Mini project report

Student name: Wang, Chaoran

Student ID: 5118930

CONTENT

Abstract

This report is based three different methods of speech enhancement. In the report, some basic techniques are introduced, and three in use methods are partly explained. Decisions on parameters are shown with experiment results. Analysis with different method and different types of noise with different conditions is shown based on experiment results.

1. Introduction

Speech enhancement aims at improving the performance of speech communication systems in noisy environments. Speech enhancement may be applied, for example, to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired.

Most speech enhancement methods may be broadly classified into three categories:

1.Filtering techniques: The basic principle is to design a linear filter/transformation such that when noise speech is passed through it, the noise component will be attenuated.

2.Spectral Restoration: Techniques that formulate the problem as estimation of clean speech spectrum from noisy speech spectrum.

3.Model based speech enhancement: Techniques that utilize speech models to convert the problem into one of parameter estimation (for these models).

In this report, three different filtering techniques were used to compare the quality of speech enhancement, which are spectral subtraction based on minimum statistics (Martin 1994), Suppression of Acoustic Noise (Boll 1979), Weiner filter.

2. Theoretical Background

The basic assumption in three methods is that in noisy speech noise and clean speech are independent with each other, so that a noisy speech can be represent by the following equation:

$$y(n) = s(n) + \eta(n)$$

where $y(n)$ is the noisy speech, $s(n)$ is the clean speech, $\eta(m)$ is the noise speech.

Taking DFT is given:

$$Y(k) = S(k) + N(k)$$

Where $Y(k)$ is the DFT of $y(n)$, $S(k)$ is the DFT of $s(n)$, $N(k)$ is the DFT of $\eta(n)$.

In general speech would be divide into short time frames to analyze the features, so i is introduced to represent the frame number. The Energy of each frame is given below:

$$E\{y^2(i)\} = E\{s^2(i)\} + E\{\eta^2(i)\}$$

where i denotes the frame index.

In the spectral subtraction methods, the general idea is that if we can estimate the noise

speech $\eta(i)$ or $\eta(n)$, then we can eliminate noise by spectral subtracting it from noisy speech.

2.1 Minimum statistics

In this method, Martin (1994) believed that we can estimate the noise speech firstly by finding the minimum energy or power for each frequency in a D length window, where D is a set of frames, and then we can estimate the noise speech by given a factor that compensate the bias of minimum estimate. D length window should be large enough to bridge any peak of speech activity, but short enough to follow non stationary noise variations.

$$P_n(i,k) = omin * P_{min}(i,k)$$

Where $Pn(i,k)$ is the estimated noise power for each frame and each frequency, $omin$ is the compensate factor, and $Pmin(i,k)$ is the minimum power in D length window for each frequency.

Following the subtraction rule of Berouti et. al. (1979) cited in Martin (1994), subtraction magnitudes with an oversubtraction factor $osub(i,k)$ and a limitation of the maximum subtraction by a spectral floor constant $subf$ $(0.01 \leq subf \leq 0.05)$.

$$|S(i,k)| = \begin{cases} \sqrt{subf * P_n(i,k)} & if\ |Y(i,k)| * Q(i,k) \leq \sqrt{subf * P_n(i,k)} \\ |Y(i,k)| * Q(i,k) & else \end{cases}$$

where $Q(i,k) = 1 - \sqrt{osub * \frac{P_n(i,k)}{|Y(i,k|^2}}$, $\overline{|Y(i,k)^2}$ is the smoothed noisy speech subband power, $P_n(i,k)$ is the estimate noise power. $osub(i,k)$ should be a factor that related to i,k,SNR , basically it should have a smaller value for high SNR and for high frequency than that for low SNR and for low frequency. However, there is no certain equation for this factor.

2.2 Suppression of Acoustic Noise in speech

The difference between this method and minimum statistic method is how the noise speech is estimated and how the noisy speech been subtracted. Boll (1979) described

$$S(k) = H_R(k) * Y(k)$$

with

$$H_R(k) = \frac{H(k) + |H(k)|}{2}$$

$$H(k) = 1 - \frac{\mu(k)}{|Y(k)|}$$

where $\mu(k)$ is the mean value of $N(k)$,$N(k)$ can be found in non-speech region. However, we use first 5 frames as non-speech region.

In order to reduce residual noise, a scheme was introduced.

$$\begin{cases} |S_i(k)| = |S_i(k)| & for\ |S_i(k)| \geq \max|N_R(k)| \\ |S_i(k)| = \min\{|S_j(k)|\ j = i-1, i, i+1\ for\ |S_i(k)| < \max|N_R(k)| \end{cases}$$

where $N_R(k)$ is the max value of noise speech in non-speech region.

2.3 Weiner Filter

In general, the noisy speech is divided into M banks of frequency by passing band-pas filters $p_i = 1,2,3 \dots M$, then using gain $k_m\ m = 1,2,3 \dots M$ to eliminate noise speech in each bank, finally the subband signal pass synthesis filters $g_i\ i = 1,2,3, \dots M$ to reconstruct the signal. To implement this method we use Gammatone filter bank with critical frequency fc= [50 150 250 350 450 570 700 840 1000 1170 1370 1600 1850 2150 2500 2900 3400] as p_i , and $g_i = p_i(-n)$ as synthesis filters, and k_m with following equation.

$$K_m = \frac{\sigma_{sm}^2}{\sigma_{sm}^2 + \mu\sigma_{wm}^2}, \qquad \mu \geq 0$$

where, μ is an arbitrary constant that allows a trade-off between signal distortion and noise reduction.

$$\sigma_{sm}^2 = E[s_m^2(n)]$$
$$\sigma_{wm}^2 = E[w_m^2(n)]$$
$$s_i(n) = q_i(n) * s(n)$$
$$w_i(n) = q_i(n) * \eta(n)$$

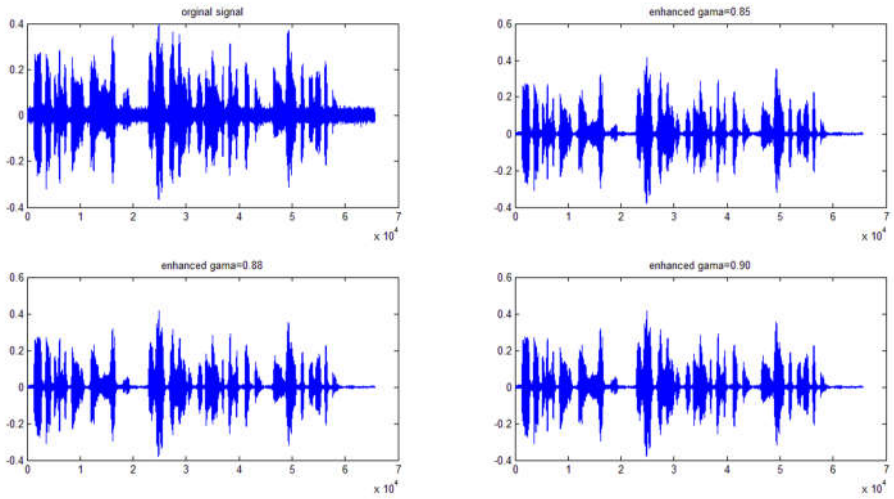
In the experiment, we estimate $\eta(n)$ by using mean value from first 5 frames in speech signal.

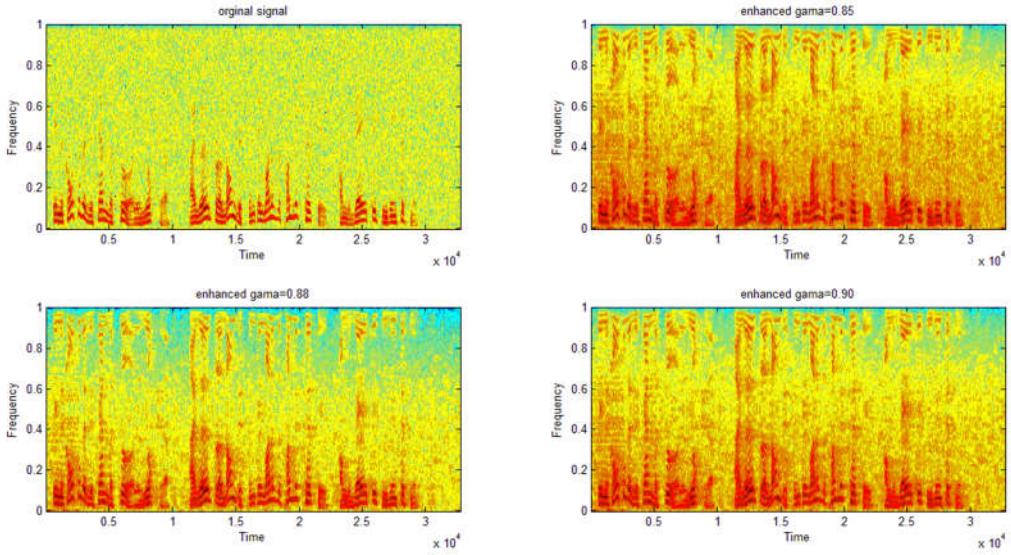
3.Experiment Design and Procedure

3.1 Minimum statistic

Basically, the design just following the paper of Martin. The parameter $omin$ uses the same value mentioned in the paper which is 1.5.

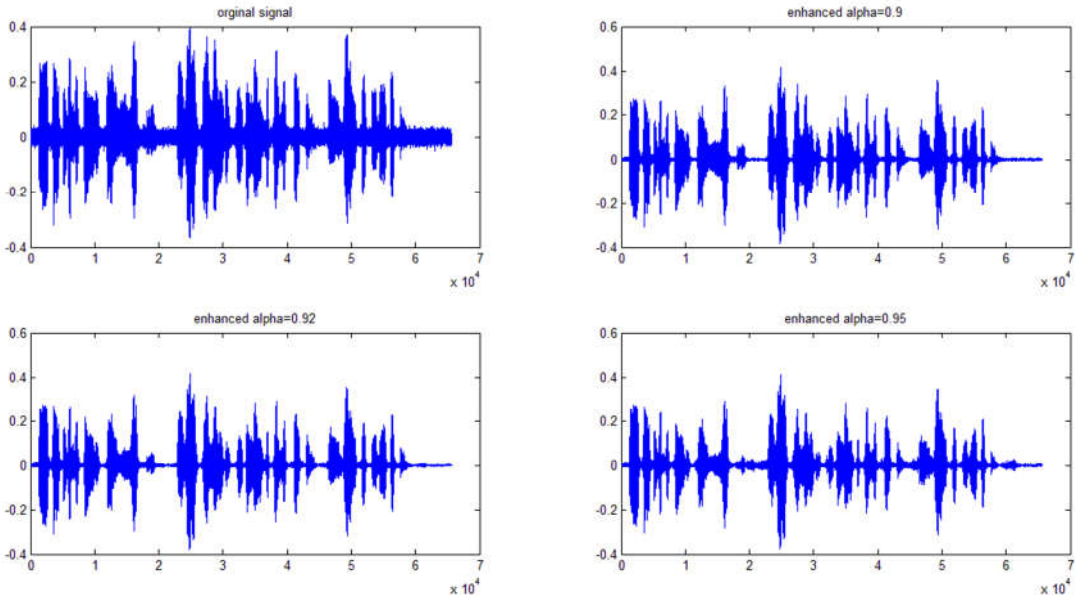
To determine the γ , the following experiment has been done:

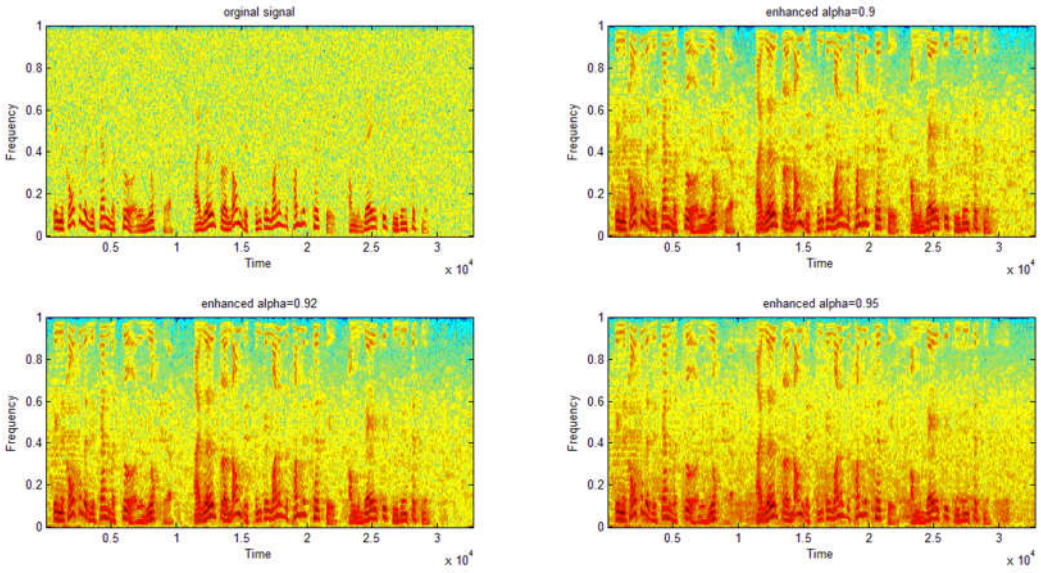




From the comparison between different $\gamma = 0.85, 0.88, 0.9$, the results show significant differences. Therefore, we decide to use $\gamma = 0.88$.

