**SINGLE-CHANNEL SPEECH ENHANCEMENT**

Table of Contents

[1. Introduction 4](#_Toc422831505)

[1.1. Purpose 4](#_Toc422831506)

[1.2. Scope 4](#_Toc422831507)

[1.3. Reference Documents 4](#_Toc422831508)

[1.4. Assumptions, Constraints & Risk](#_Toc422831509) 5

[1.5. Overview](#_Toc422831510) 6

[2. Description 7](#_Toc422831511)

[2.1. Data Specifications 7](#_Toc422831512)

[2.2. Basic Concept 7](#_Toc422831513)

[2.3. Basic and Alternate Flows 1](#_Toc422831514)1

[3. Error Analysis 1](#_Toc422831515)3

[4. Validation 1](#_Toc422831516)4

[4.1. Approaches 1](#_Toc422831517)4

[4.2. Measurements 1](#_Toc422831518)6

[5. Abbreviations 1](#_Toc422831519)8

# Introduction

## Purpose

Speech signal is often accompanied by the background noise of the environment. This algorithm uses a modified spectral over-subtraction approach that allows better and more suppression of the noise than the one obtained in the classical spectral subtraction method.

## Scope

The spectral subtraction is a well-known single channel noise reduction technique. The implementation of this basic technique applies subtraction of the noise spectrum estimate over the speech spectrum. The conventional power spectral subtraction method substantially reduces the noise levels in the noisy speech. However, it also introduces an annoying distortion in the speech signal called musical noise.

The algorithm, implemented, uses a modified spectral over-subtraction approach that allows better and more suppression of the noise. The power spectrum of the clean speech is estimated by subtracting an overestimate of the noise power spectrum from the speech power spectrum. In addition, a psycho-acoustically motivated spectral weighing rule was incorporated to find the best trade-off between speech distortion and noise reduction. The existing residual noise can be masked by exploiting the masking properties of the human auditory system. A masking threshold estimation was used to eliminate noise influence when determining the speech masking threshold.

## Reference Documents

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Name** | **Document Code** |
| 1 | An improved spectral subtraction method for speech enhancement using a perceptual weighting filter. | <https://doi.org/10.1016/j.dsp.2007.08.002> |
| 2 | Single channel noise suppression for speech enhancement. | <http://plaza.ufl.edu/hejiaxiu/speech.htm> |
| 3 | Transform coding of audio signals using perceptual noise criteria. | DOI: [10.1109/49.608](https://doi.org/10.1109/49.608) |
| 4 | Noise estimation techniques for robust speech recognition | DOI: [10.1109/ICASSP.1995.479387](https://doi.org/10.1109/ICASSP.1995.479387) |
| 5 | Single channel speech enhancement based on masking properties of the human auditory system. | DOI: [10.1109/89.748118](https://doi.org/10.1109/89.748118) |
| 6 | Optimizing digital speech coders by exploiting masking properties of the human ear. | DOI: [10.1121/1.383662](http://dx.doi.org/10.1121/1.383662) |
| 7 | Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. | <https://doi.org/10.1016/j.procs.2015.06.066> |
| 8 | Multimedia signals and systems | http://engrwww.usask.ca/classes/EE/862/notes/mycme462-MP3.pdf |
| 9 | Enhancement of speech corrupted by acoustic noise. | DOI: [10.1109/ICASSP.1979.1170788](https://doi.org/10.1109/ICASSP.1979.1170788) |
| 10 | Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. | https://pdfs.semanticscholar.org/0025/225f5110c7e1053ce08885b06a766f39d976.pdf |

Table 1: List of Reference Documents

## Assumptions, Constraints & Risk

* The properties of the noise as well as the nature of corruption vary in various scenarios, so we make the following assumptions in our project. The background noise is slowly varying, additive and independent with the speech. It is relatively long-time stationary compared with the speech so that its spectral magnitude expected values during the speech activity remain almost the same as prior to the speech activity.
* A usable voice frequency band ranges from approximately 300 Hz to 3400 Hz. As per the Nyquist-Shannon sampling theorem, the sampling frequency must be at least twice the highest component of the voice frequency. The bandwidth allocated for a single voice-frequency transmission channel is usually 4 kHz. So we can safely assume a sampling frequency of 8kHz, without loss of any information.

