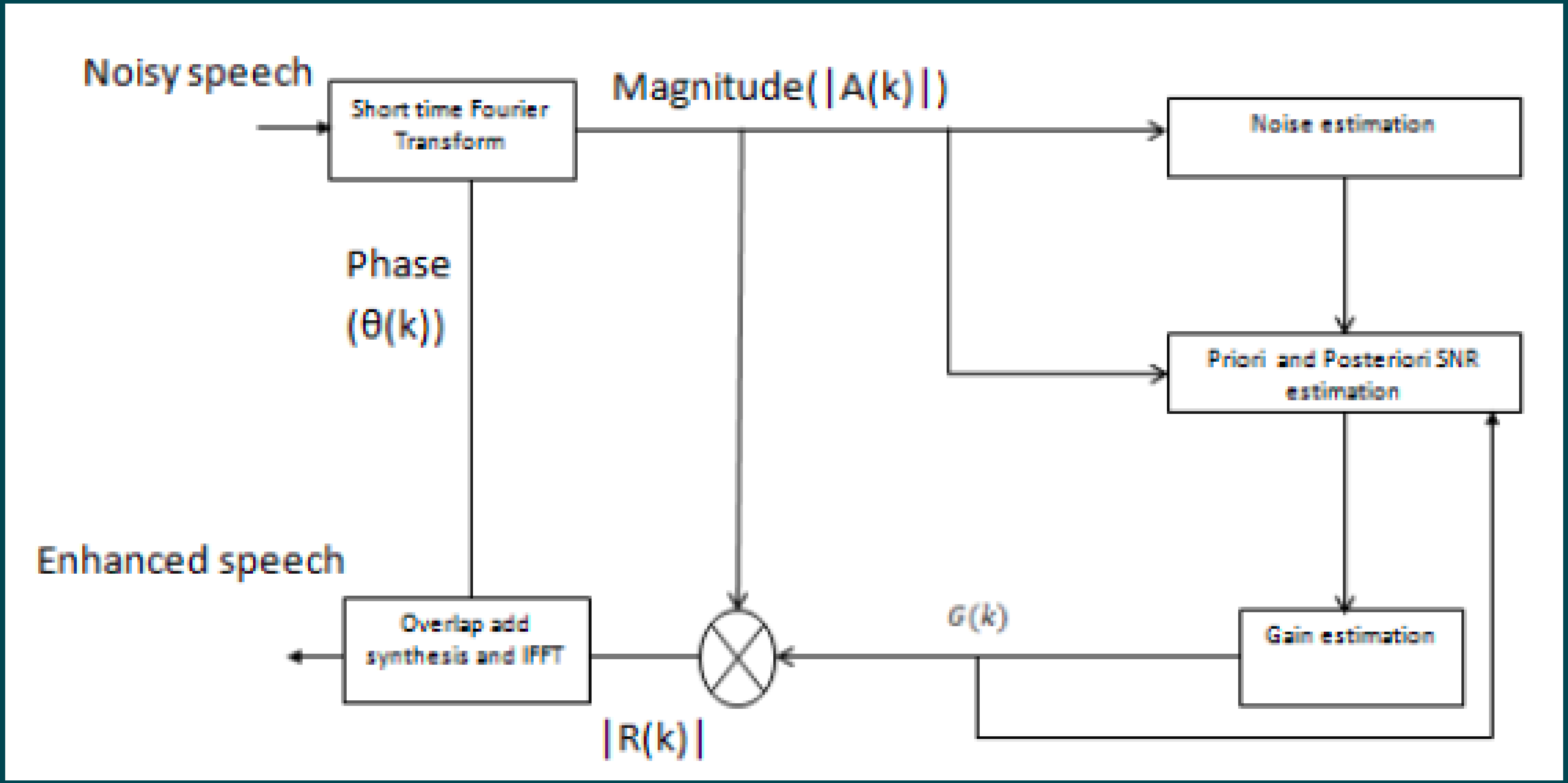
- Spectral subtraction **needs only noisy speech as input**. For this, an estimator is obtained by **subtracting an estimate of the noise spectrum from the noisy speech spectrum.** The signal collected during nonspeech activity provides the spectral information needed to define the noise spectrum.
- The enhanced signal is obtained by computing the **inverse discrete Fourier transform of the estimated signal spectrum** using the **phase of the noisy signal**. The algorithm is computationally simple as it only involves a forward and an inverse Fourier transform.
- The proposed spectral enhancement is performed only on the **high SNR regions** of the spectrally processed speech. This require an estimate of pitch information and is computed from the **autocorrelation of the HE of temporally processed LP residual**.

Spectral Subtraction Outputs



MMSE Estimator

Block diagram for STSA based statistical models



MMSE-STSA Principle

- It is statistical model that a distortion measure by mean square error of spectral amplitude of clean speech and estimate speech.