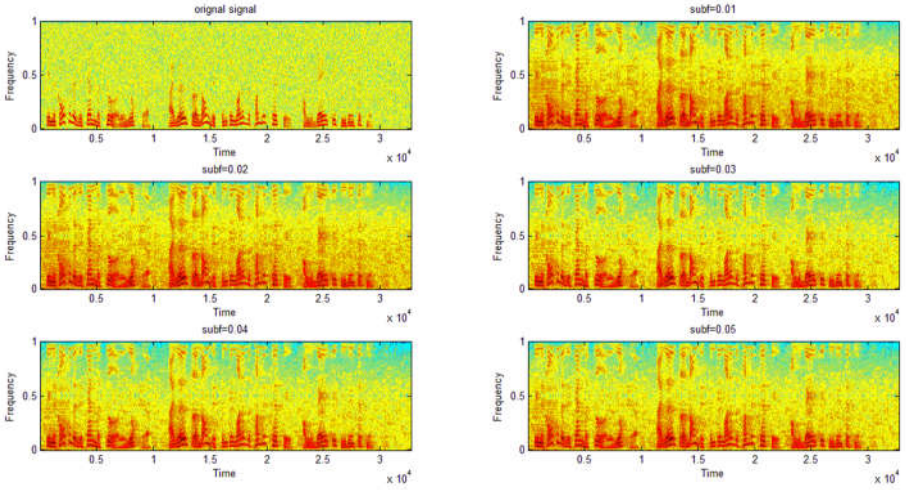
To determine α the following experiment has been done:





These results are done by set $\gamma = 0.88$ with α varying in 0.9 , 0.92 and 0.95 . Results do not show significant differences, therefore, $\alpha = 0.92$ is been used.

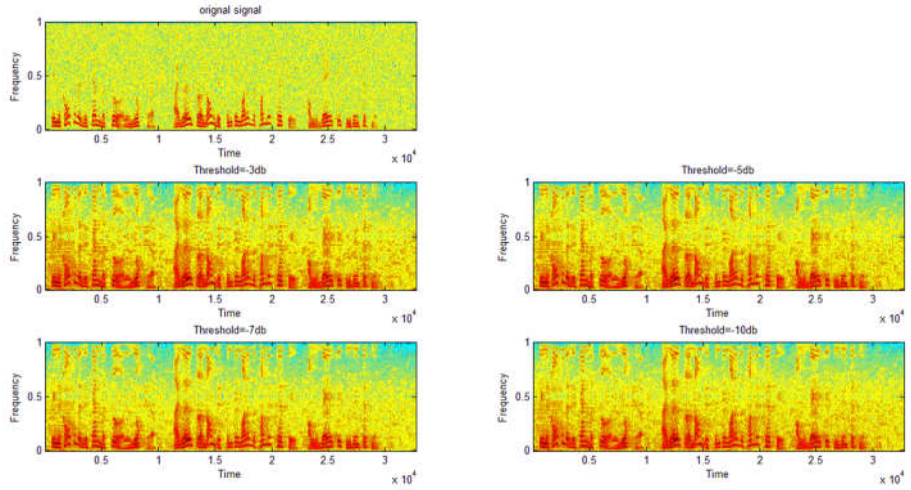
To determine $subf$:



Results do not have significant result, hence, $subf = 0.04$ is used.

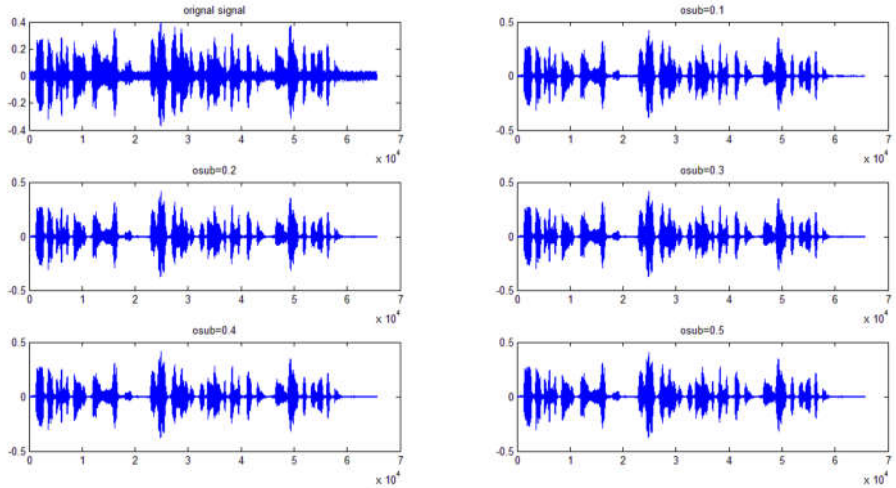
Because of there is no certain equation to determine $osub$ in paper, we simply use noise-to-signal ratio to determine this parameter.

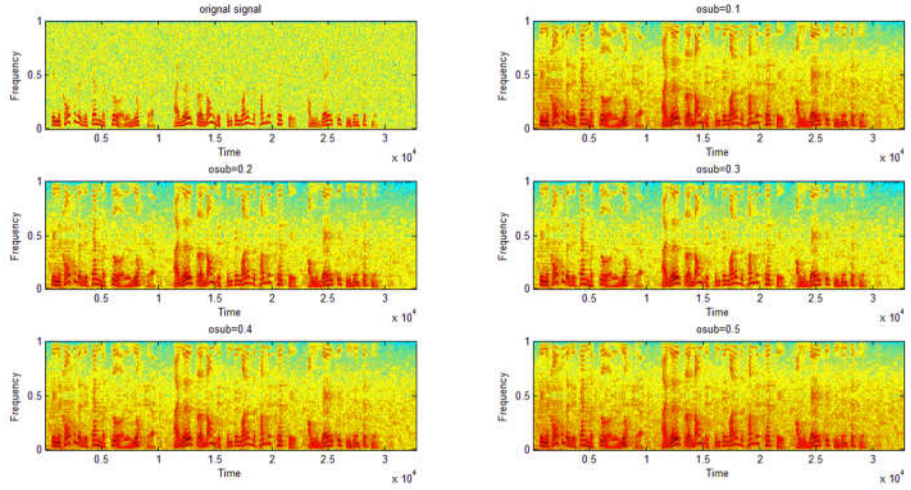
In code, the equation $NSR = 10 * \log_{10} \frac{(P_n(k)}{|X(k)|^2}$ is used to determine the threshold



Eventually, threshold use -7dB.

Then $osub$ is been set to 0.2 when NSR is larger than -7dB, by the following results:

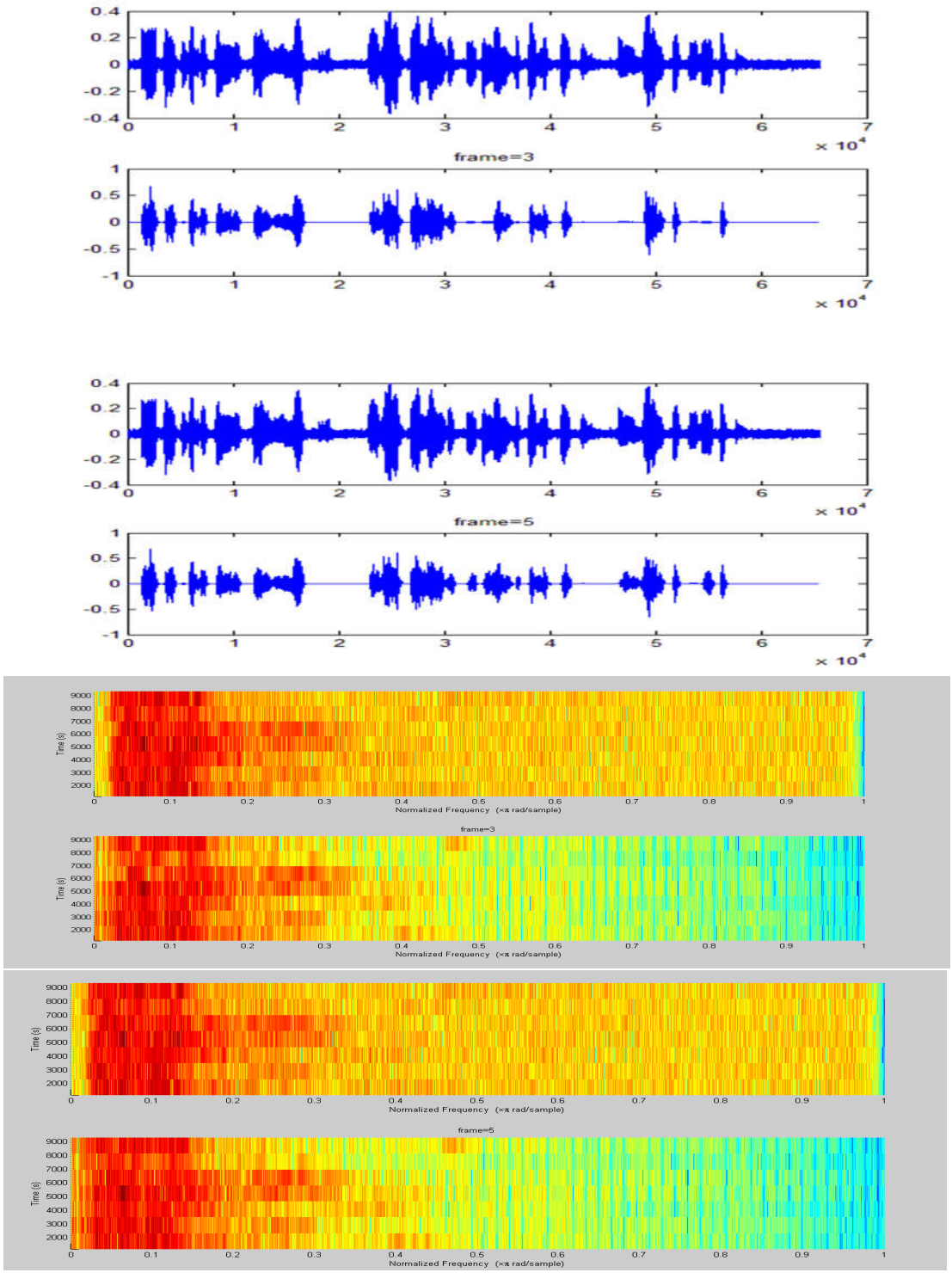




The result of using this method to eliminate different type of noise will be in next section.

3.2 Suppression of Acoustic Noise in speech

In the paper, the noise speech is estimated by calculating mean value of the non-speech region. However, because of bad performance on voice activity detection which we have, we decided to use first several frames to estimate the noise speech. The result shows that it shows a better result when using 5 frames dealing with white noise and babble noise.



The result of using this method to eliminate different type of noise will be in next section.