* Frame Size : The frame size has been set to 32 ms. At a sampling frequency of 8 kHz, this gives us a frame size of 256 points. The noise is assumed to be stationary over this frame. Using a frame shorter than 20 ms results in roughness while if the frame is too long slurring results.
* Window Overlap: Hanning window is used in order to overlap and add adjacent segments of processed speech. The overlap is necessary to prevent discontinuities at frame boundaries. The amount of overlap is taken to be 50%.
* FFT Order : 256 point FFT is done in order for frequency transformation (minimum order FFT corresponding to the given frame size)
* If the background noise is not stationary enough, then our assumption of stationarity over the given frame size is not obeyed and one can observe a “musical noise” being introduced into the signal after filtering. The musical noise is reduced by using psycho-acoustic masking.
* It has been observed that the algorithm introduces some distortion at very low SNR (<0 dB).
* The absolute threshold of hearing of Human Auditory system has been factored into the algorithm. It is advised to use that, at very low SNR (<5 dB) and comment it out at SNR > 10 dB. This will ensure that the distortion introduced during the process of denoising by the algorithm is least.

## Overview

The algorithm aims to enhance to enhance the quality of speech degraded by additive background noise. The goal is to improve the listen-ability of the speech signal by decreasing the background noise, without affecting the intelligibility of the speech. The noise is at such levels that the speech is essentially unintelligible out of context. The average segmental signal-to-noise ratio (SNR) is used to measure the noise level of the noise-corrupted speech signal.

After an initial investigation of several methods of speech enhancement, we conclude that the method of spectral subtraction for speech enhancement using a perceptual weighting filter is more effective than others. The initial aim of the algorithm was to remove noise which is stationary over a long period. However, it has been observed that the current algorithm performs quite well in noise conditions such as babble, factory floor and car which have limited stationarity.

# Description

## Data Specifications

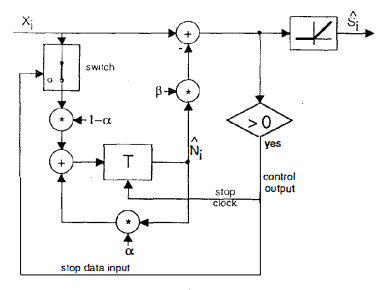
* Xi(k) *:*  Spectral magnitude at time k in sub-band i (of the input noisy signal)
* Ňi(k) *:* Estimation of the noise magnitude at time k in sub-band i
* *α :* Over estimation factor
* *β :* Spectral floor factor
* *NSNR :*  Noisy Signal to Noise Ratio
* Ŝ(ejw) : Estimate of the clean speech
* Ť(ejw) *:* Masking threshold of clean speech
* G(ejw) *:* Perceptual weighting filter

## Basic Concept

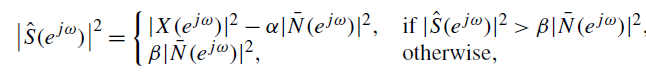
* **Frequency Transform** : A fast Fourier transform (FFT) [algorithm](https://en.wikipedia.org/wiki/Algorithm) computes the [discrete Fourier transform](https://en.wikipedia.org/wiki/Discrete_Fourier_transform) (DFT) of a sequence, or its inverse (IFFT). [Fourier analysis](https://en.wikipedia.org/wiki/Fourier_analysis) converts a signal from its original domain (in time) to a representation in the [frequency domain](https://en.wikipedia.org/wiki/Frequency_domain) and vice versa. An FFT rapidly computes such transformations by [factorizing](https://en.wikipedia.org/wiki/Matrix_decomposition) the [DFT matrix](https://en.wikipedia.org/wiki/DFT_matrix) into a product of [sparse](https://en.wikipedia.org/wiki/Sparse_matrix) (mostly zero) factors. ( MATLAB command : fft() )
* **Segmentation into frames**: The input audio signal is divided into frames of size 256 (32 ms). The windowing is done using a Hanning window and 50% overlap.
* **Voice Activity Detection (VAD)** : In each speech frame, the energy in the frame, E, the linear prediction error normalized with respect to the energy of the signal, LPE, and the zero-crossing rate, ZCR, are calculated. Using all these three parameters, a compound parameter, D, is calculated as D=E(1-ZCR)(1-LPE). From all the frames of the signal, Dmax is computed. Then the value of  D/Dmax is used to determine whether a signal has speech activity or not. The threshold values have to be obtained empirically. VAD fails at very low SNR, so adaptive noise spectral estimation is used along wit VAD. (The index obtained from VAD is used in time frequency filtering)
* **Adaptive noise spectral estimation** **[4]** : A recursive system used to estimate the noise spectral magnitude. This method calculates the weighted sum of past spectral magnitude values Xi in each sub-band i. The weighting system is done by a simple first order recursive system.