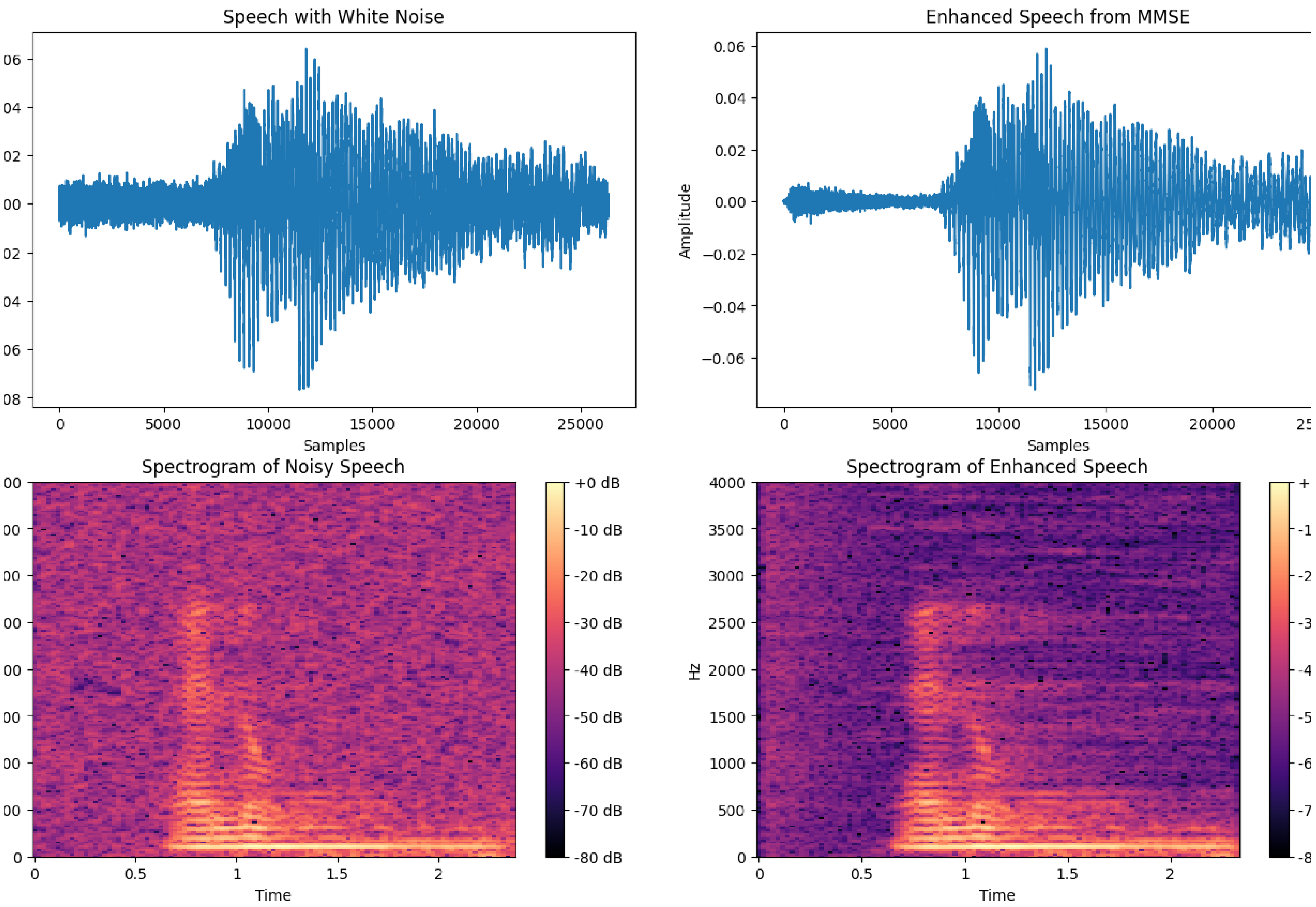
- In particular, optimal estimators were sought that minimized the mean-square error between the estimated and true magnitudes:

$$e = E\left\{\left(\hat{X}_k - X_k\right)^2\right\}$$

- The minimization of the equation can be done in two ways, depending on how we perform the expectation.
- Gain function of MMSE-spectral amplitude in terms of bessel function is given by the equation:

$$\hat{X}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_o\left(\frac{v_k}{2}\right)+v_k I_1\left(\frac{v_k}{2}\right)\right]Y_k$$

MMSE Estimator Outputs



LID Using GMM

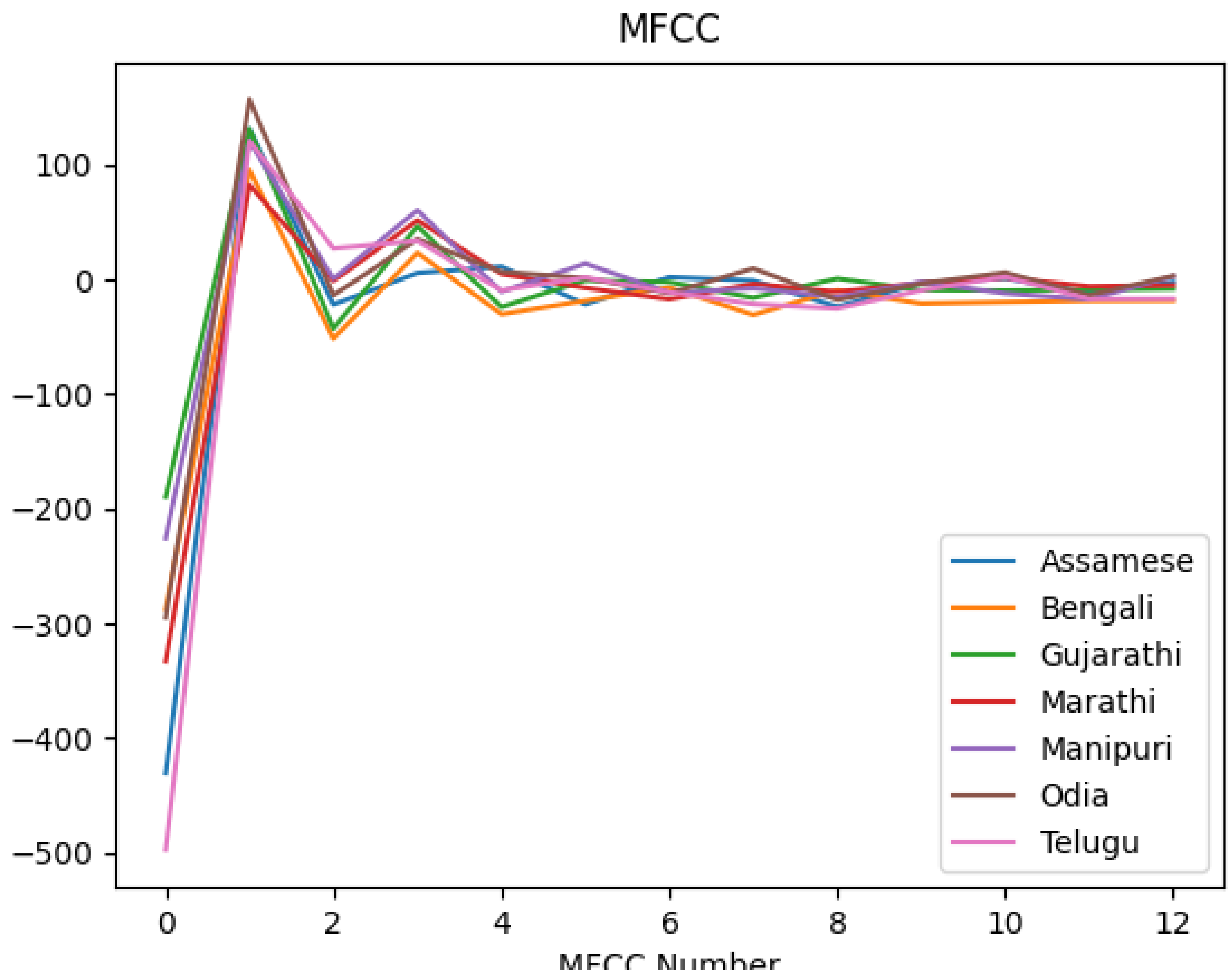
GMM Based Language Identification System