3.3 Weiner Filter

In this part, Gammatone filter bank is used. The impulse response of Gammatone Filter is given:

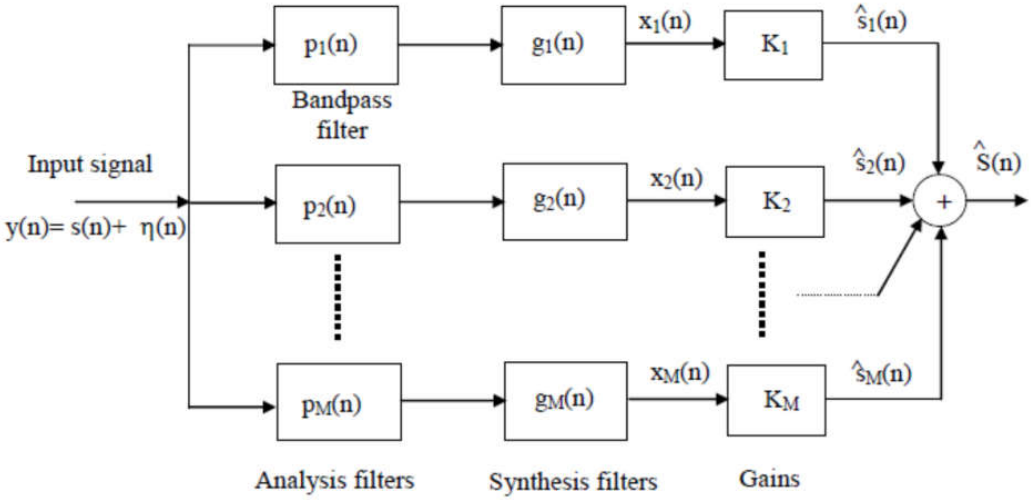
$$p(t) = a * t^{N-1} e^{-2*\pi*b*ERB(fc)*t} \cos(2 * \pi * fc * t) * u(t)$$

where $ERB(fc) = 24.7 + 0.108 * fc$ is the Equivalent Rectangular Bandwidth of the critical band, $a = 1, b = 1.109, N = 4, t = nT$. The center frequency (fc in Hz) of the filters are as follows:

$$fc = [50\ 150\ 250\ 350\ 450\ 570\ 700\ 840\ 1000\ 1170\ 1370\ 1600\ 1850\ 2150\ 2500\ 2900\ 3400]$$

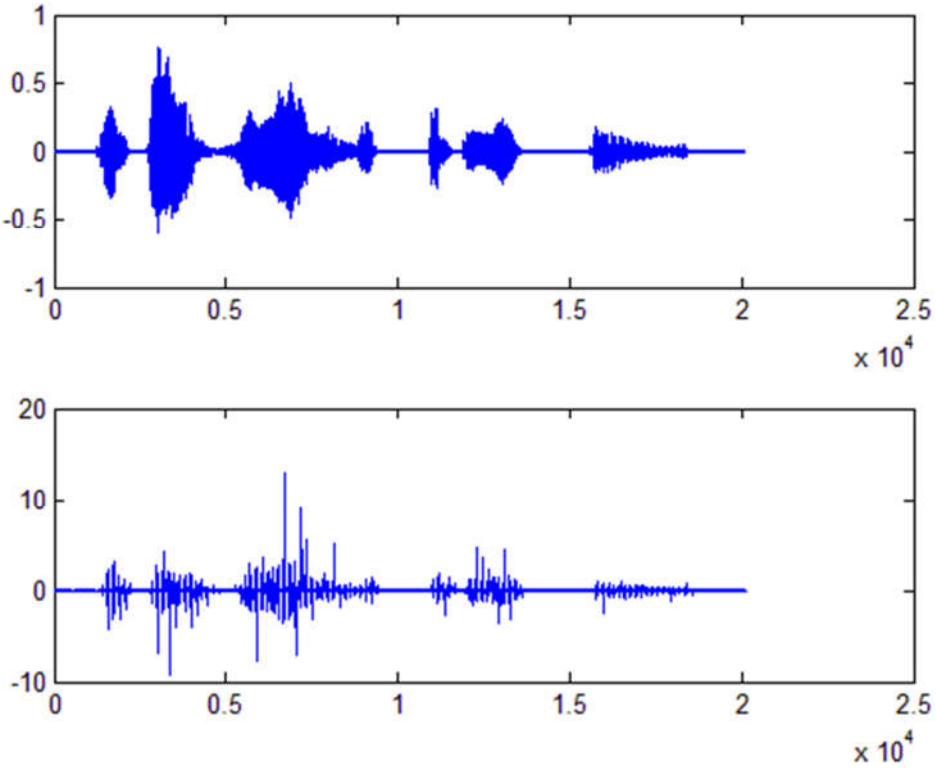
The analysis filters $p_m(n)$ are FIR Gammatone filters, with filter length of L= 160 samples each. The synthesis filters, $g_m(n)$ are obtained by time reversing the corresponding analysis filters,$p_m(n)$.

The whole system looks like this:

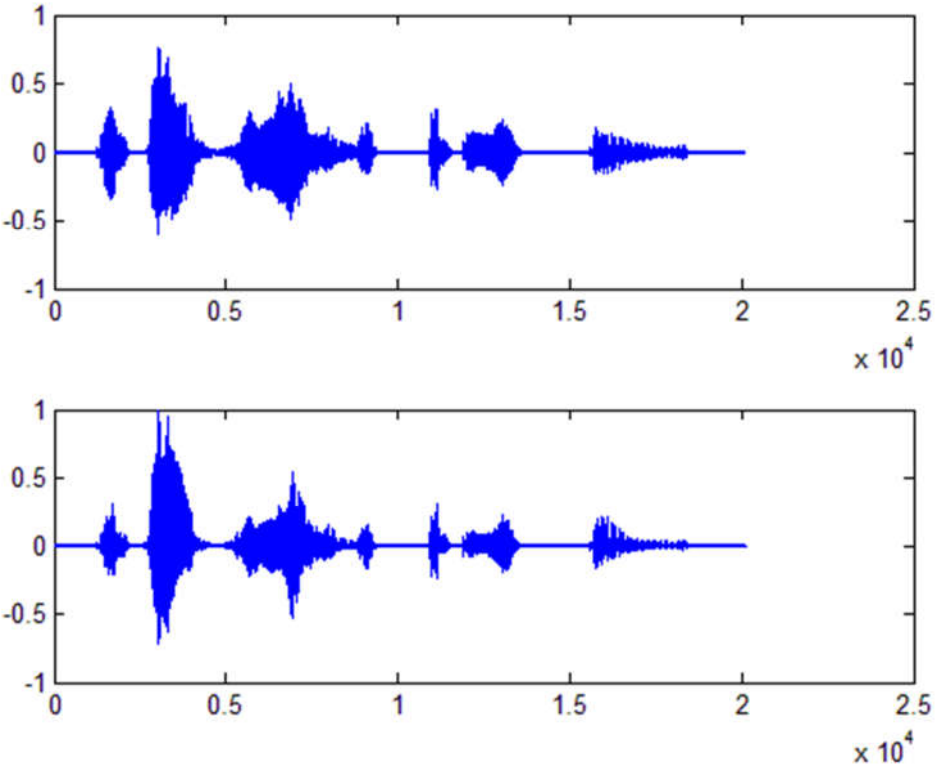


Because of every point in speech signal is not independent with each other, when using filter, it is necessary to set final condition for the filters for each frame as the initial condition of those filters for the next frame.

Figures below show when not set final condition as initial condition and when set initial condition respectively.



Without initial condition



With initial condition

4.Results of experiment

4.1 Minimum statistic method

4.1.1 White noise

Using minimum statistic method to eliminate noise shows a pretty good result when dealing with white noise, from 0dB to 10dB.Figure 4.1-4.3 illustrate the original signal of noisy speech and its spectrogram of white noise with 0dB,5dB,10dB respectively.