Ňi(k) = (1-α) \* Xi(k) + α \* Ňi(k-1)

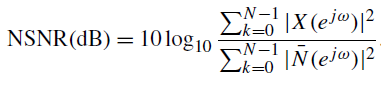
Where Xi(k) denotes the spectral magnitude at time k in sub-band i and Ňi(k) is an estimation of the noise magnitude. The noise estimate Ňi is calculated with a first order recursive system. Ňi is multiplied with an over estimation factor in the usual range of about 1.5 to 2.5. For positive values of (Xi - βŇi ) the data input as well as the recursive accumulation are stopped. This indicates an onset of speech. Negative values of (Xi - βŇi ) are set to zero to get an estimate Si of the clean speech.

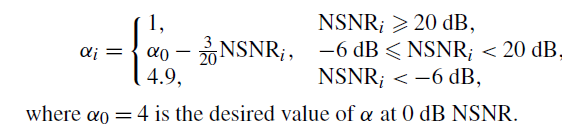


* **Adaptive Spectral Over-subtraction [1]** : Due to the differences between estimated and effective noise a residual noise appears after applying spectral subtraction method. This is perceived as a distortion in the speech signal called musical noise. To reduce this effect, an overestimate of the noise power spectrum is subtracted and the resulted spectrum is limited from going below a preset minimum level (spectral floor). The proposed algorithm can be expressed as :



Where α is the subtraction factor and β is the spectral floor parameter. |Ŝ(ejw)| is the estimate of the spectral magnitude of the clean speech signal. The value of α is adapted from frame to frame depending on the segmental noisy signal to noise ratio (NSNR) of the frame, in order to apply less subtraction with high NSNRs and vice versa.





* **Psycho-acoustic Masking (Noise masking threshold calculation)** : In order to further enhance the quality of speech, a psycho-acoustically motivated spectral weighting rule has been incorporated. Some musical noise remains in the estimated clean speech after spectral over-subtraction. The existing musical noise can be masked by exploiting the masking properties of the human auditory system. Masking is present because the auditory system is incapable of distinguishing two signals close in time and frequency domain.

The noise masking threshold is obtained through modeling the frequency selectivity of the human ear and its masking property. The different calculation steps are :-

1. *Frequency analysis along a critical band scale, or Bark scale :* The critical band analysis is performed on the fast Fourier transform (FFT) power spectrum by adding up the energies in each critical and *k*, according to the values given in the following table 1.
2. *Convolution with a spreading function SF(k) to take into account masking between different critical bands:*  The function used in this algorithm has been taken from **[6]**.
3. *Subtraction of a relative threshold offset O(k) depending on the noise-like or tone-like nature of the masker and the maskee:* The method which has been followed in this algorithm has been taken from **[3]**. The noise-like or tone-like nature of the signal is determined by using the Spectral Flatness Measure (SFM). The SFM is further used to generate the coefficient of tonality which helps in determining the threshold offset.
4. *Renormalization and comparison with the absolute threshold of hearing* **[3]***:* The convolution of the spreading function must be undone i.e., the threshold calculated should be deconvoluted. This process is very unstable due to the shape of the spreading function and often leads to artifacts. In place of deconvolution, renormalization is used. The renormalization multiplies each Ti by the inverse of the energy gain, assuming a uniform energy of 1 in each band.

