Combined CTSP Speech Enhancement for Language Identification in Noisy Environments

Speech Signal Processing

P15

Jewel Benny - 2020102057

Srujana Vanka - 2020102005

Shreeya Singh - 2020102011

Abstract

Acknowledgement

Objective

- To 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
- To test the enhanced speech using a GMM based Language Identification system and compare the outputs.
- To test with several LID modifications (MFCC + MFCC delta + MFCC delta delta + ZFCC) and check the accuracy of the models with mix and match.

Theory

1. Spectral Processing of Noisy Speech

- 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 in two stages: attenuation of spectral characteristics of background noise and enhancement of speech-specific spectral features.
- In the first stage, the spectral characteristics of the background noise is estimated and attenuated using conventional spectral processing methods based on spectral subtraction or MMSE estimators.

2. Spectral Subtraction

- The spectral subtraction algorithm is historically one of the first algorithms proposed for noise reduction, and is perhaps one of the most popular algorithms.

- It is based on a simple principle. Assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum.
- The noise spectrum can be estimated, and updated, during periods when the signal is absent.
- 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 single forward and inverse Fourier transform.

Basic spectral subtraction:

- Assume that y(n), the noise-corrupted input signal, is composed of the clean speech signal x(n) and the additive noise signal, d(n), that is,

$$y(n) = x(n) + d(n)$$

- Taking the discrete-time Fourier transform, Y(w) can be expressed in polar form as:

$$Y(\omega) = |Y(\omega)| e^{j\phi_y(\omega)}$$

- The magnitude noise spectrum |D(ω)| is unknown, but can be replaced by its average value computed during nonspeech activity (e.g., during speech pauses). Similarly, the noise phase ϕd(ω) can be replaced by the noisy speech phase ϕy(ω). This is partly motivated by the fact that phase that does not affect speech intelligibility may affect speech quality to some degree.
  Thus we get the following equation:

$$\hat{X}(\omega) = [|Y(\omega)| - |\hat{D}(\omega)|]e^{j\phi_y(\omega)}$$

- We use the symbol "∧" to indicate the estimated spectrum or estimated parameter of interest.
- The enhanced speech signal can be obtained by simply taking the inverse Fourier transform of X^(ω).
- Spectral subtraction could in principle be performed in the autocorrelation domain as well as the cepstrum domain.

Demerits of Spectral Subtraction

- The subtraction process needs to be done carefully to avoid any speech distortion. If too much is subtracted, then some speech information might be removed, whereas if too little is subtracted, then much of the interfering noise remains.

- One of the serious drawbacks of this method is that it produces musical noise in the enhanced speech. This noise arises because of randomly spaced peaks in the time frequency plane due to the deviation of the estimated spectrum of noise from the instantaneous noise spectrum.
- Another relatively minor shortcoming of the spectral subtraction approach is the use of noisy phase that produces a roughness in the quality of the synthesized speech. The phases of the noise-corrupted signal are not enhanced before being combined with the modified spectrum to regenerate the enhanced time signal. This is because the presence of noise in the phase information does not contribute much to the degradation of speech quality. Although this is especially true at high SNRs (>5dB) and at low SNRs (<0dB), the noisy phase can lead to a perceivable roughness in the speech signal, contributing to the reduction in speech quality.

3. MMSE Estimator

- MMSE-Spectral amplitude is one of best estimation method for estimating clean speech amplitude from corrupted speech signal amplitude. It is a statistical model of 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:

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

4. Zero Frequency Filtering

- Zero frequency filtering is a simple and effective technique **used to estimate the glottal closure instants accurately from the speech signal**.

- The basic assumption that is followed in this method is that the excitation to the vocal-tract system can be approximated by a sequence of impulses of completely varying strengths.
- The effect due to an impulse signal in the time-domain is spread uniformly across the spectrum in frequency-domain including at zero-frequency.
- The main feature of a Zero Frequency Resonator is to extract the excitation source characteristics from speech signals by the process of filtering out most of the time-varying vocal-tract information.
- The following steps are involved in processing the speech signal to derive the zero frequency filtered signal.

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

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

2) Pass the differenced speech signal *x[n]* twice through an ideal resonator at zero frequency. That is

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

and

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

Here, $a_1 = -2$ and $a_2 = 1$. This is equivalent to successive integration four times but we prefer to call the process as filtering at zero frequency.

3) Remove the trend in $y_2[n]$ by subtracting the average over 10 ms at each sample.

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

The resulting signal is called the zero-frequency filtered signal, or simply the filtered signal. Here, 2N+1 is the number of samples in the interval (say 10ms).

Procedure

1. Spectral Subtraction

Assumption:

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.