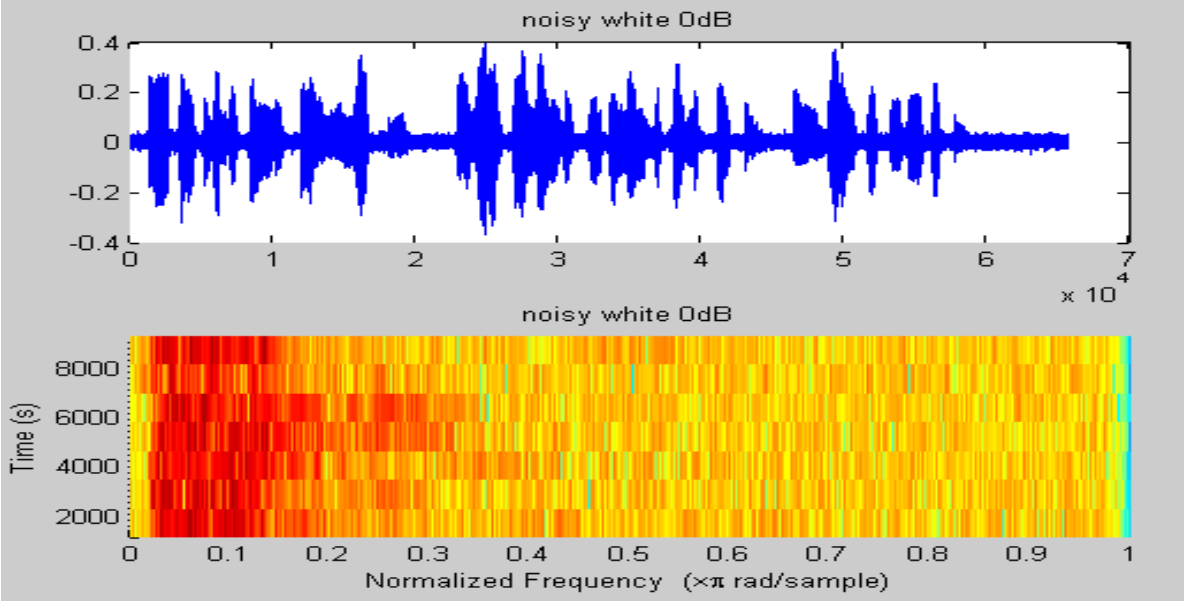


Figure 4.1

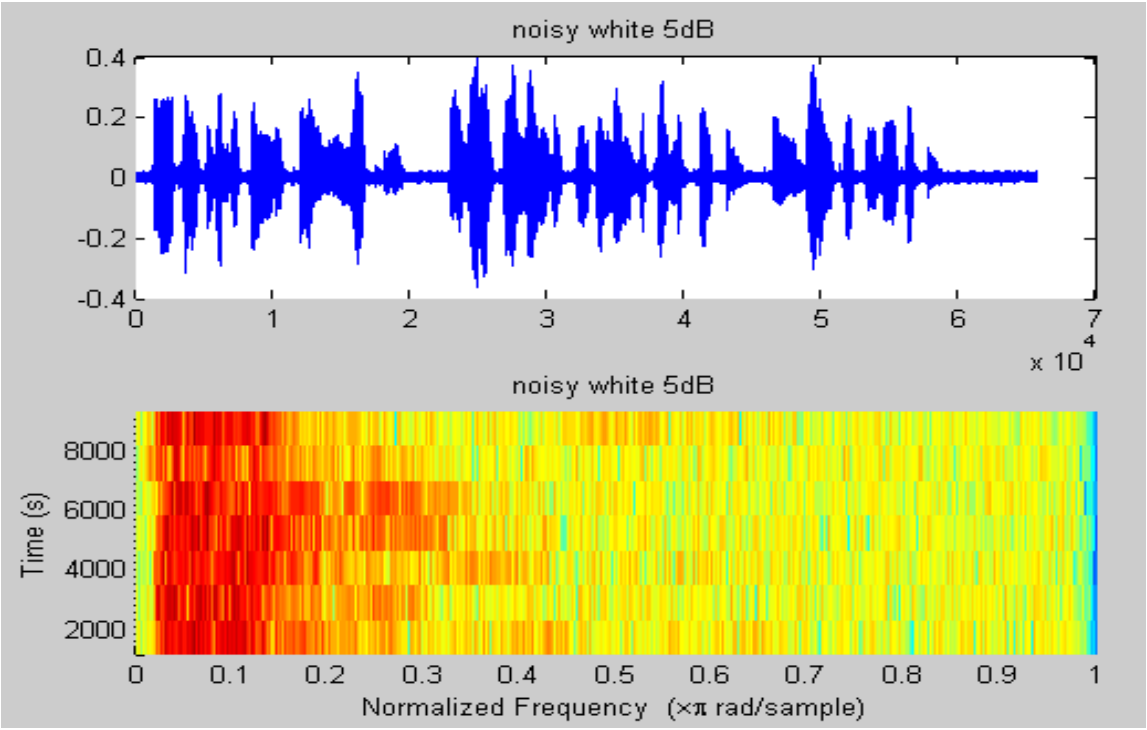


Figure 4.2

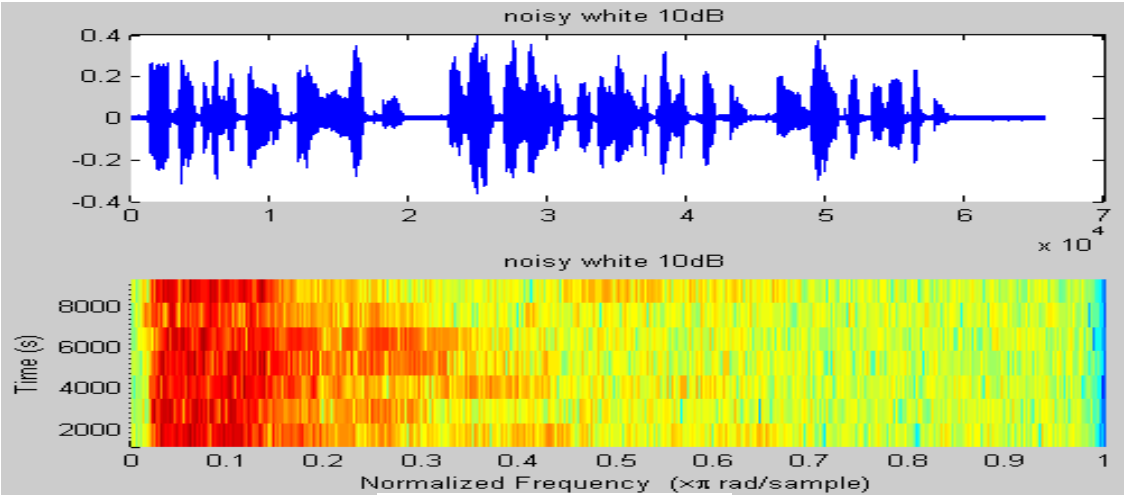
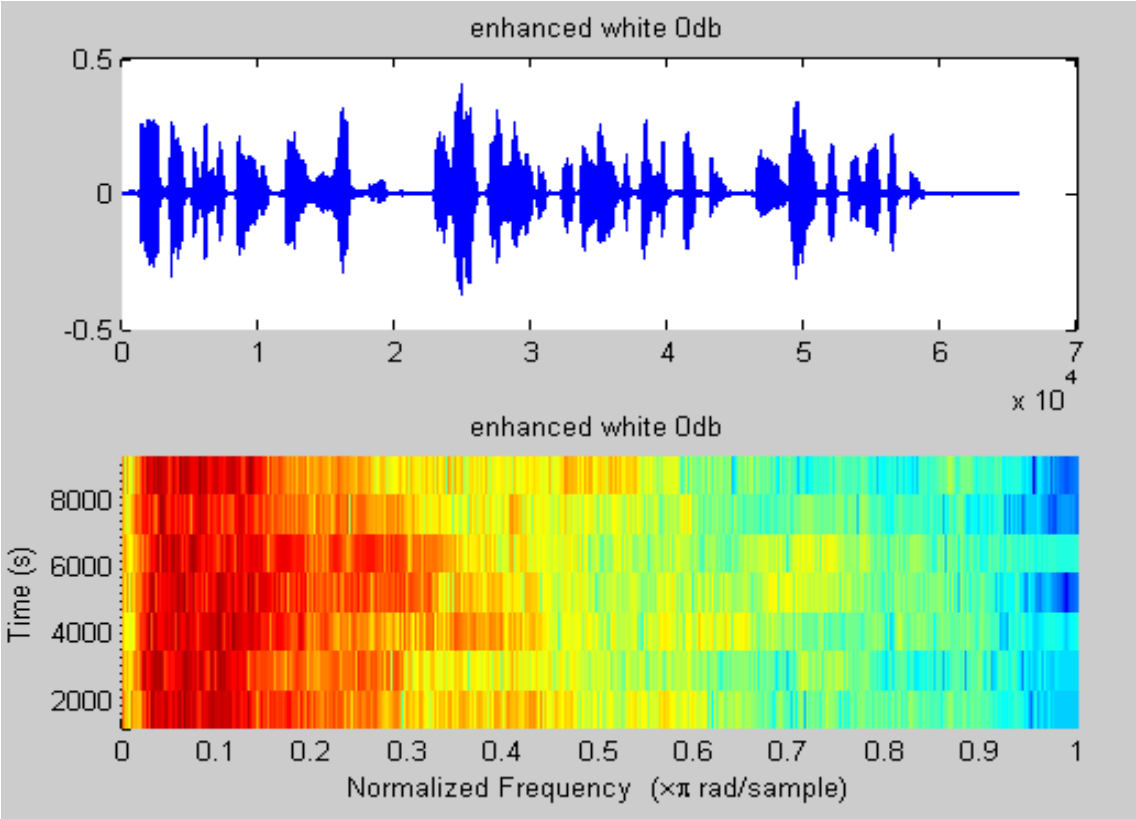
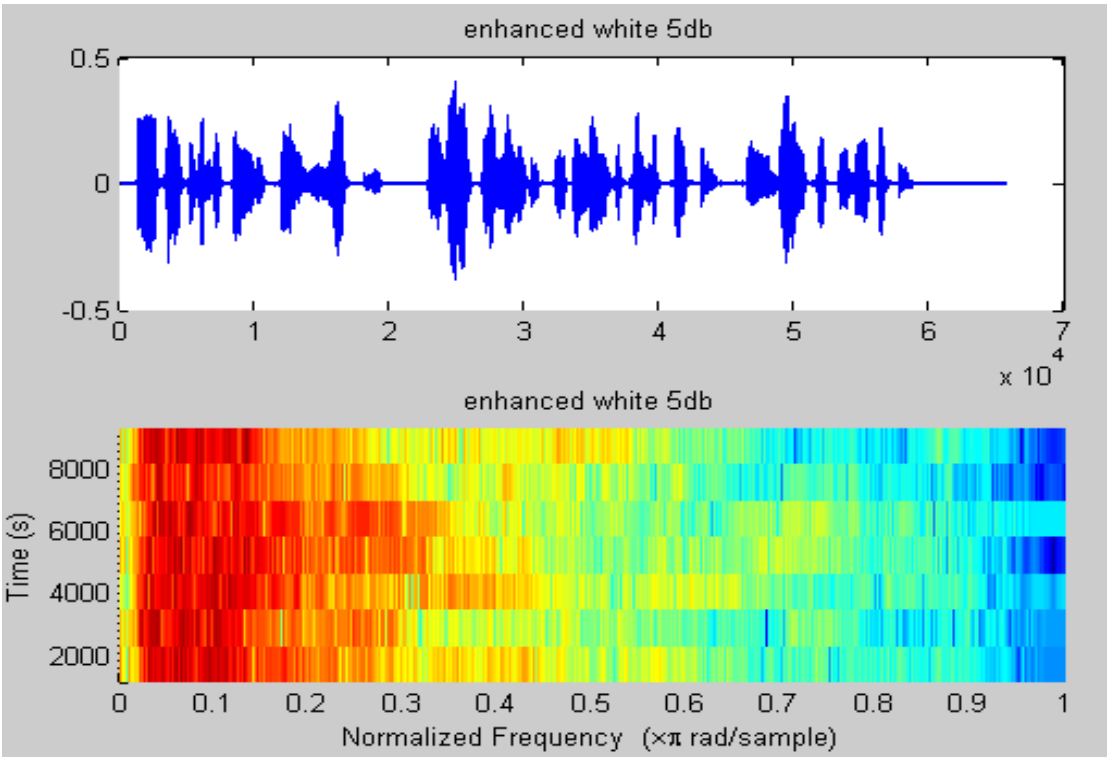


Figure 4.3

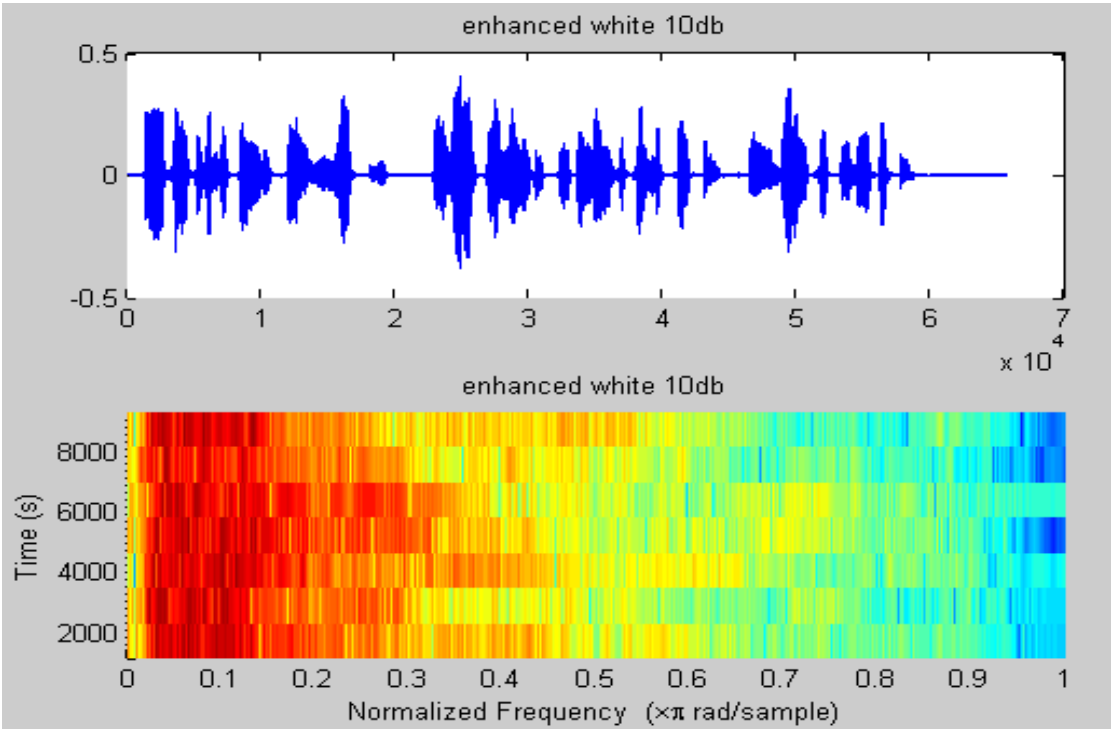
The results of using minimum statistic method are as follows.



Form time domain, it shows a good result at noise elimination. Moreover, it shows same result at frequency domain, it eliminates noise at high frequency. SNR is also introduced to evaluate the performance of speech enhancement, with the equation $SNR = 10 * \log_{10}(P_s/P_n)$, where P_s is the power of the speech, P_n is the power of the noise. The SNR of original speech, noisy white 0dB, is 38.5dB. SNR of enhanced speech, enhanced white 0dB, is 50.1dB. The increase in SNR also shows a good enhancement result.



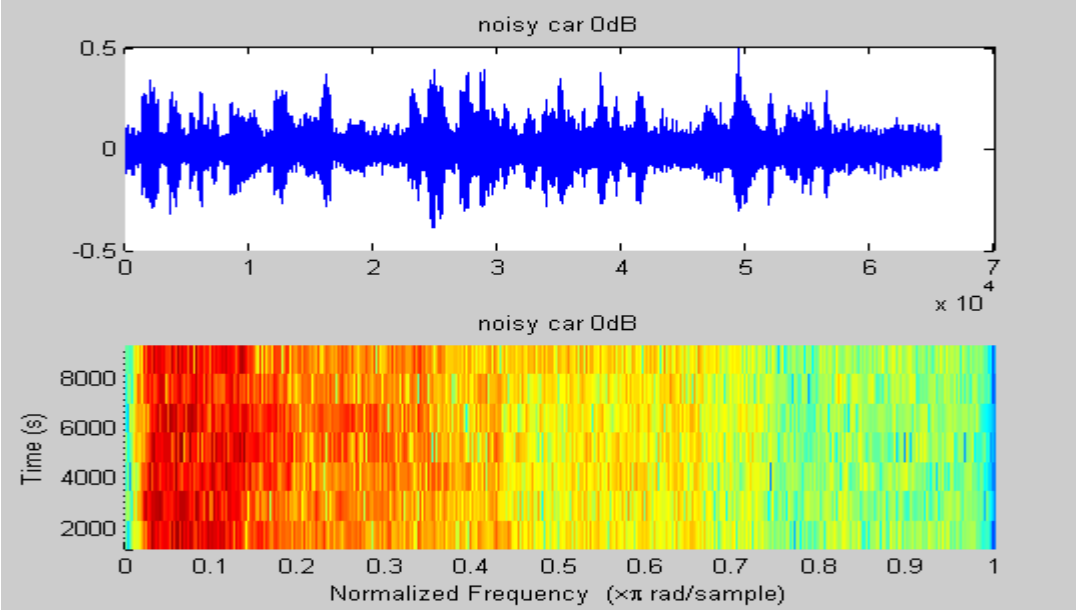
Form time domain, it shows a good result at noise elimination. Moreover, it shows same result at frequency domain, it eliminates noise at high frequency. Compared with white 0dB speech, speech at high frequency is cleaner than that in white 0dB speech. The SNR of original speech, noisy white 5dB, is 43.5dB. SNR of enhanced speech, enhanced white 0dB, is 55.9dB. The increase in SNR also shows a good enhancement result.