- This technique is generalized by using Gaussian mixture models as the basis for tokenizing.
- GMM based approach has been proposed for language recognition using new feature vectors derived from MFCC feature vectors and formants. Formants are extracted using LP spectrum of the speech signal.
- Formant and MFCC feature vectors represent the acoustic features of speech signals so that LID performance is improved.
- the GMM tokenizer is computationally less expensive

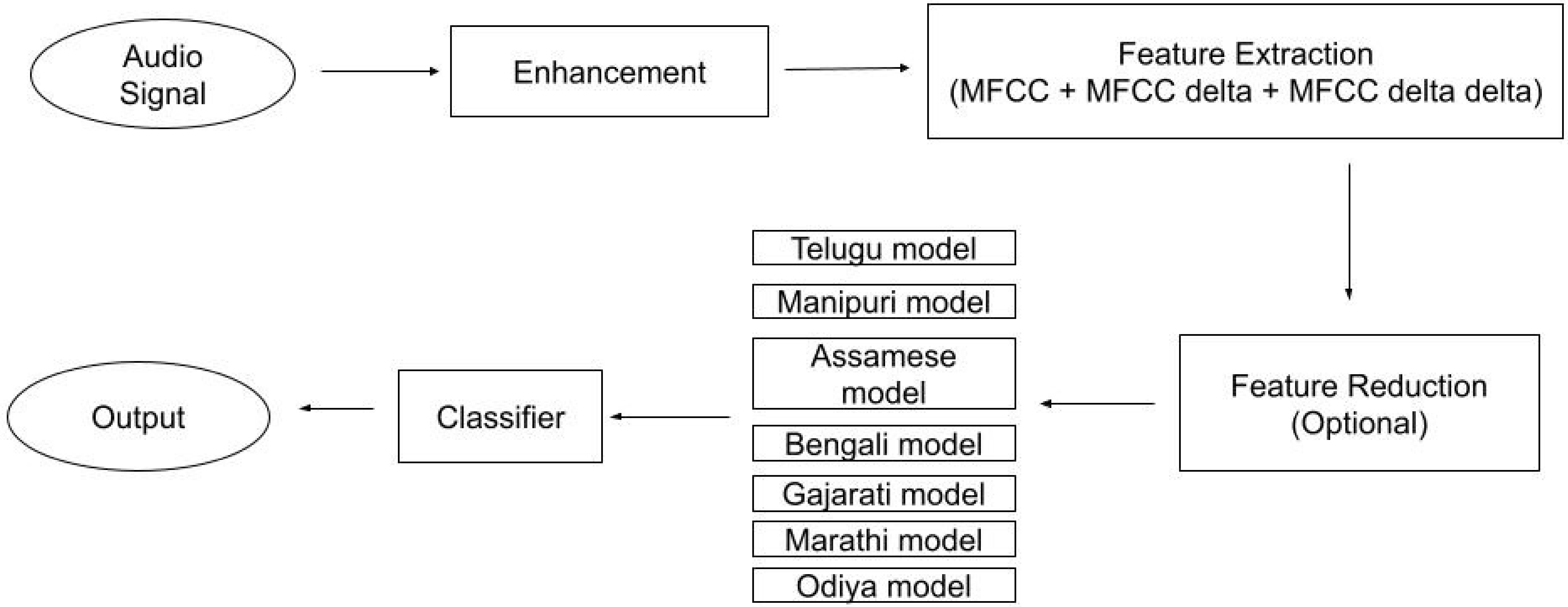
Mel-frequency cepstrum Coefficients

- The MFCC feature extraction technique basically includes:

1. windowing the signal
2. applying the DFT
3. taking the log of the magnitude
4. and then warping the frequencies on a Mel scale
5. followed by applying the inverse DCT.



LID Using GMM - Flowchart



Generated Datasets

- Clean Dataset - 80:20 Split
- Noisy Dataset (White) - 80:20 Split
- Enhanced Dataset (White) - 80:20 Split
- Noisy Dataset (Babble) - 80:20 Split
- Enhanced Dataset (Babble) - 80:20 Split
- Noisy Dataset (Factory) - 80:20 Split
- Enhanced Dataset (Factory) - 80:20 Split

ZFF and ZFCC

Zero Frequency Filtering Steps

- Difference the speech signal s[n] (to remove any timevarying low frequency bias in the signal)

$$x[n] = s[n] - s[n-1]$$

- Pass the differenced speech signal x[n] twice through an ideal resonator at zero frequency.