2. MMSE Estimator

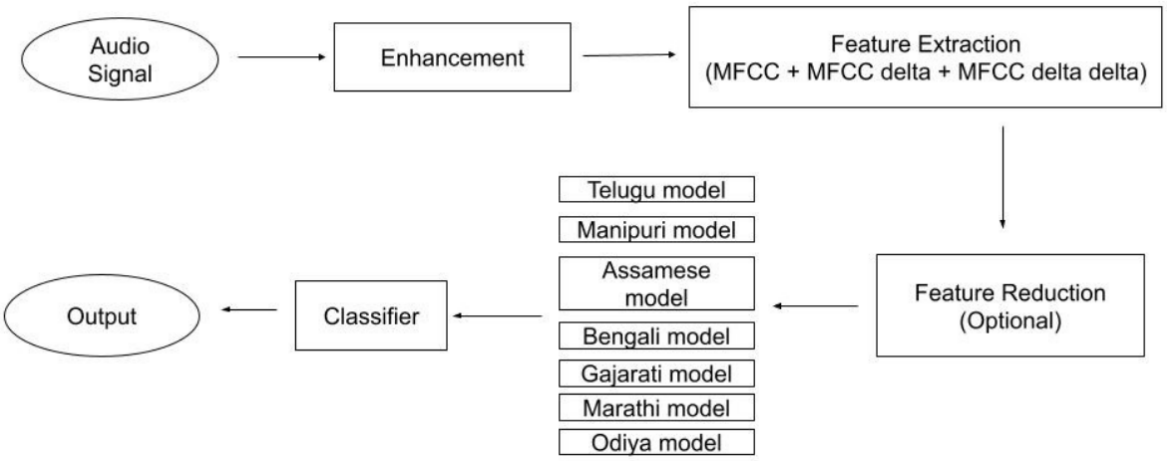
3. GMM Based Language Identification System

- 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.

MFCC Feature Extraction

- The MFCC feature extraction technique basically includes:
    1. Windowing the signal
    2. Applying the DFT
    3. Taking the log of the magnitude
    4. Warping the frequencies on a Mel scale
    5. Followed by applying the inverse DCT
    6. MFCC deltas and MFCC delta deltas can also be obtained from MFCCs.

LID using GMM flowchart

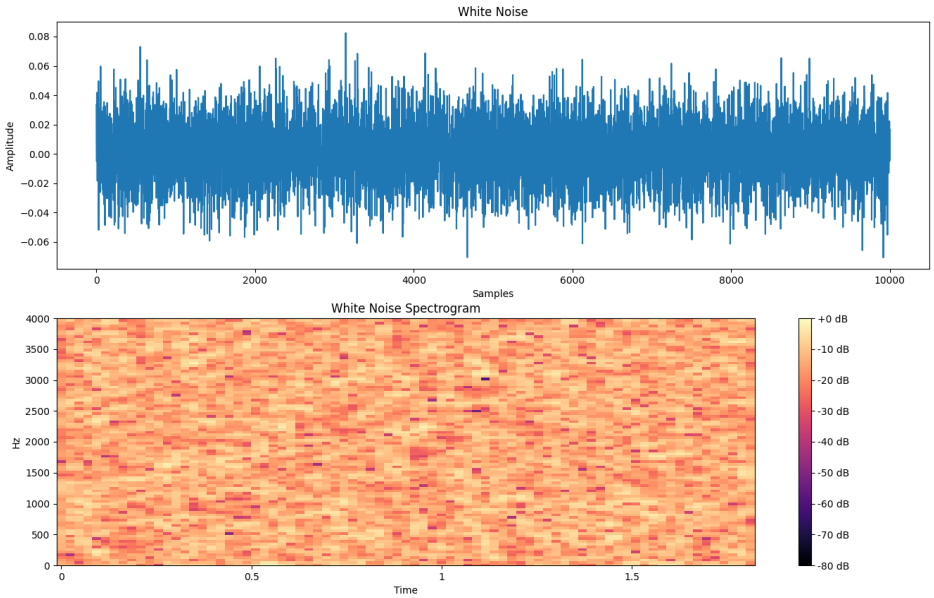


In addition to the MFCC feature vectors, we also try to use MFCCs of the ZFF signal (ZFCCs) also as feature vectors.