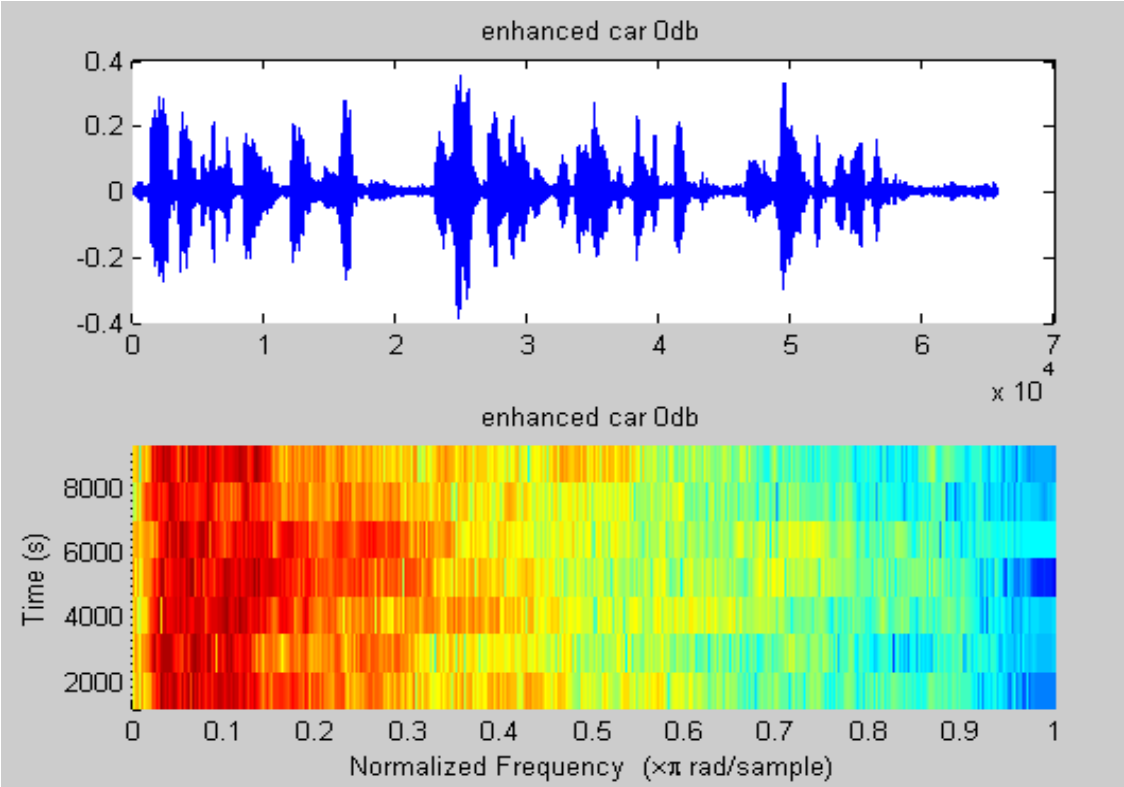


The SNR of original signal, noisy white 10dB, is 48.4dB. SNR of enhanced speech is 60.98dB. These three figure all illustrate a good stable enhancement when dealing with

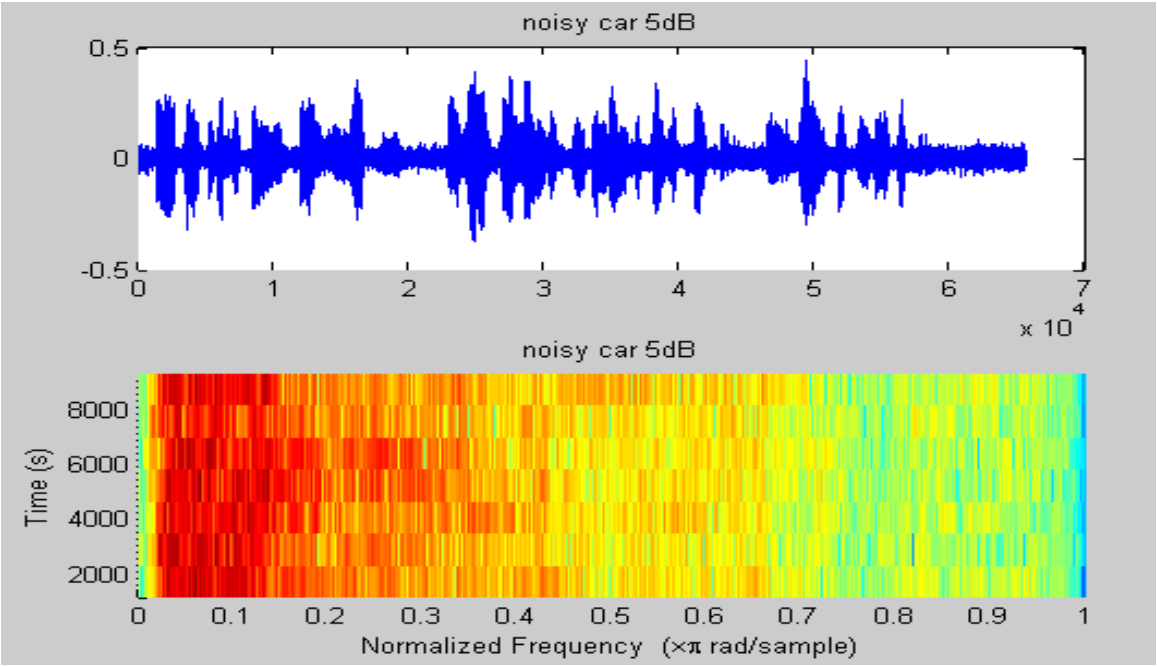
white noise, with an approximate 12dB increase.

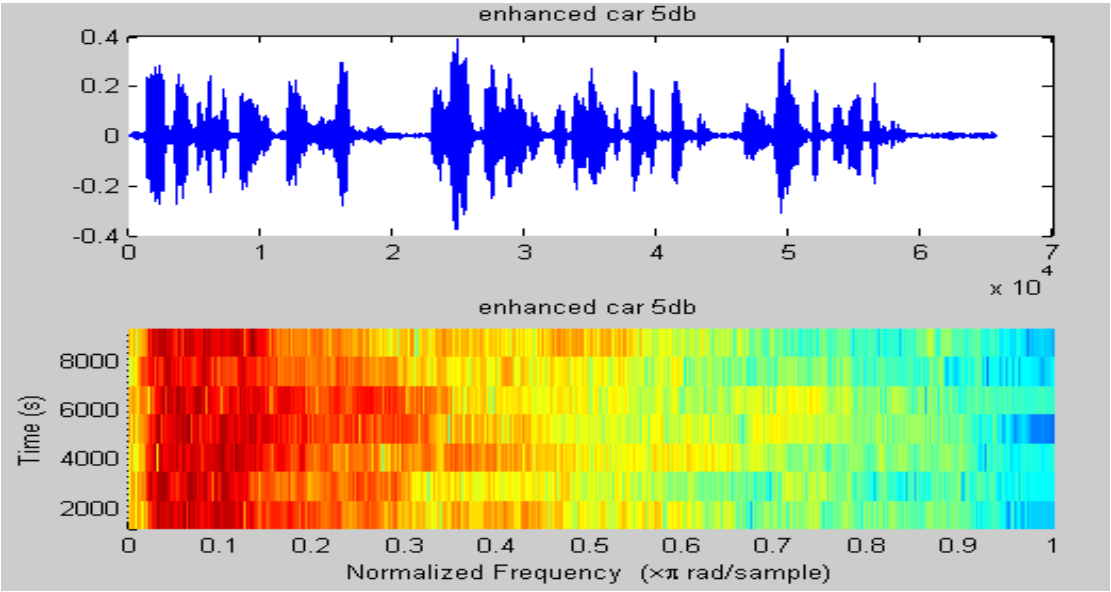
4.1.2 Car noise



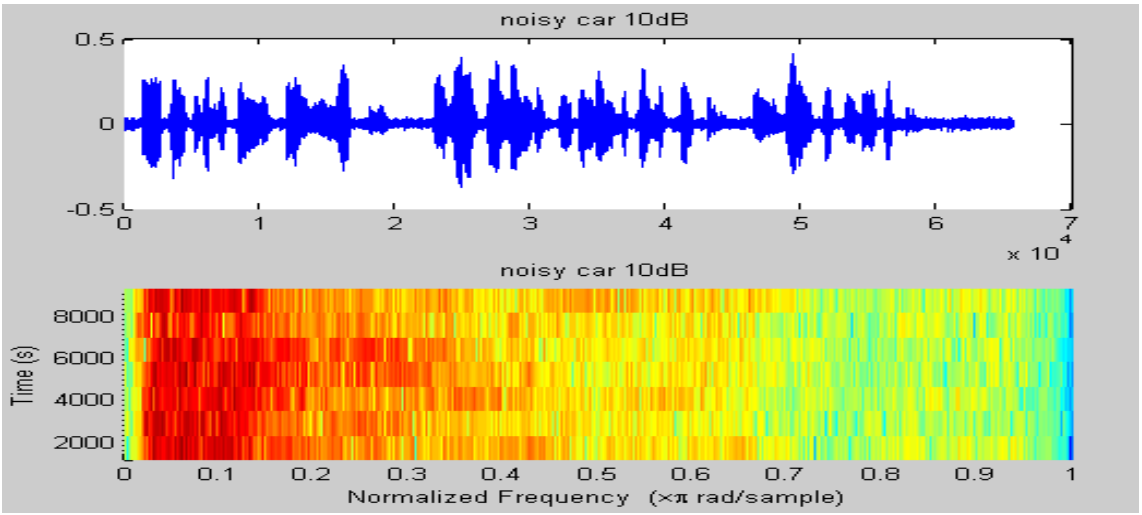


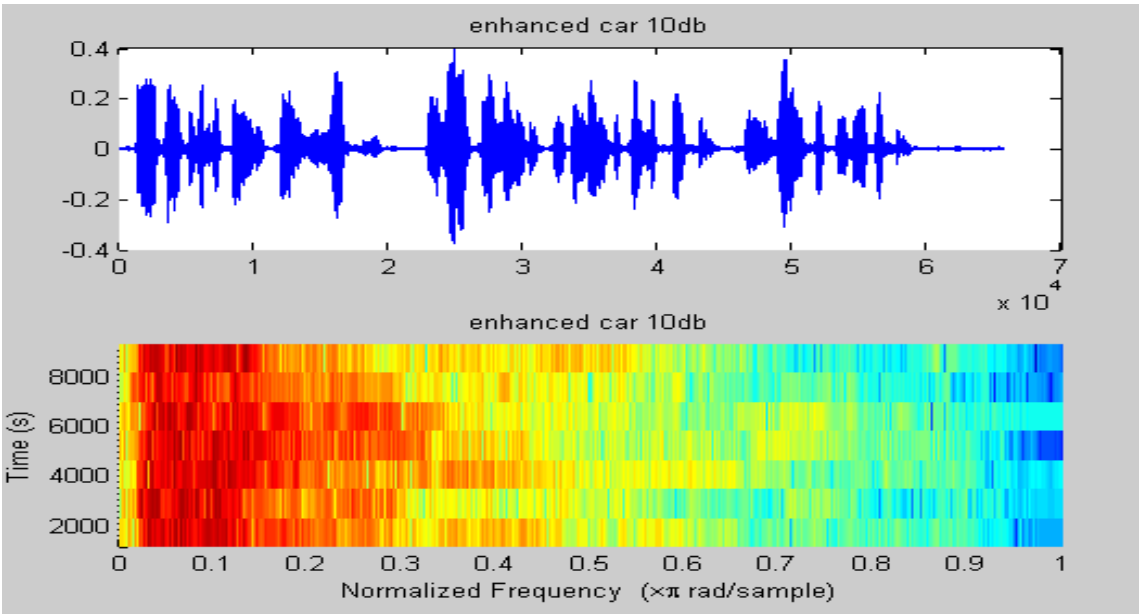
These two figures are original noisy car 0dB speech and enhanced car 0dB speech, respectively. From frequency domain, result shows a good enhancement, noise at high frequency and some of noise at low frequency has been eliminated. However, it still has some noise, which shown in time domain. The SNR of original signal is 28.4dB. The SNR of enhanced speech is 39.1dB. It has a quite good enhancement despite of the fact that some noise has not been eliminated.





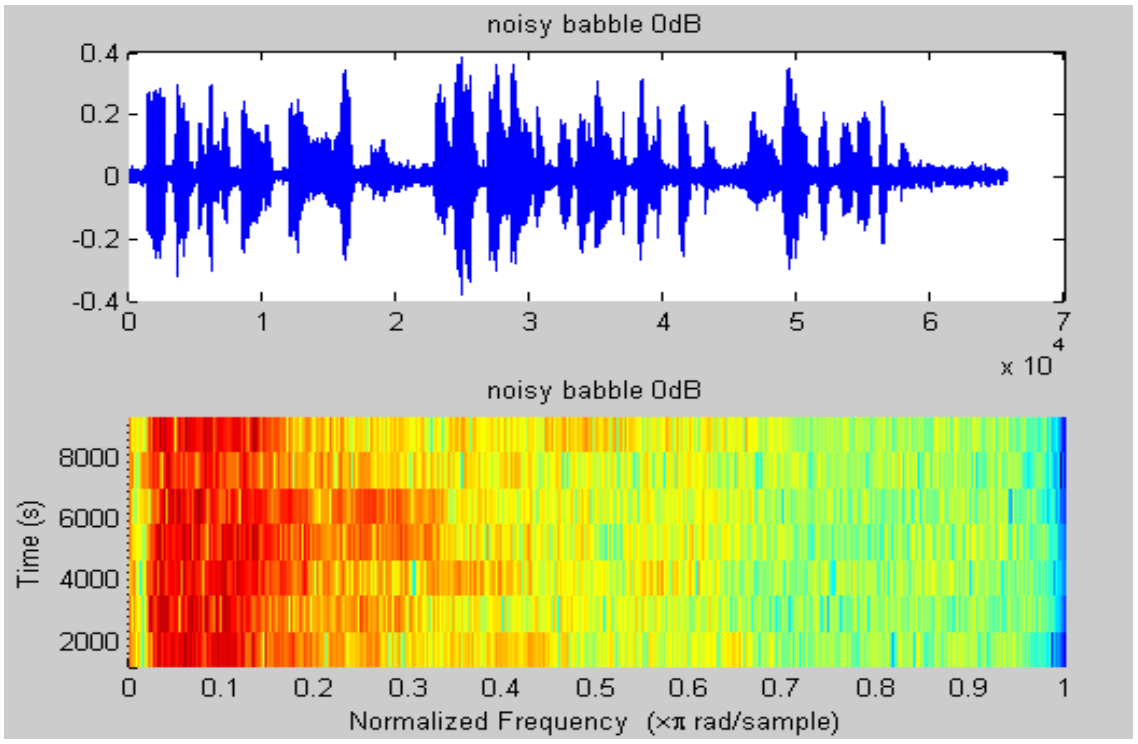
The result of car 5dB speech is a little better than that of car 0dB speech, which can be seen from time domain. In frequency domain, it shows a similar result with former one. The SNR of original signal is 33.02dB. The SNR of enhanced speech is 44.91dB.