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n]$$

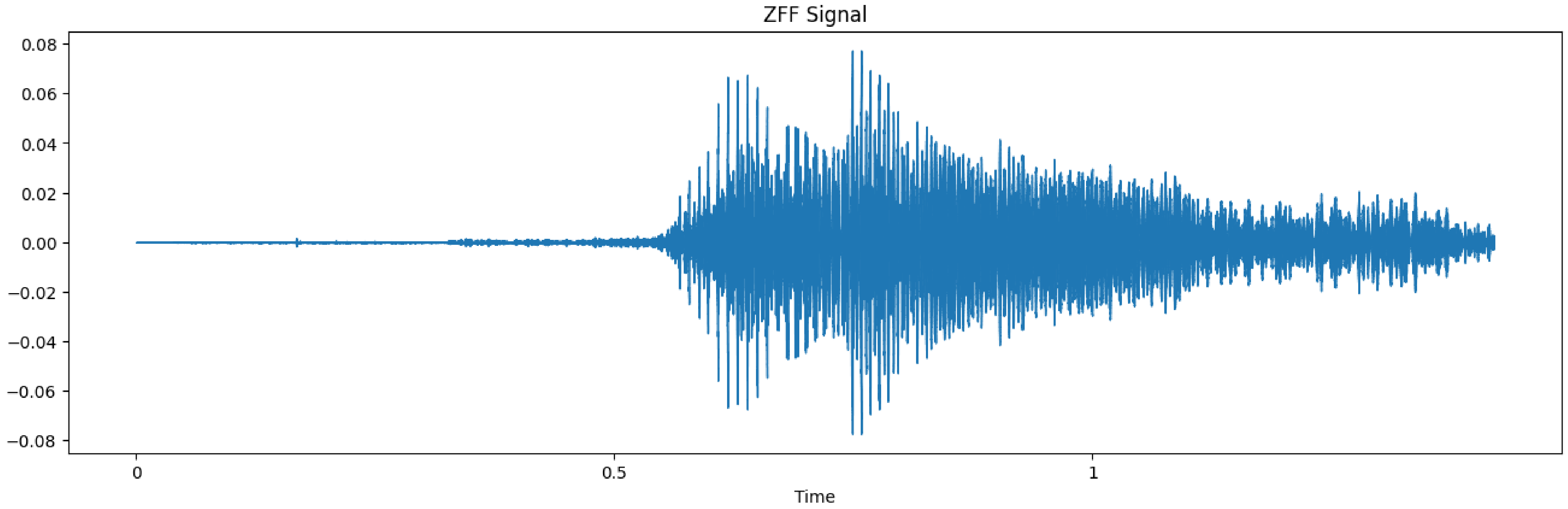
and

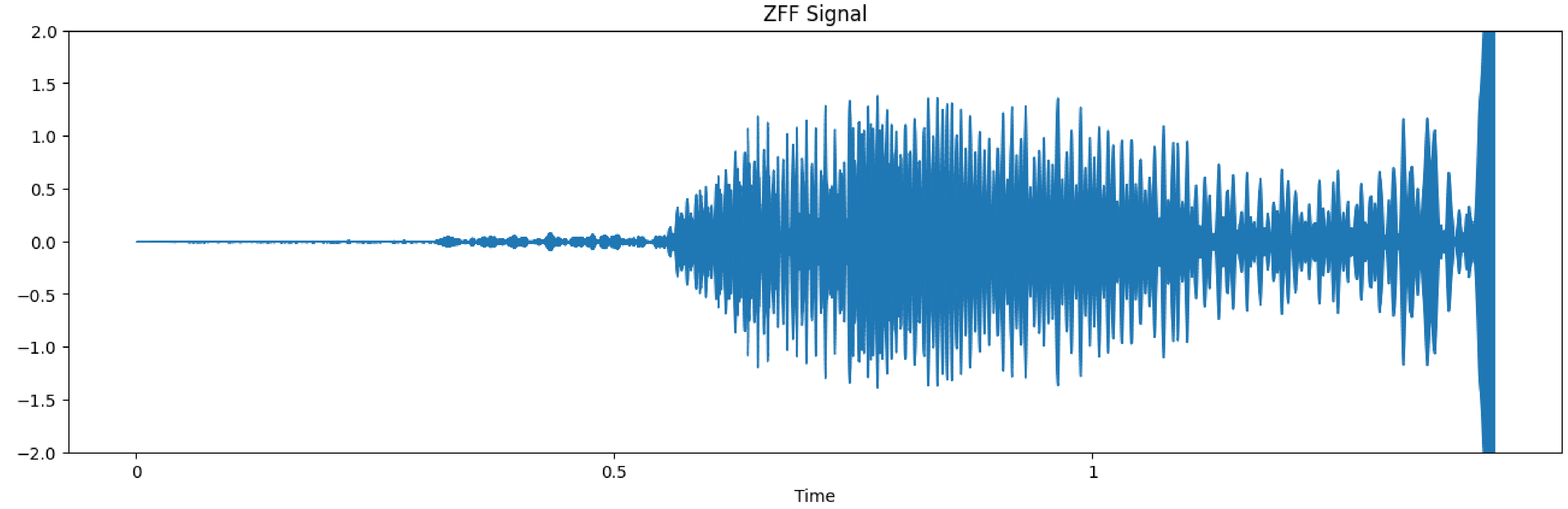
$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n]$$

- Remove the trend in y2[n] by subtracting the average over a window duration (say 10ms) at each sample.