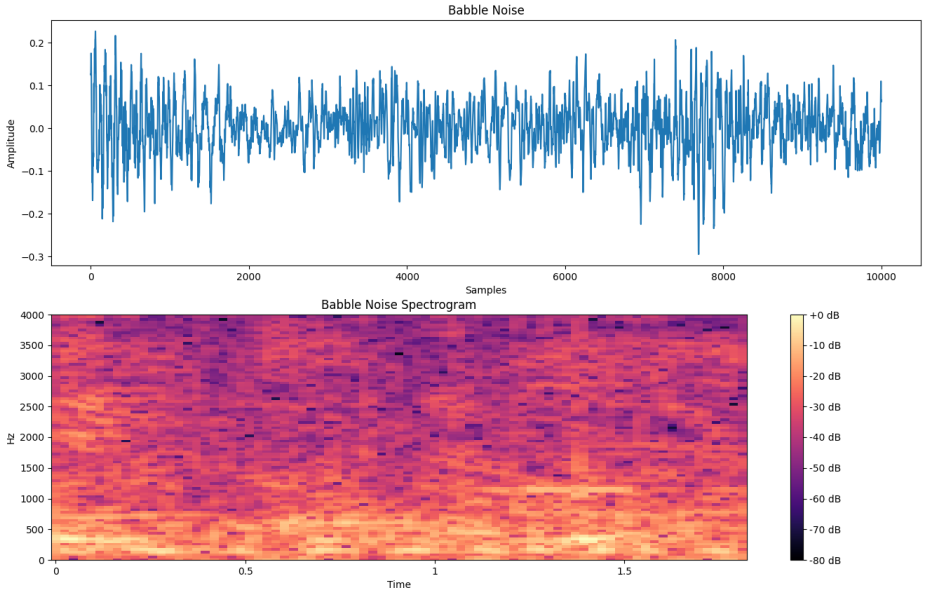
Analysis

1. Noise Spectrograms

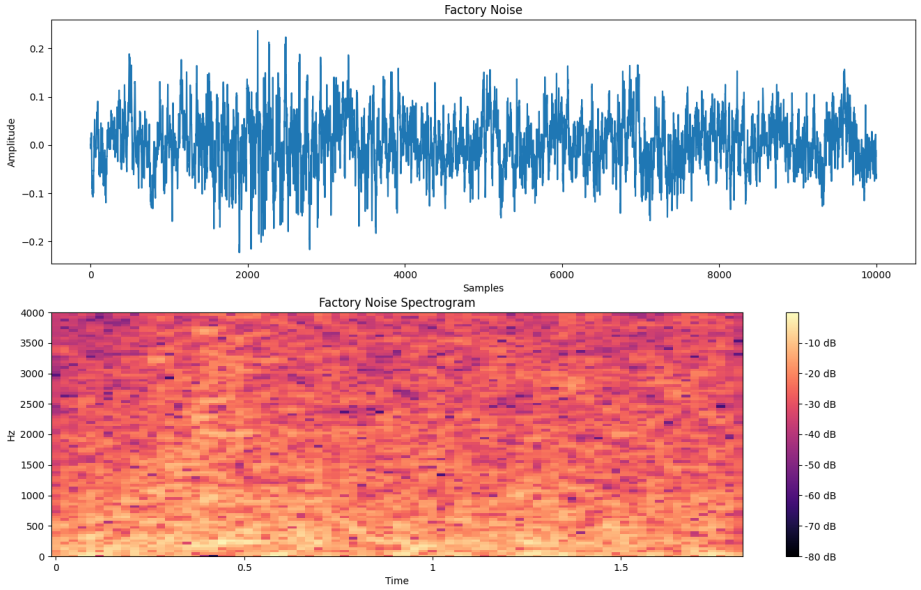
White Noise



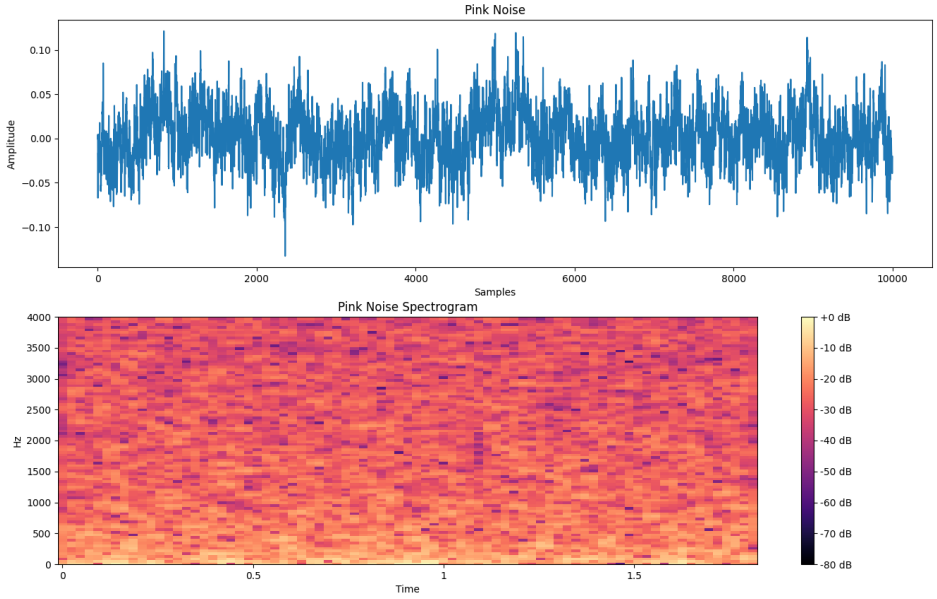
Babble Noise



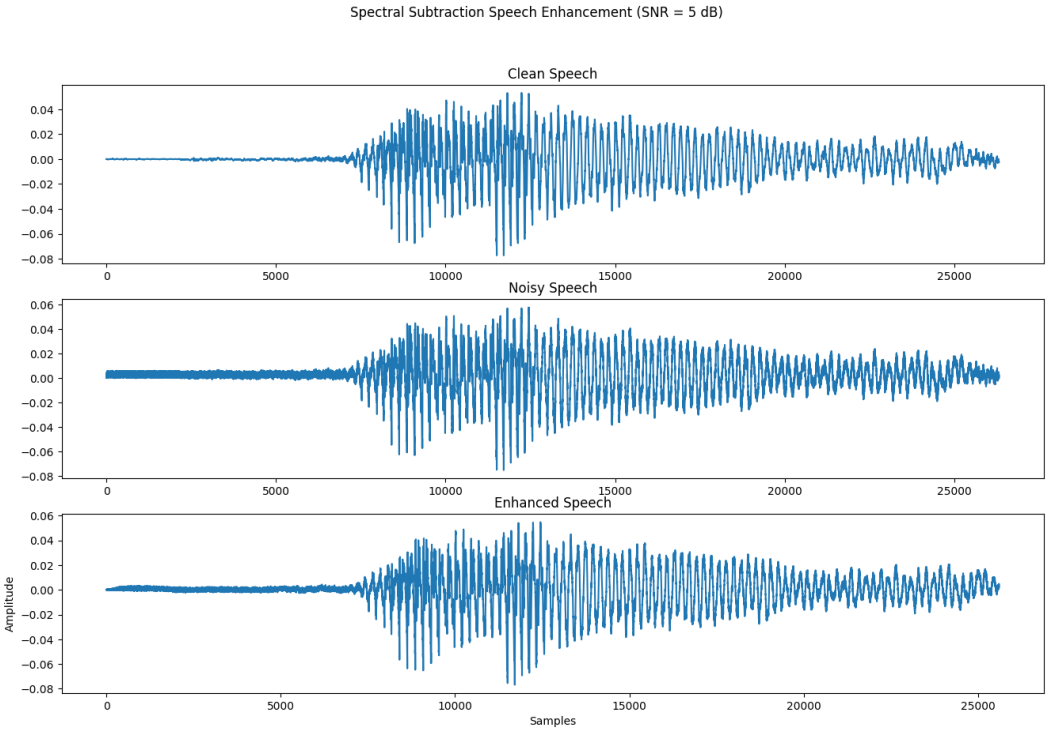