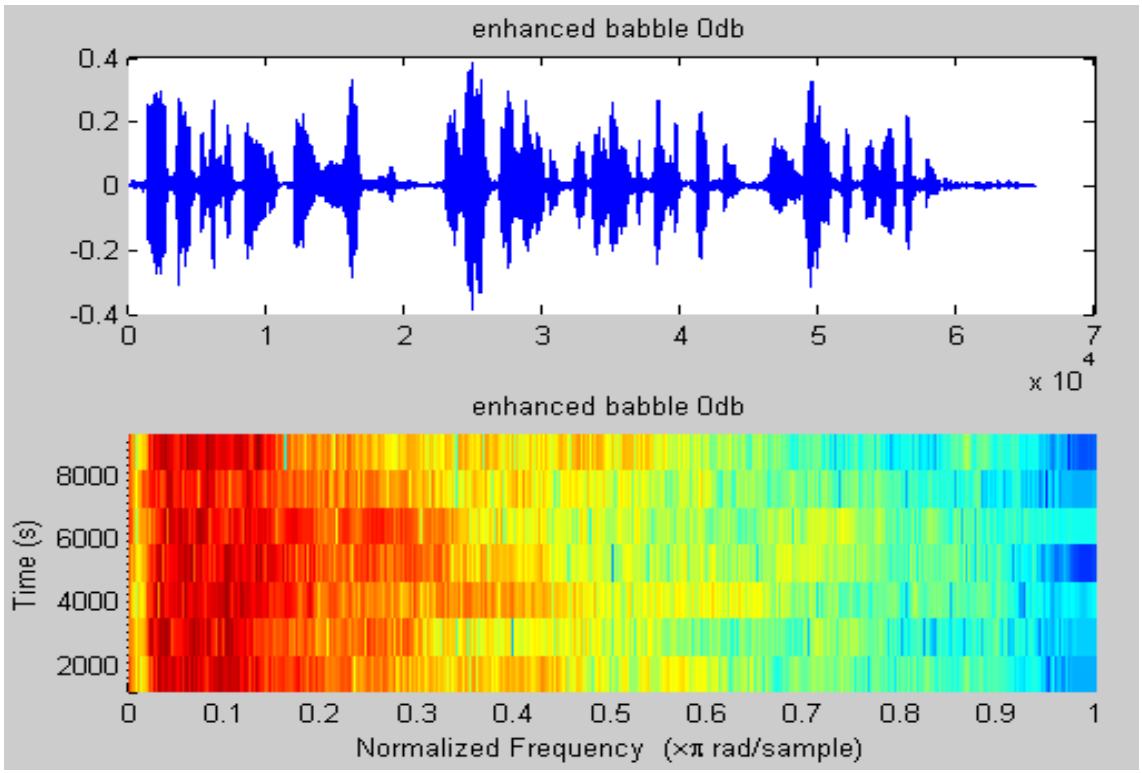




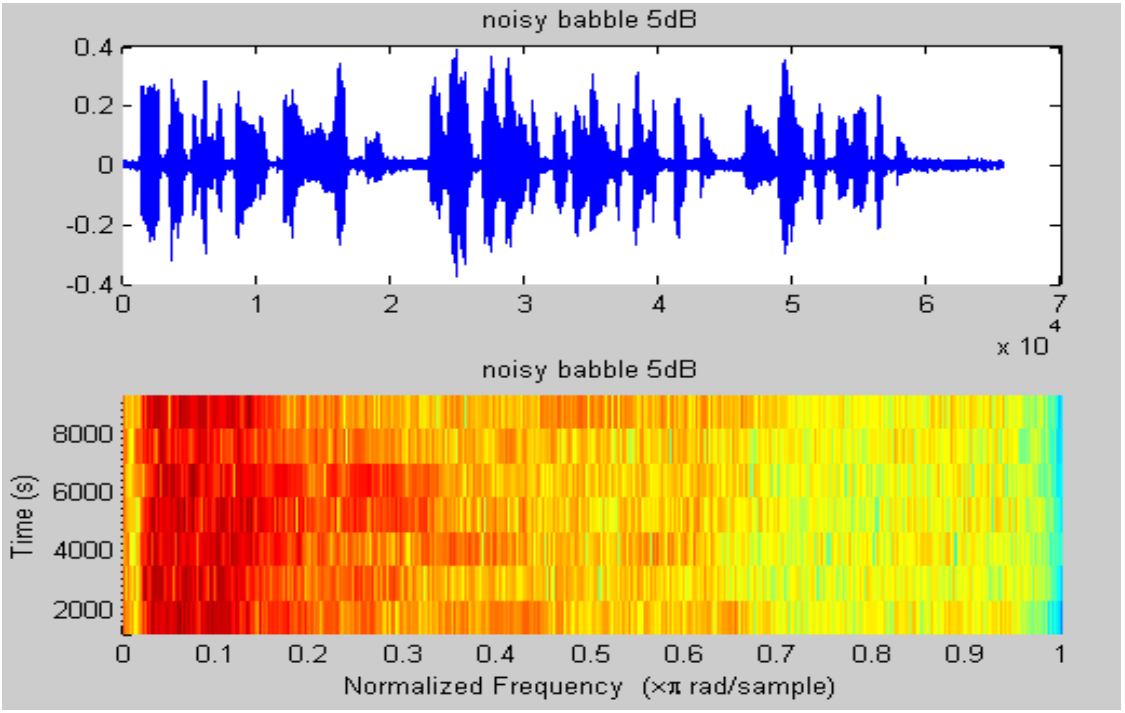
Similar with former comparison, in time domain, it shows a better result than those in former two speeches. It shows a very similar time-frequency graph with car 0dB and car 5dB speeches. The SNR of original car 10dB speech is 37.8dB. SNR of enhanced car 10dB speech is 50.6dB. The increases in SNR in three speeches are different, with approximate 10dB, around 11dB, and 12dB, respectively. It seems that when dealing with car noise, performance of enhancement will decrease while car noise level increase.

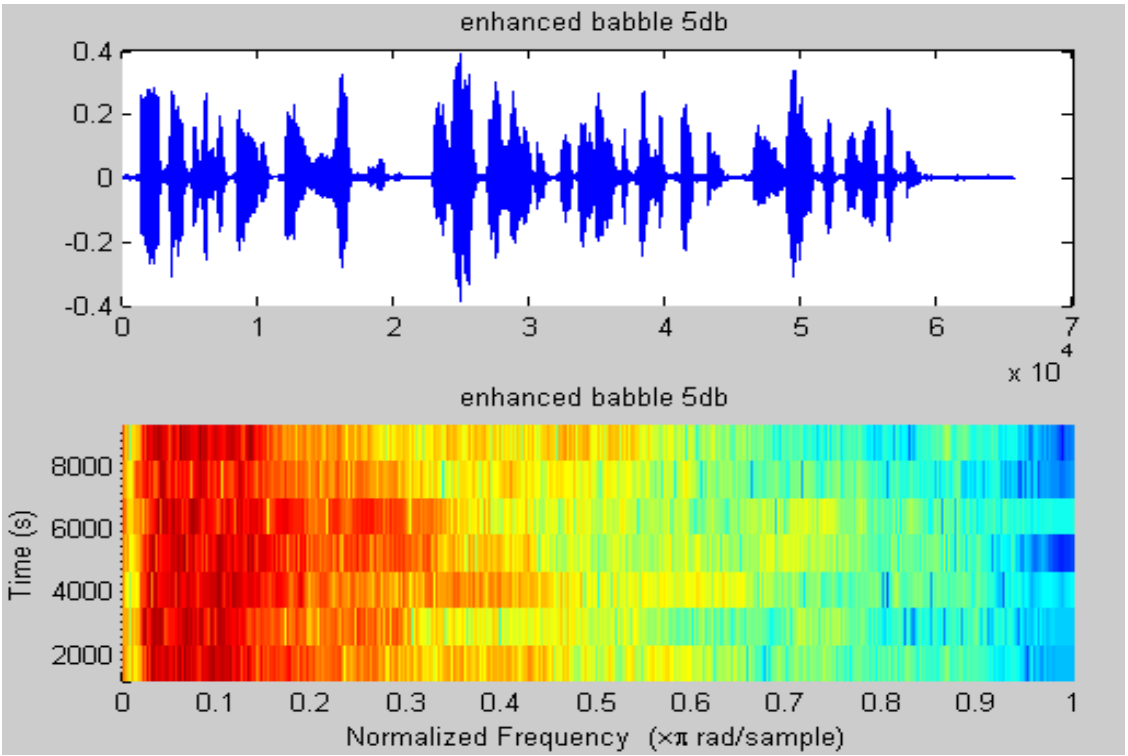
4.1.3 Babble noise



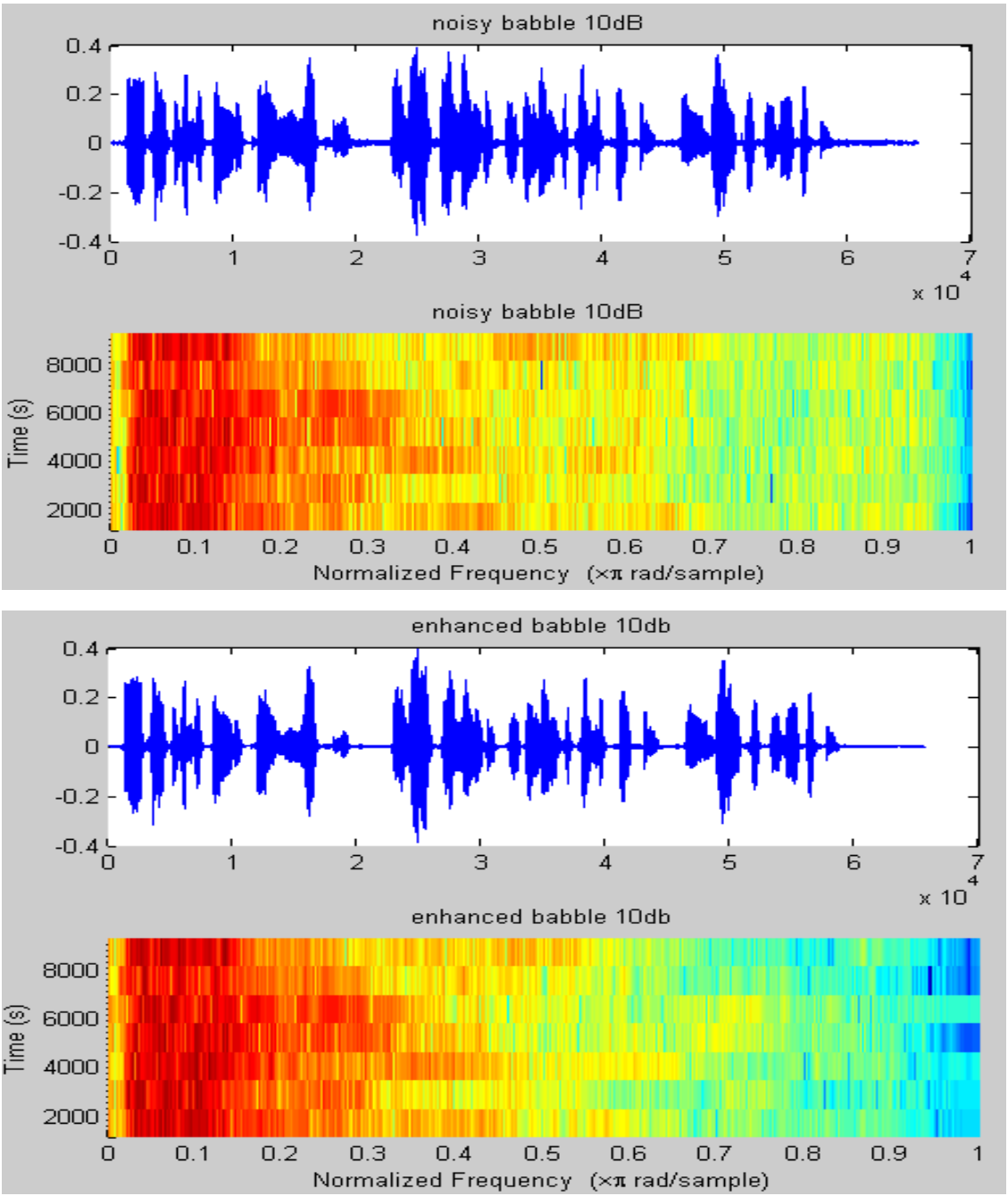


From the picture, it shows a good result in time domain. However, it only has a good performance at first half. At the other half performance of enhancement is limited. There is still some noise at high frequency which should be eliminated. Because of non-stationary noise, we do not calculate the SNR of this type of noise. When hearing the sound, it confirms this result, at former part the performance is good, at second half the performance is limited.