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m]$$

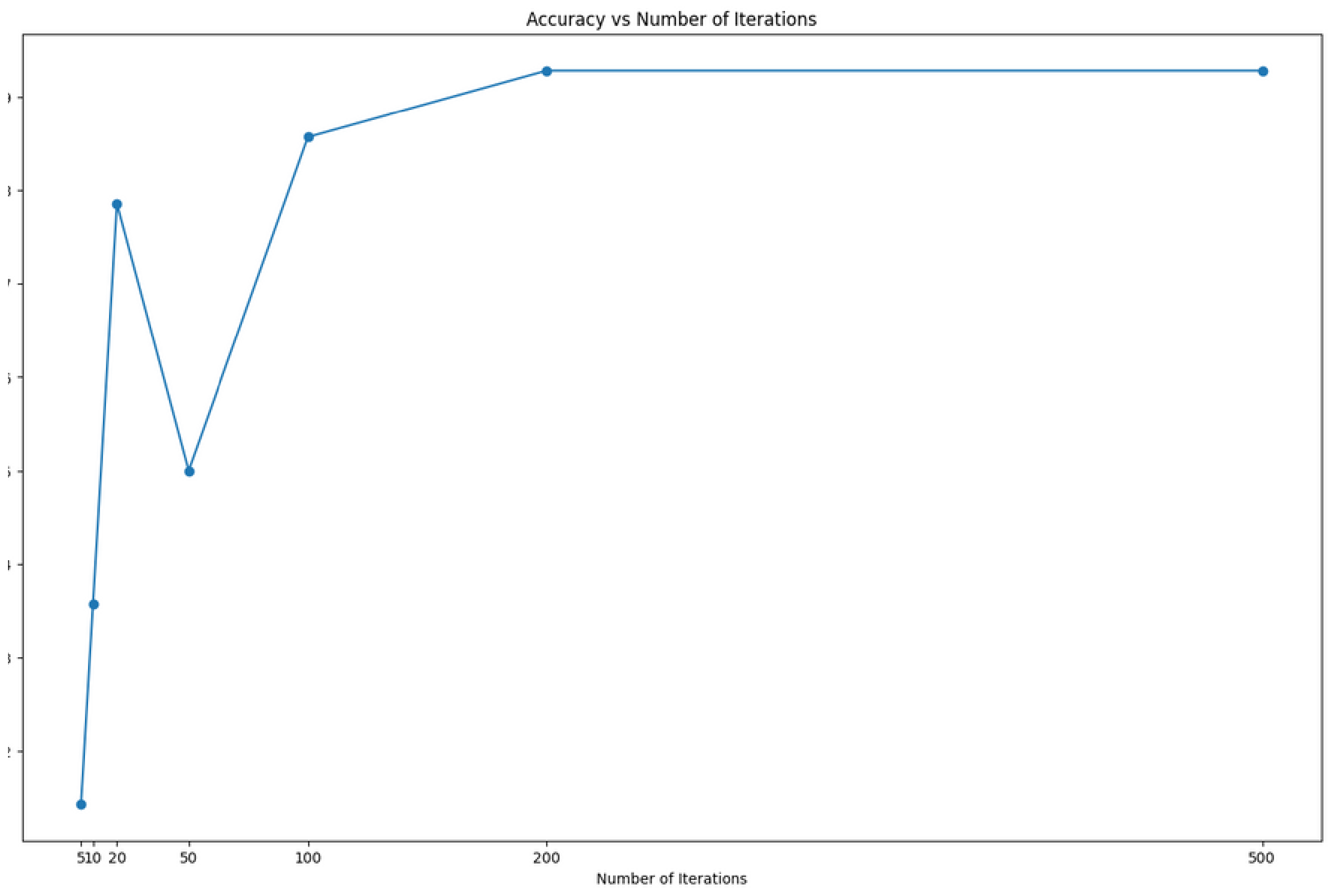
The MFCCs of ZFF signal are the ZFCCs.