Factory Noise



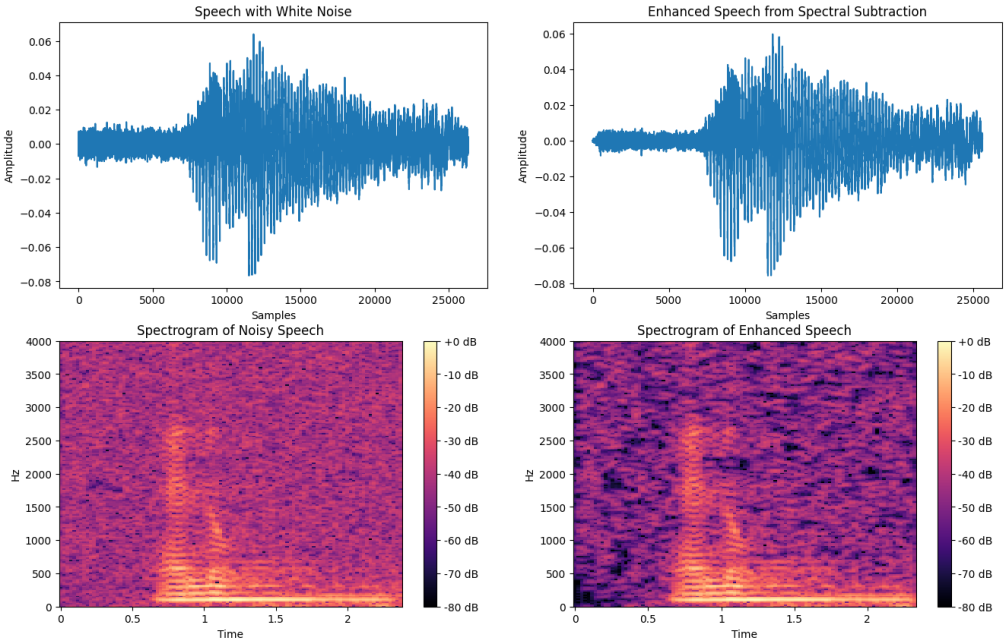
Pink Noise



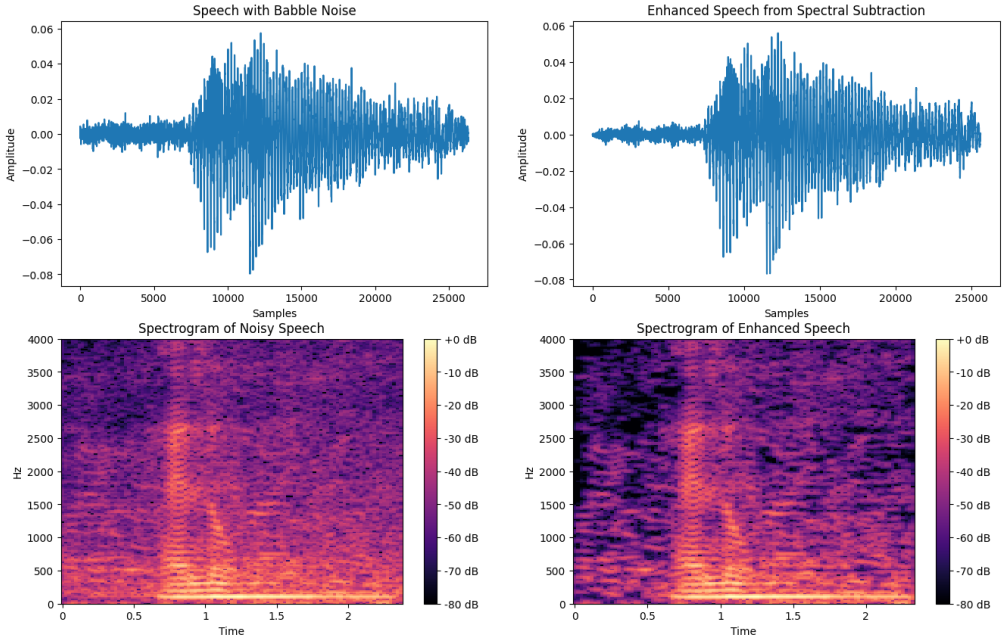
2. Spectral Subtraction Outputs



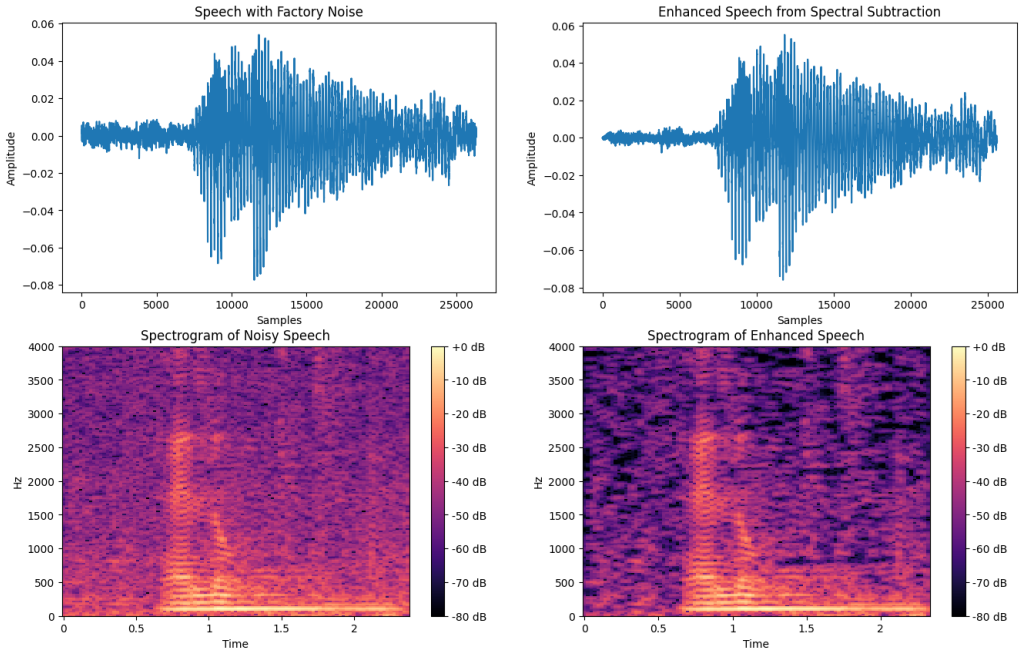
White Noise



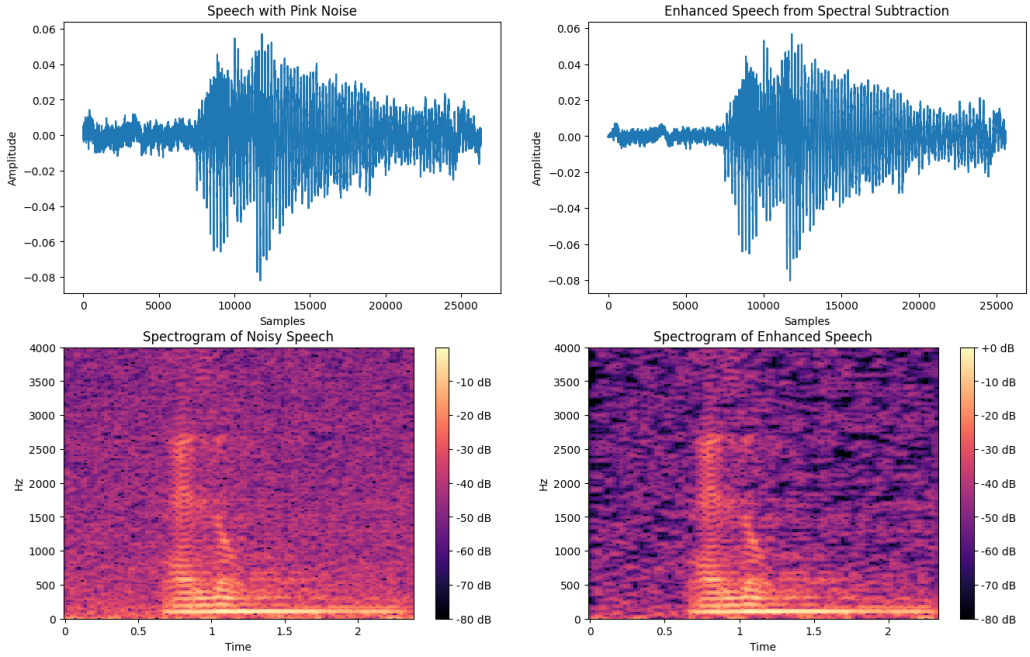