Since the masking thresholds have thus far been calculated without reference to absolute level, they must be checked to make sure that they do not demand a level of noise below the absolute limits of hearing. The current algorithm should use the absolute threshold of hearing only at very low SNRs (has to be done manually).

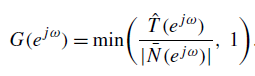
1. *Modification in the masking threshold :* We are performing the threshold calculation on the rough estimate of the clean speech signal. The residual noise modifies the tonality of the signal and the masking threshold is slightly different from the one obtained from the clean speeches especially for high frequencies (if the critical band number k > 15). Refer “Estimation of the masking threshold for perceptual weighting” in paper **[1]**.

* **Perceptual Weighting [1] :** After following the calculations as described in [1] , we arrive at the following conclusions. Let Ŝ(ejw) = G(ejw)X(ejw) be the enhanced speech spectrum and G(ejw) the perceptual weighting filter.

The weighting function G(ejw) has to satisfy the following criteria :

Capture

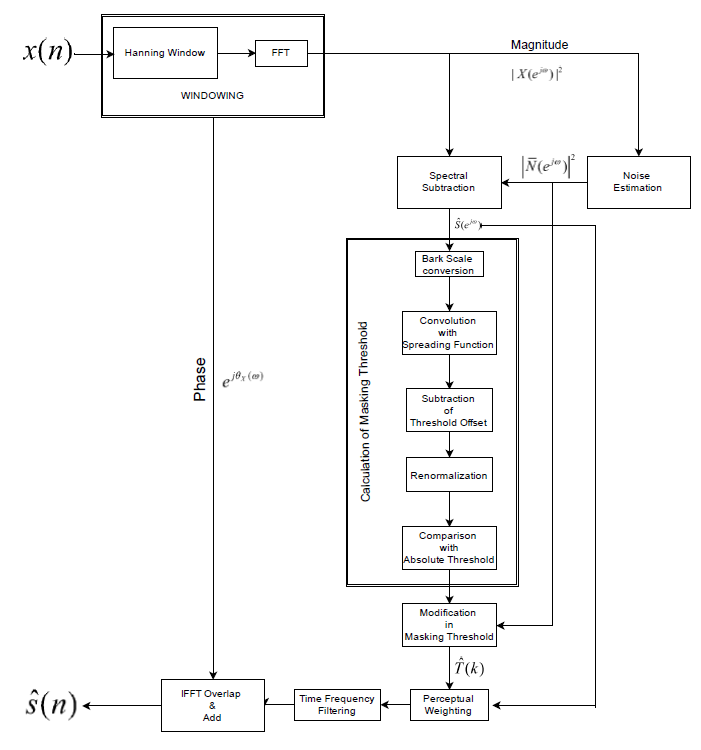
With the constraint 0 ≤ G(ejw) ≤ 1, where Ť(ejw) is the clean speech estimated masking threshold. Therefore the psychoacoustically motivated weighted filter is obtained as follows:

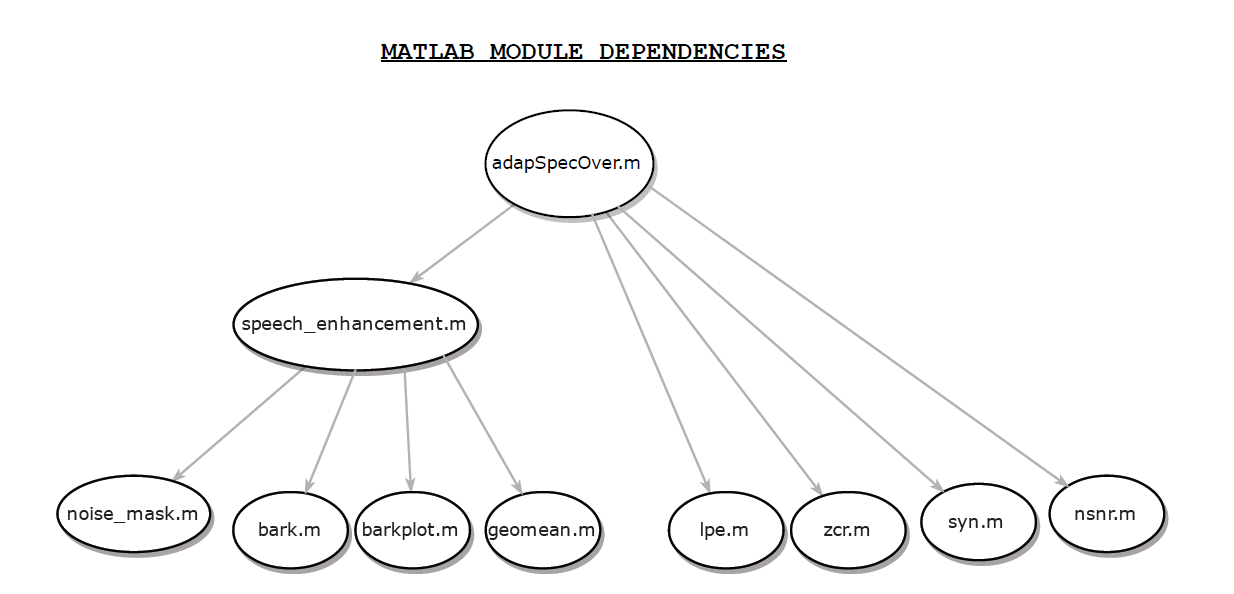


* **Time-Frequency filtering :** Time-frequency filtering is used to reduce the resultant musical noise. It is performed using several preceding frames and several frames following the frame of interest.

## Basic and Alternate Flows

*Block diagram of the proposed spectral subtraction method with perceptual weighting.*

**



# Error Analysis

The algorithm introduces musical noise at very low SNRs. To explain the nature of musical noise, one must realize that peaks and valleys exist in the short-term power spectrum of white noise; their frequency locations for one frame are random and they vary randomly in frequency and amplitude from frame to frame. When we subtract the smoothed estimate of the noise spectrum from the actual noise spectrum, all spectral peaks are shifted down while the valleys (points lower than the estimate) are set to zero (minus infinity on the logarithmic scale). Thus, after subtraction there remain peaks in the noise spectrum. Of those remaining peaks, the wider ones are perceived as time varying broadband noise. The narrower peaks, which are deep valleys that define them, are perceived as time varying tones which we refer to as musical noise.

The worst case situation occur at very low SNRs and when the assumption of stationarity of noise becomes void. In that case maximum musical noise is introduced by the algorithm. Pschoacoustic masking covers this musical to to a large extent.

For analyzing the algorithm’s error methods are discussed in the Section 4 (Validation).

# Validation

## Approaches

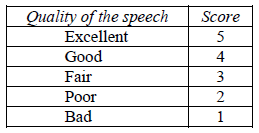
For validation of the algorithm, two approaches have been used, SNR improvement and the PESQ **[10]** algorithm. The SNR improvement is the performance evaluation for calculating the amount of noise reduction in the background noise level conditions. The main drawback of SNR is the fact that it has a poor correlation with subjective quality assessment results. Therefore, the SNR of enhanced speech is not a sufficient indicator of speech quality.

The Perceptual Evaluation of Speech Quality (PESQ) is an objective quality measure designed to predict the subjective opinion score of a degraded audio sample and it is recommended by ITU-T for speech quality assessment. In PESQ measure, a reference signal and the processed signal are first aligned in both time and level. The PESQ measure is reported to be highly correlated with subjective listening tests. The PESQ is one of the best measures of signal’s quality.

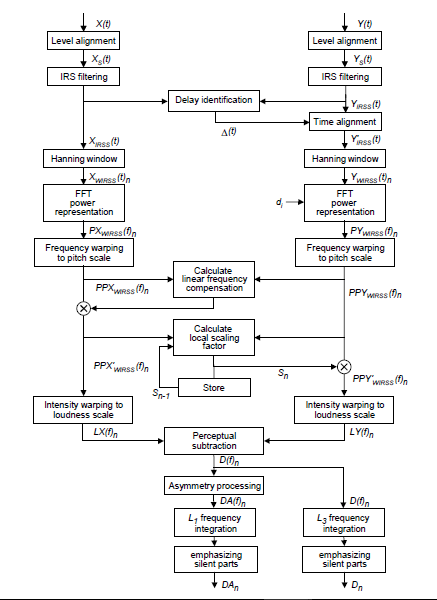
In PESQ algorithm the original and the degraded signals are mapped onto an internal representation using a perceptual model. The difference in this representation is used by a cognitive model to predict the perceived speech quality of the degraded signal. The perceived listening quality is expressed in terms of Mean Opinion Score, an average quality score over a large set of subjects.