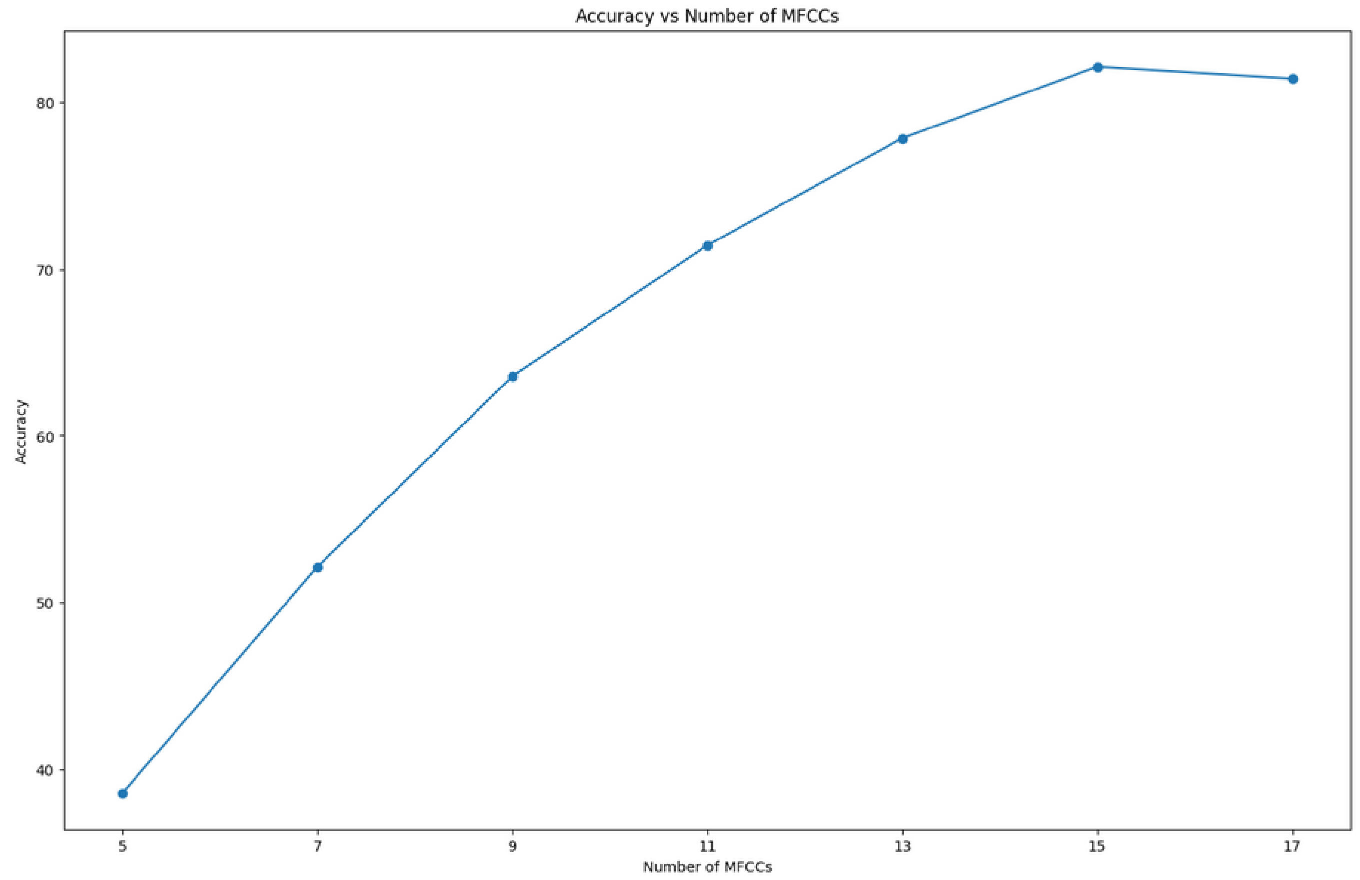


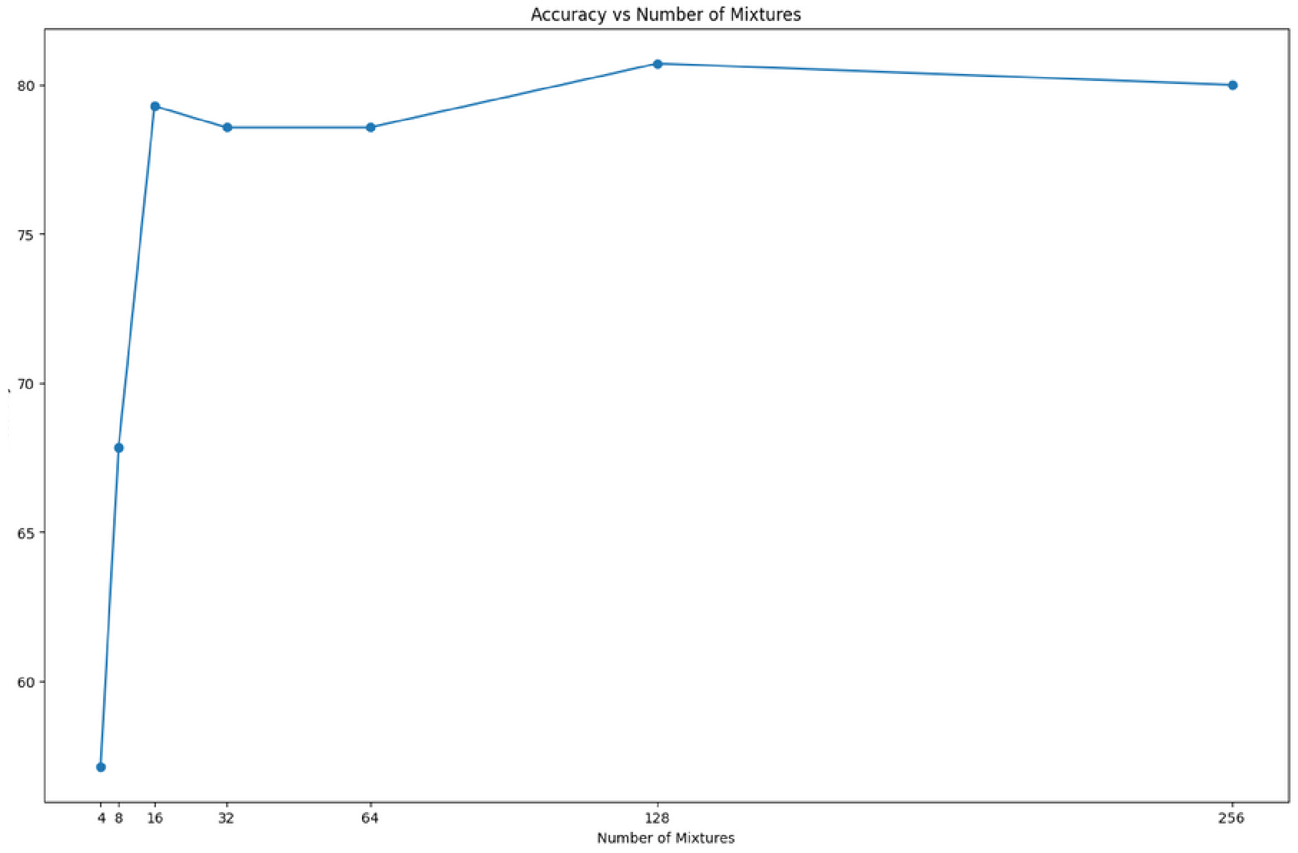


LID Analysis

LID Accuracy vs Various Paramaters



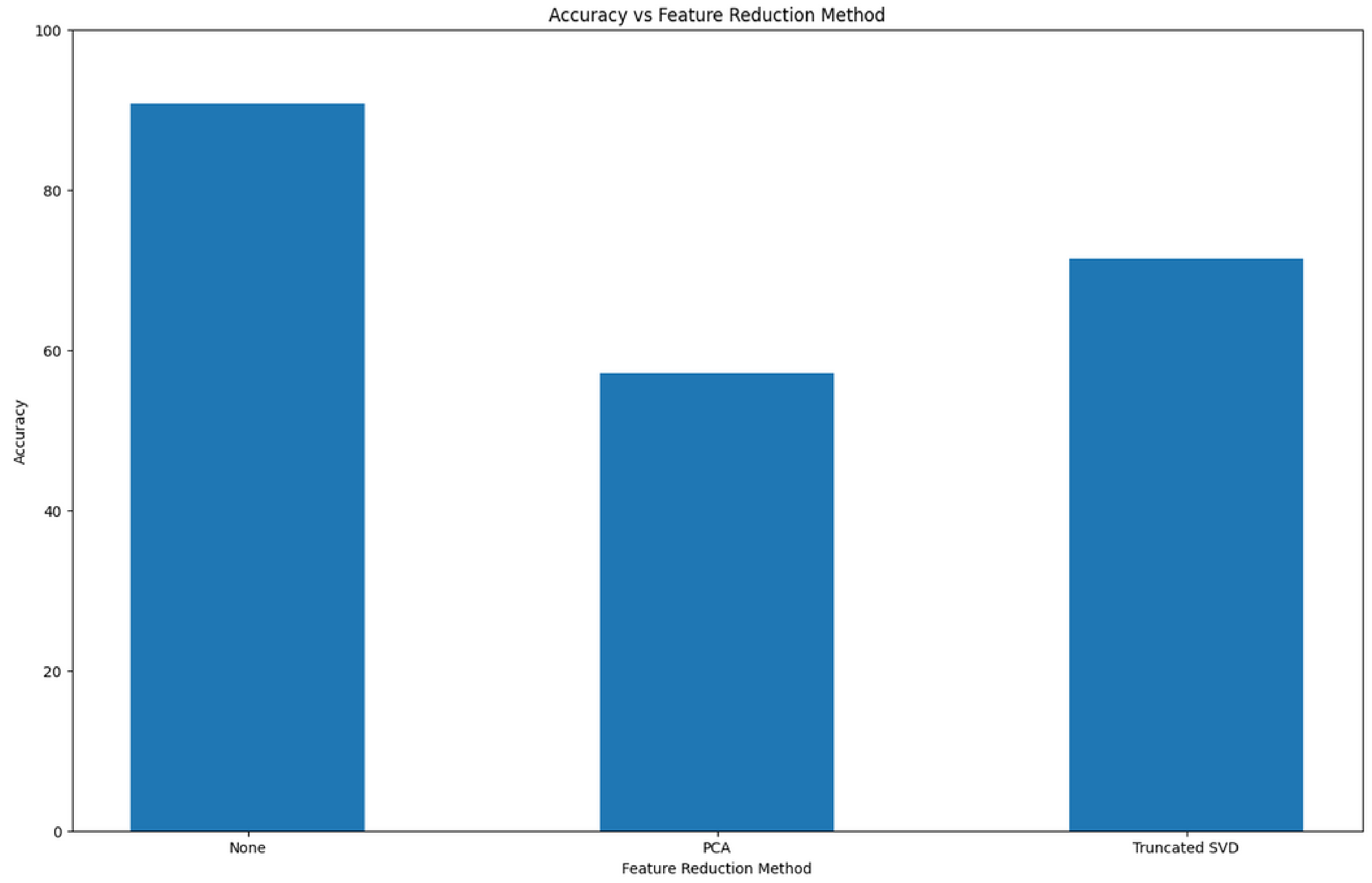


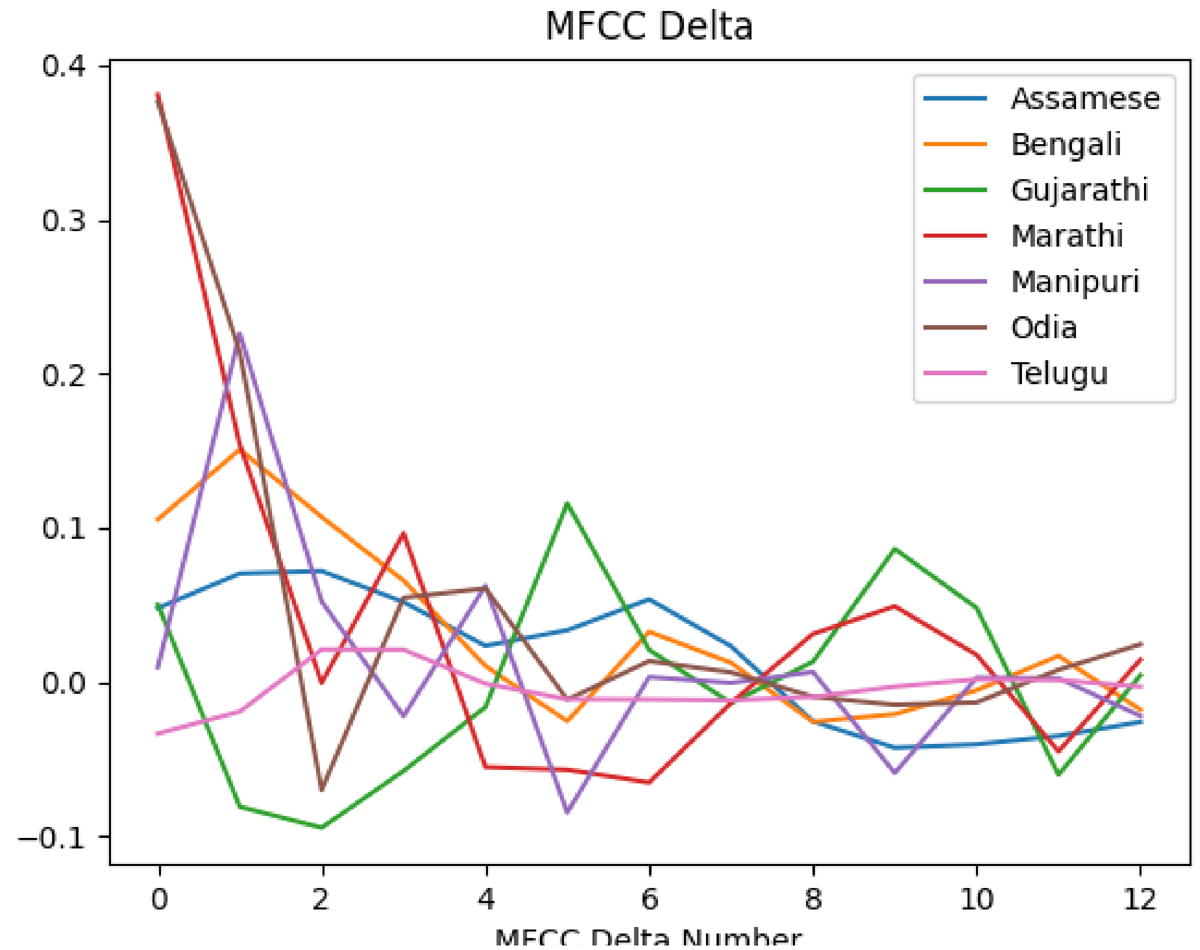


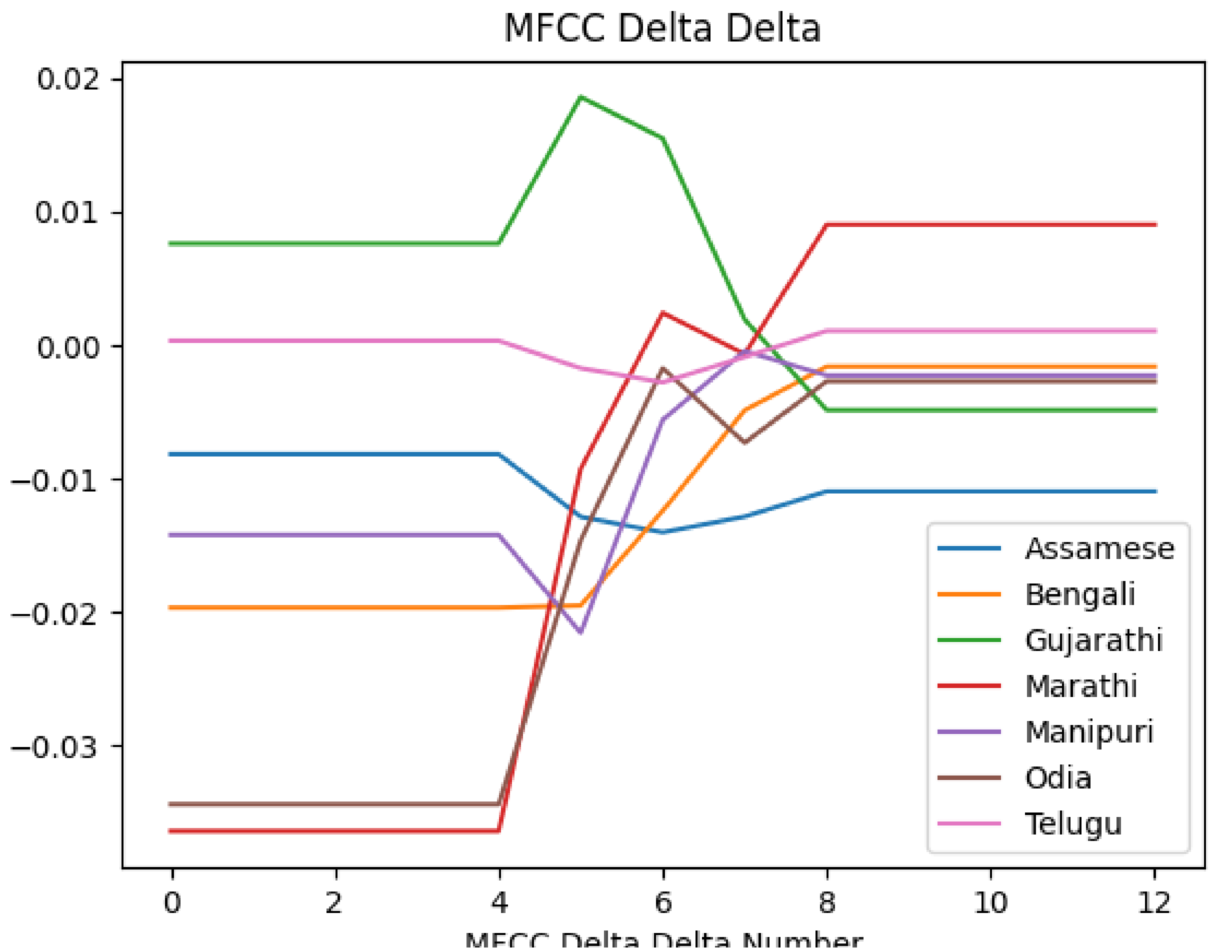
General Trend:
Prediction accuracy increases as the number of the parameter increases but saturates.

LID Modifications

MFCC Delta and MFCC Delta Delta Outputs + Feature Dimensionality Reduction + ZFCC







LID Results

Best Accuracy Obtained = 99.2857 %

	Telugu	Odia	Marathi	Assamese	Manipuri	Gujarathi	Bengali
Telugu	19	0	0	0	0	1	0
Odia	0	20	0	0	0	0	0
Marathi	0	0	20	0	0	0	0
Assamese	0	0	0	20	0	0	0
Manipuri	0	0	0	0	20	0	0
Gujarathi	0	0	0	0	0	20	0
Bengali	0	0	0	0	0	0	20

Number of Mixtures = 128
Number of Iterations = 100
Number of MFCCs = 13
Features = MFCC + MFCC Delta + MFCC Delta Delta

Accuracies		Training	
		Clean	Noisy
Testing	Clean	98.57	52.85
	Noisy	55.00	99.167