Babble Noise



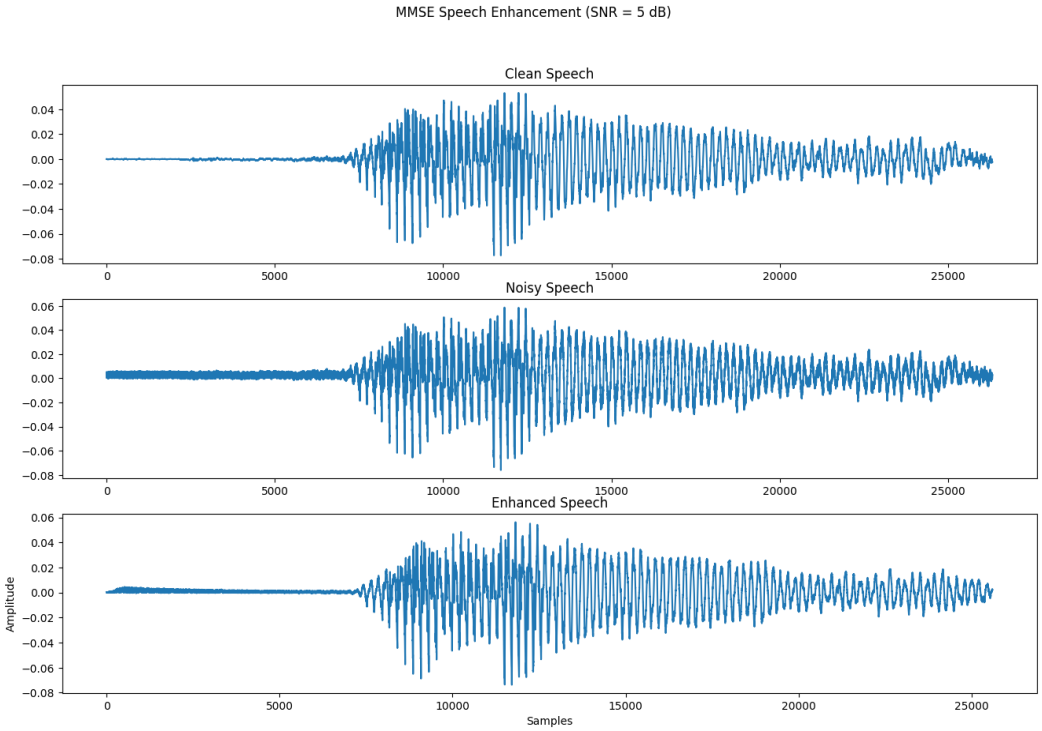
Factory Noise



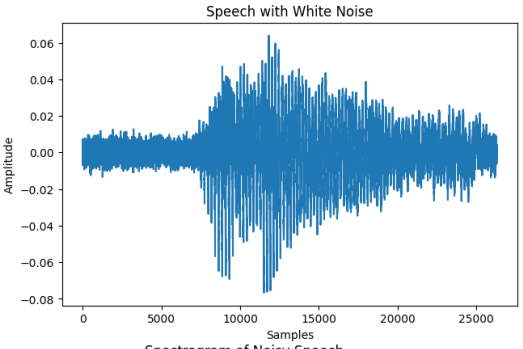
Pink Noise

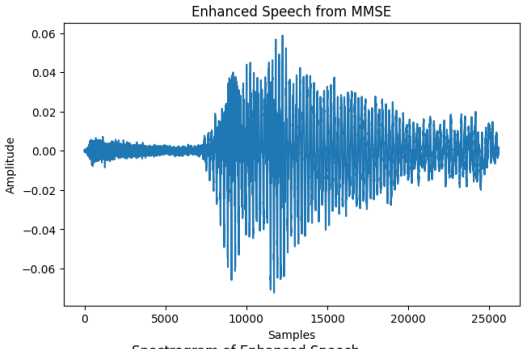


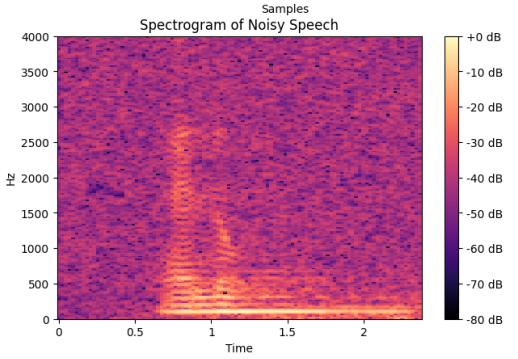