However it should be noted that the PESQ scores might not agree with the subjective listening tests in some cases.

Spectral subtractive-type speech enhancement algorithms generate two main undesirable effects: remnant noise and speech distortion. These two effects can be annoying to the human listener, and causes listener fatigue. However, they are difficult to quantify. Therefore, it is important to analyze the time-frequency distribution of the enhanced speech, in particular the musical structure of the remnant noise. The speech spectrogram is a good tool to do this work, because it can give more accurate information about remnant noise and speech distortion than the corresponding time domain waveforms.



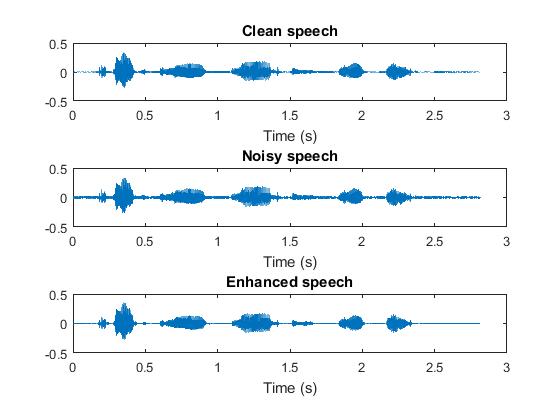
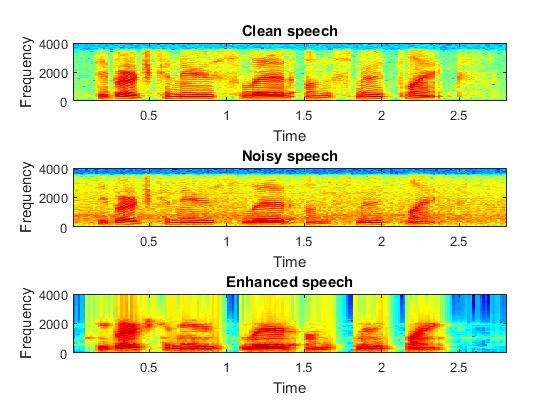
*Normal range of MOS (Mean Opinion Score) values*



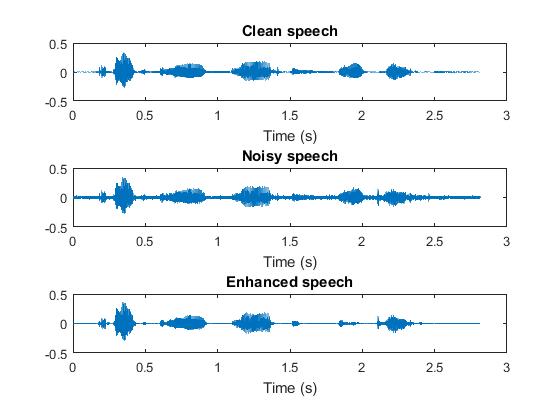
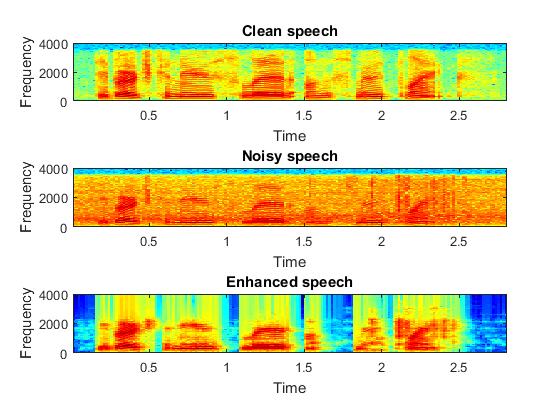
*Overview of the perceptual model (PESQ)*