Testings Done Currently

Training Data	Testing Data	Accuracy (%)
Clean	Clean	99.28
Clean	Noisy (White)	55.00
Clean	Noisy (Babble)	41.42
Noisy (White)	Clean	52.85
Noisy (White)	Noisy (White)	98.57
Clean	Enhanced (Babble)	58.57
Noisy (Babble)	Enhanced (Babble)	55.71

Contributions

Jewel

- Implementation of GMM based Language Identification system
- LID Modifications (MFCC delta + MFCC delta delta + ZFCC)
- Implementation of MMSE Estimator
- Noise Addition
- Feature Redution

Srujana and Shreeya

- Implementation of spectral subtraction
- Comparison of spectral subtraction on different noises.
- Check the accuracy of models with mix and match (eg. noisy speech is input to clean speech GMM model)

Everyone

- README file
- Worked on the presentation
- LID Modifications (MFCC delta + MFCC delta delta + ZFCC)

References

[1] P. Krishnamoorthy, S.R.M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing" , Speech Communication, Volume 53, Issue 2, 2011, Pages 154-174, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2010.08.011. (https://www.sciencedirect.com/science/article/pii/S0167639310001457)

[2] P. A. Torres-Carrasquillo, D. A. Reynolds and J. R. Deller, "Language identification using Gaussian mixture model tokenization," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, pp. I-757-I-760, doi: 10.1109/ICASSP.2002.5743828.

[3] Potla, Sreedhar. “Spoken Language Identification using Gaussian Mixture Model-Universal Background Model in Indian Context.” (2018)

[4] Steven F. Boll “Suppression of Acoustic Noise in Speech Using Spectral Subtraction” IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-27, NO. 2, APRIL 1979

[5] Yariv Ephraim “Speech Enhancement Using a- Minimum MeanSquare Error Short-Time Spectral Amplitude Estimator” IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-32, NO. 6, DECEMBER 1984

[6] Philipos C. Loizou "Speech Enhancement: Theory and Practice, 2nd Edition"