3. MMSE Estimator Outputs

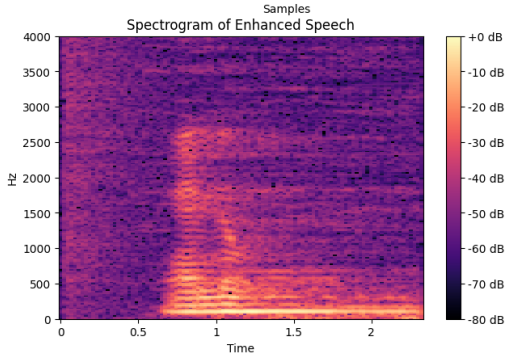


White Noise

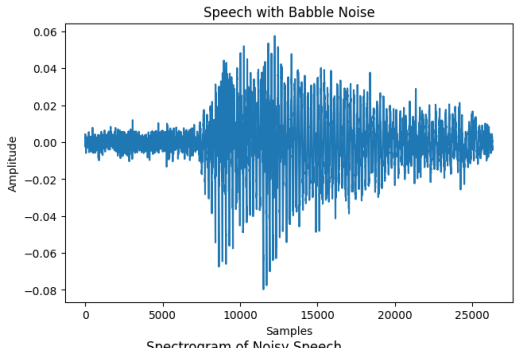


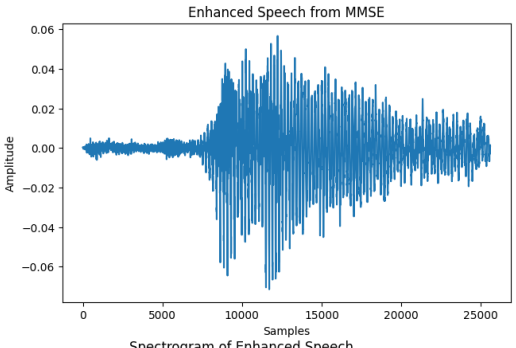


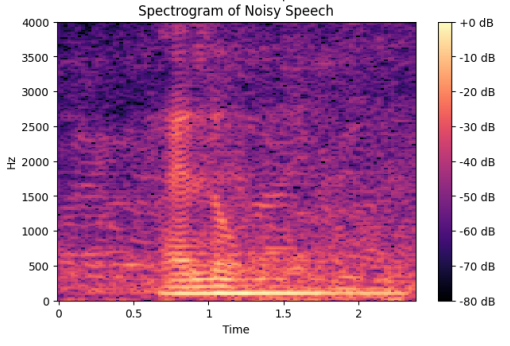


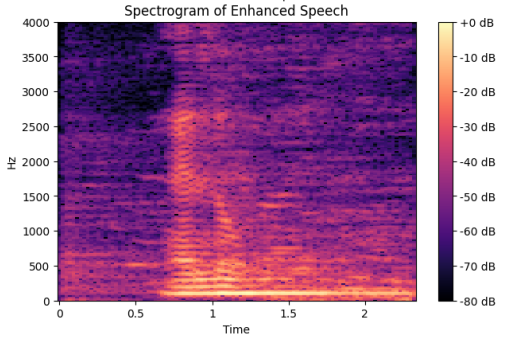