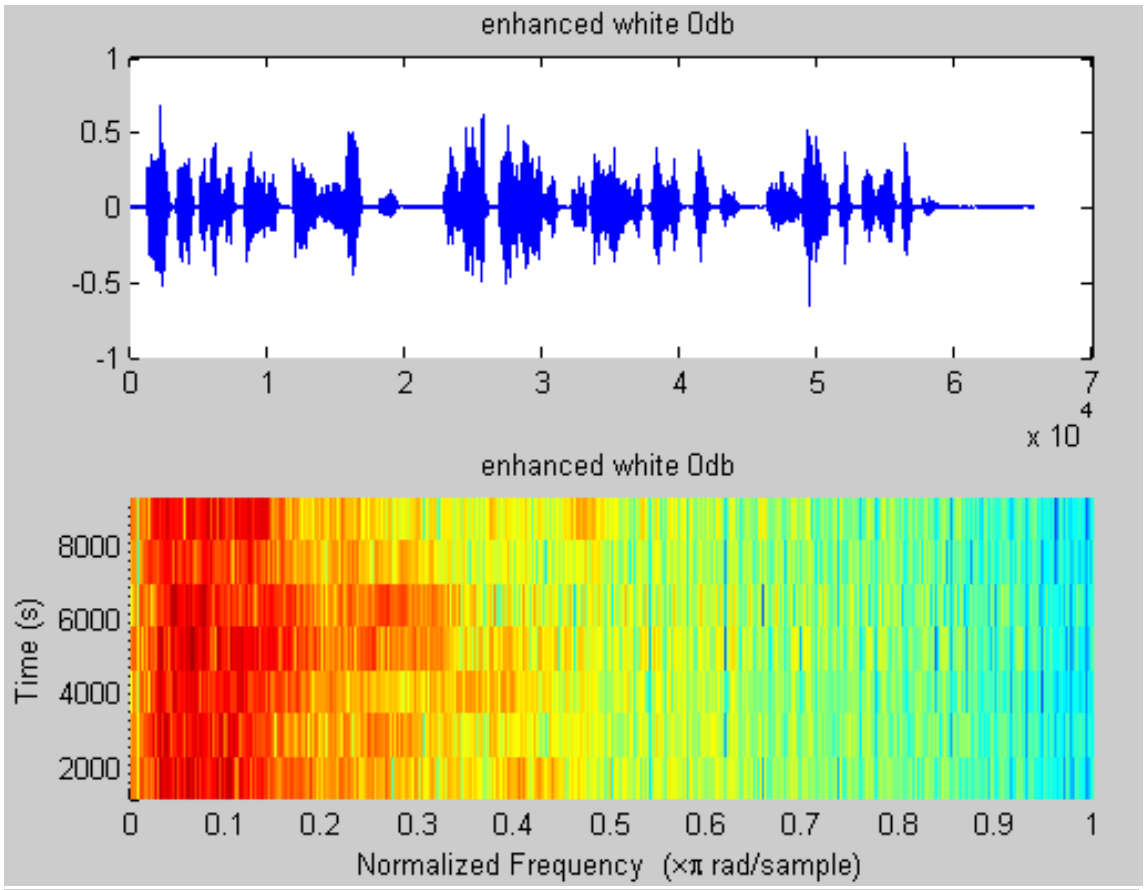
Similar with last one, it shows good performance in time domain. In time-frequency figure, there are some components at around 0.7π.frequency, which could be noise. However, when hearing sound, it seems has a better performance than the last one.

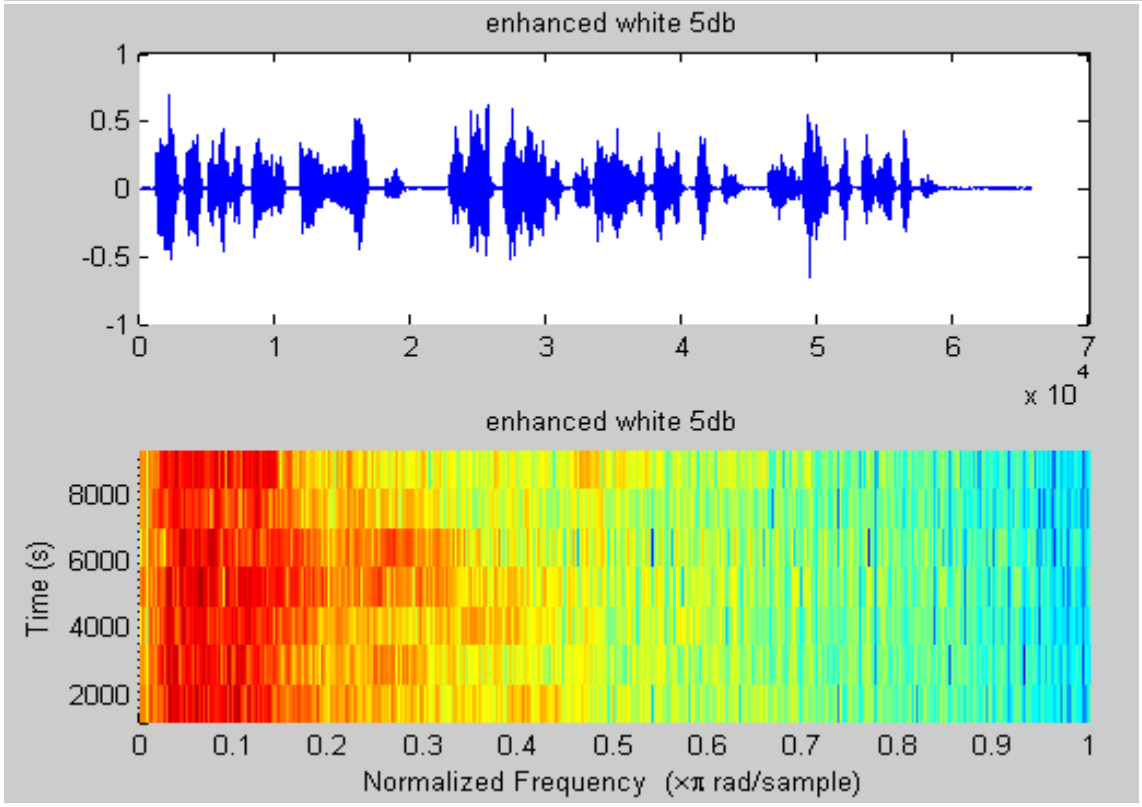


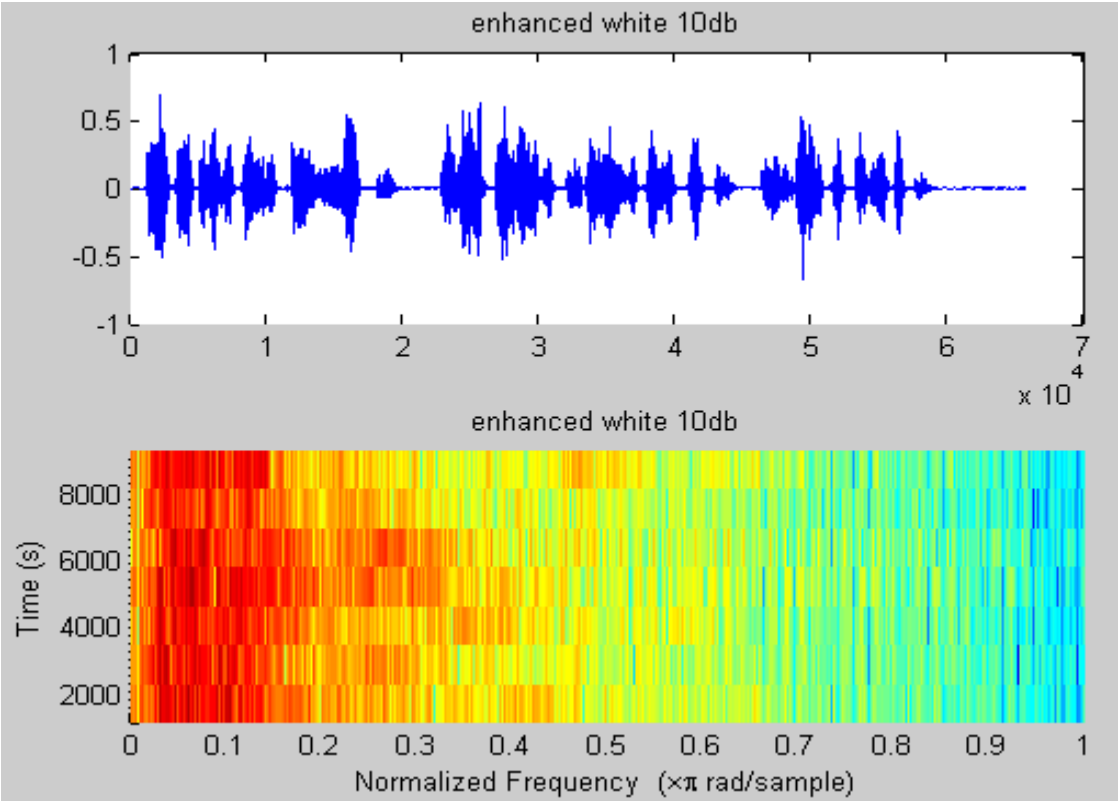
The performance of this one is much better, because of low level of noise. Although it is a non-stationary noise, the performance at both first and second half is quite good. The noise in enhanced speech can be barely heard.