## Measurements

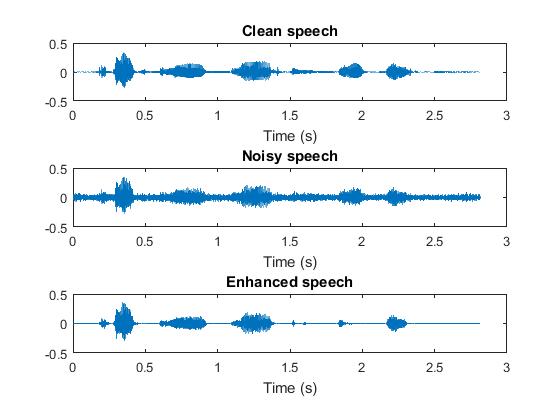
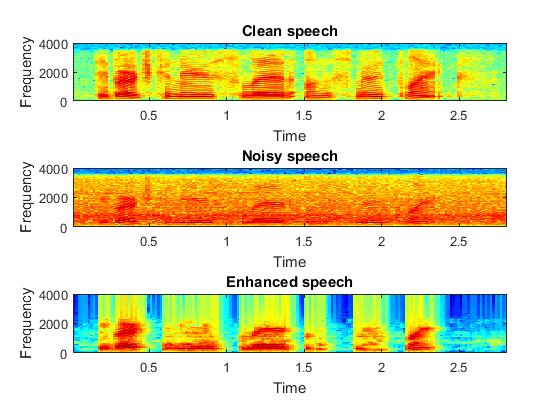
Waveform & Spectrogram for background car noise at *(i)* SNR = 15 dB *(ii)* SNR = 10 dB *(iii)* SNR = 5 dB

**

**(i)**



**(ii)**



**(iii)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SIGNAL** | **SNR\_IN (dB)** | **SNR\_OUT(dB)** | **PESQ\_IN**  **(4.5/5)** | **PESQ\_OUT**  **(4.5/5)** |
| sp01\_car\_sn0\_nom | 0 | 2.0282 | 1.593  1.370 | 1.825  1.505 |
| sp01\_car\_sn5\_nom | 5 | 6.4401 | 1.953  1.595 | 1.959  1.718 |
| sp01\_car\_sn10\_nom | 10 | 14.8112 | 2.111  1.726 | 2.202  1.957 |
| sp01\_car\_sn15\_nom | 15 | 29.5949 | 2.533  2.176 | 2.878  2.609 |
| sp01\_car\_sn0\_abs | 0 | 14.8636 | 1.593  1.370 | 1.837  1.512 |
| sp01\_car\_sn5\_abs | 5 | 32.6918 | 1.953  1.595 | 2.022  1.650 |
| sp01\_car\_sn10\_abs | 10 | 45.2474 | 2.111  1.726 | 2.349  1.959 |
| sp01\_car\_sn15\_abs | 15 | 46.8064 | 2.533  2.176 | 2.841  2.486 |

**NOTE:-**

* sp01\_car\_snxx\_nom : The signal has been processed without using the absolute threshold (till the renormalization step)
* sp01\_car\_snxx\_abs : The signal has been processed using the absolute threshold (using all the steps)

The results show that the classical spectral subtraction algorithm mostly results in audible remnant noise, which decreases speech intelligibility. The most progressive algorithm of speech enhancement is the spectral subtraction based on perceptual properties. This algorithm takes advantage of how people perceive the frequencies instead of just working with SNR. It results in appropriate remnant noise suppression and acceptable degree of speech distortion.

# Abbreviations

|  |  |
| --- | --- |
| ***Abbreviation*** | ***Long form.*** |
| ***FFT*** | ***Fast Fourier Transform*** |
| ***SNR*** | ***Signal to Noise Ratio*** |
| ***LPE*** | ***Linear Prediction Error*** |
| ***ZCR*** | ***Zero Crossing Rate*** |
| ***NSNR*** | ***Noisy Signal to Noise Ratio*** |
| ***SFM*** | ***Spectral Flatness Measure*** |
| ***PESQ*** | ***Perceptual Evaluation of Speech Quality*** |
| ***MOS*** | ***Mean Opinion Score*** |