Babble Noise

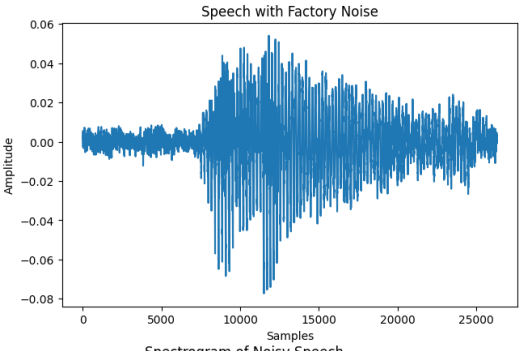


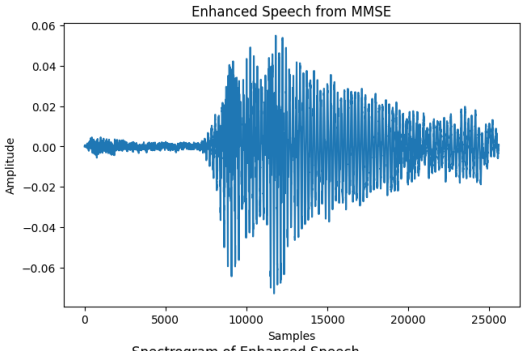


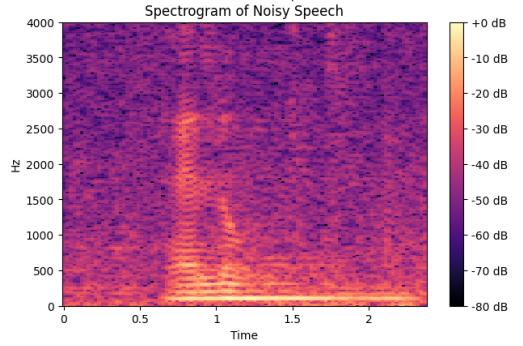


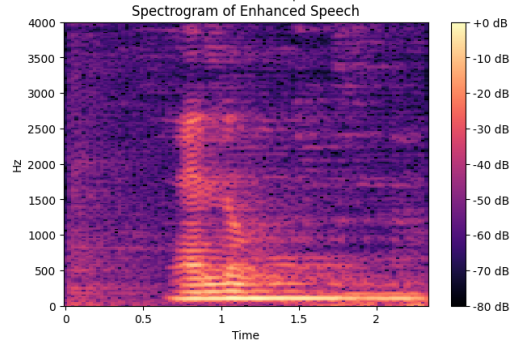


Factory Noise

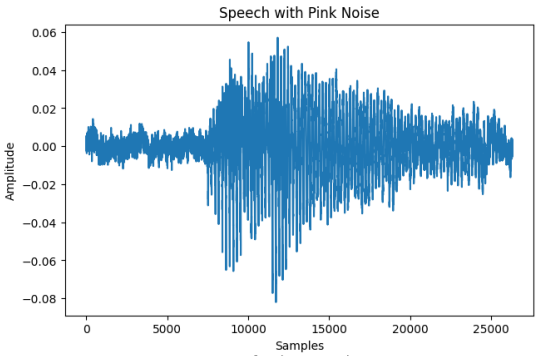


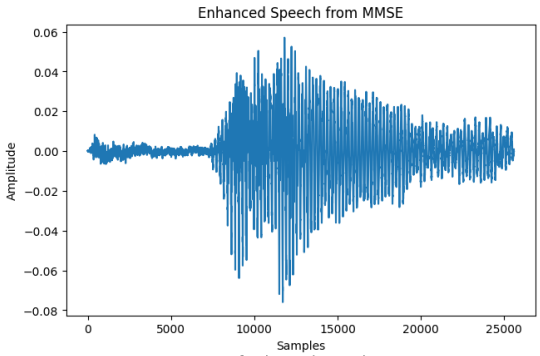


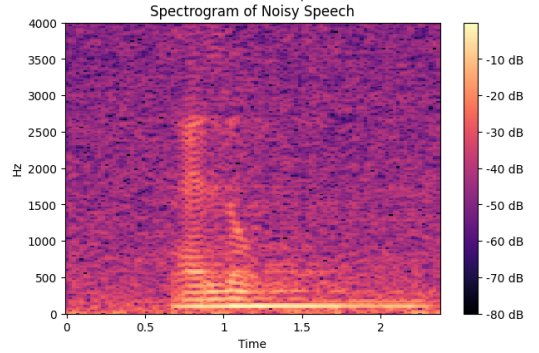


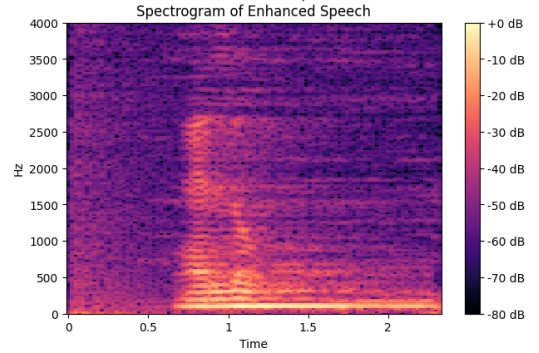


Pink Noise









4. **LID Analysis**
5. **LID Modifications**

Results