4.2 Suppression of Acoustic Noise in speech

4.2.1 White noise



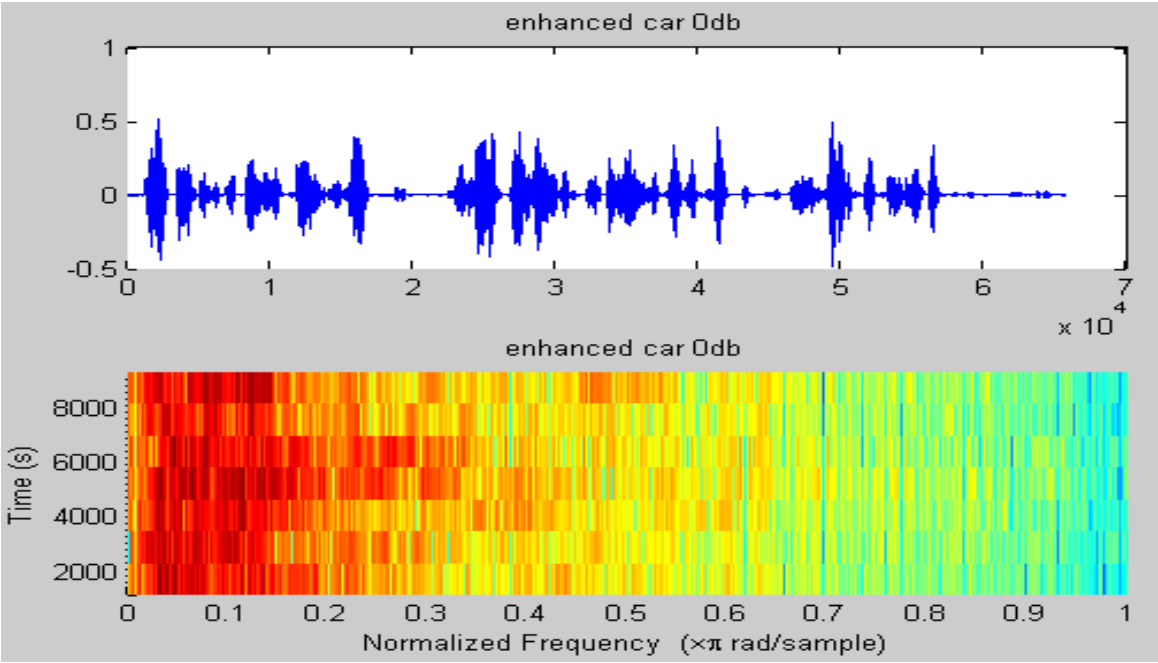


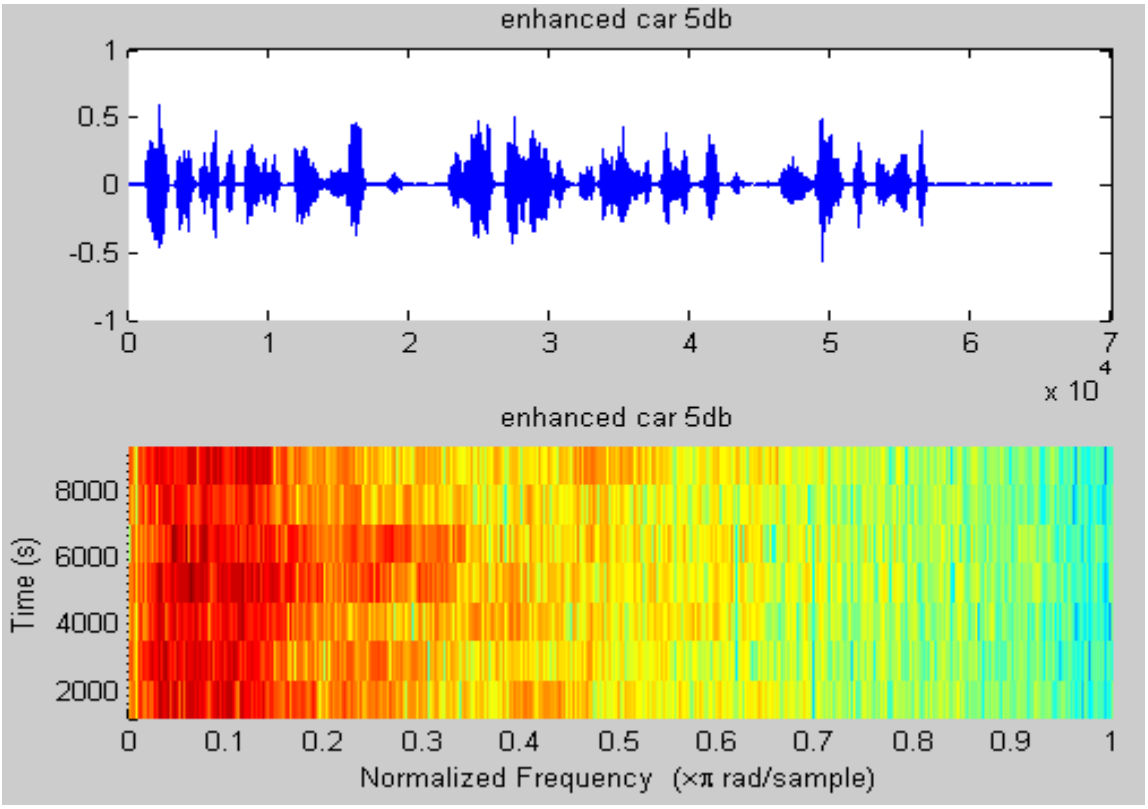


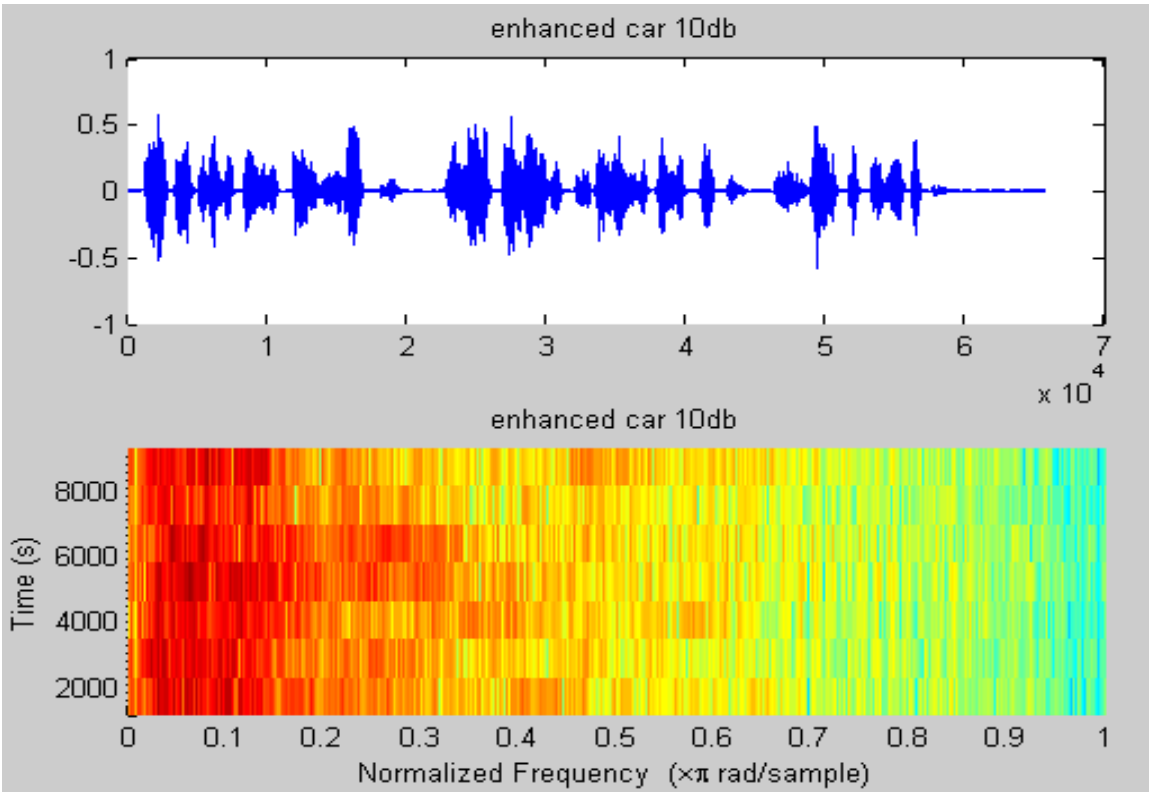
The SNRs of these speech are 64.73dB, 69.87dB,74.91dB, respectively. However, despite of high SNR, these all have some level of distortion, which can be seen from time domain.

The shape of wave has changed in three speeches. Moreover, compare with minimum statistic method, components in high frequency have higher values.

4.2.2 Car noise

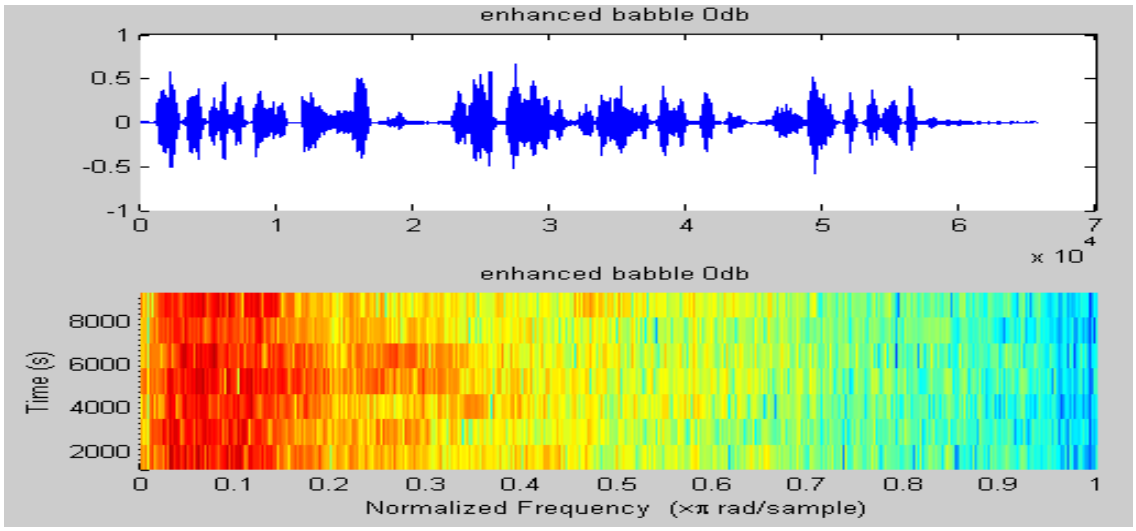


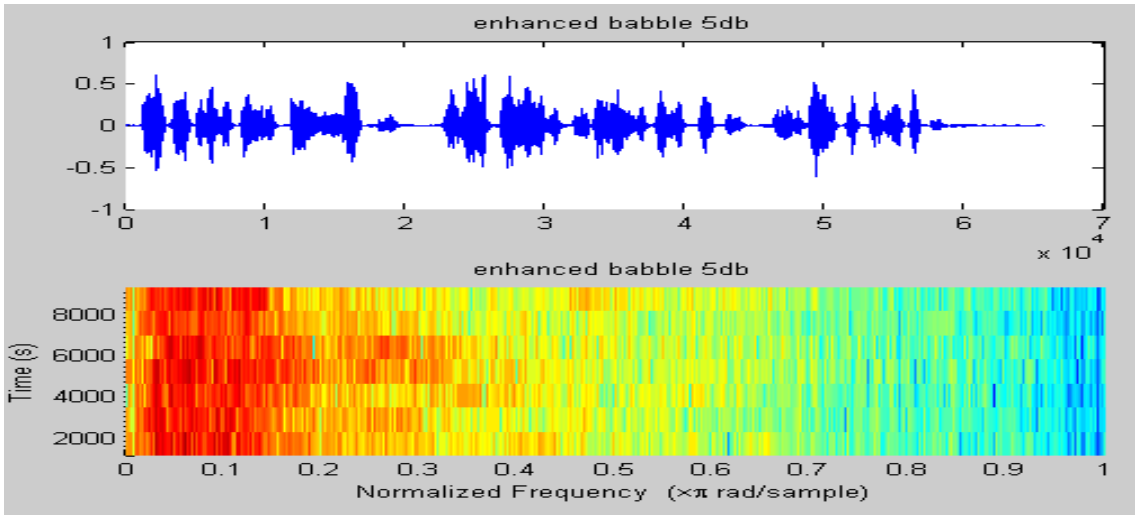


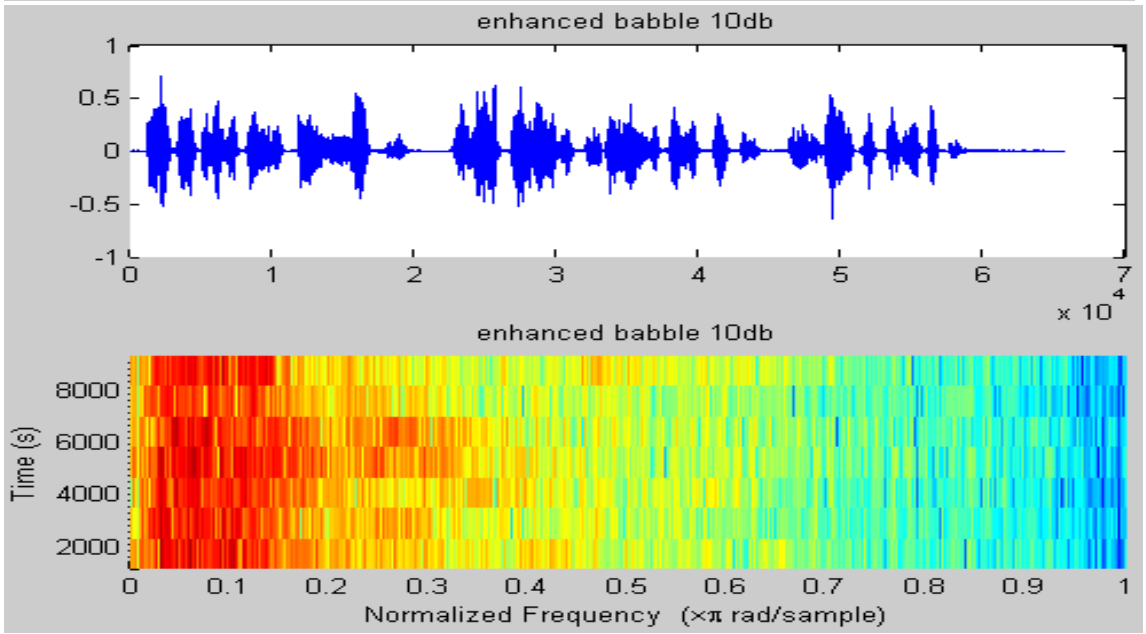


It seems have a better result than using minimum statistic method, with SNR 50.46dB, 56.89dB, 58.61dB, repectively. It has good performance at elimination the noise part. However, it causes distortion as well in speech part. Its performance will change when facing different level of car noise, similarly, performance will decrease with the increase of noise level.

4.2.3 Babble noise





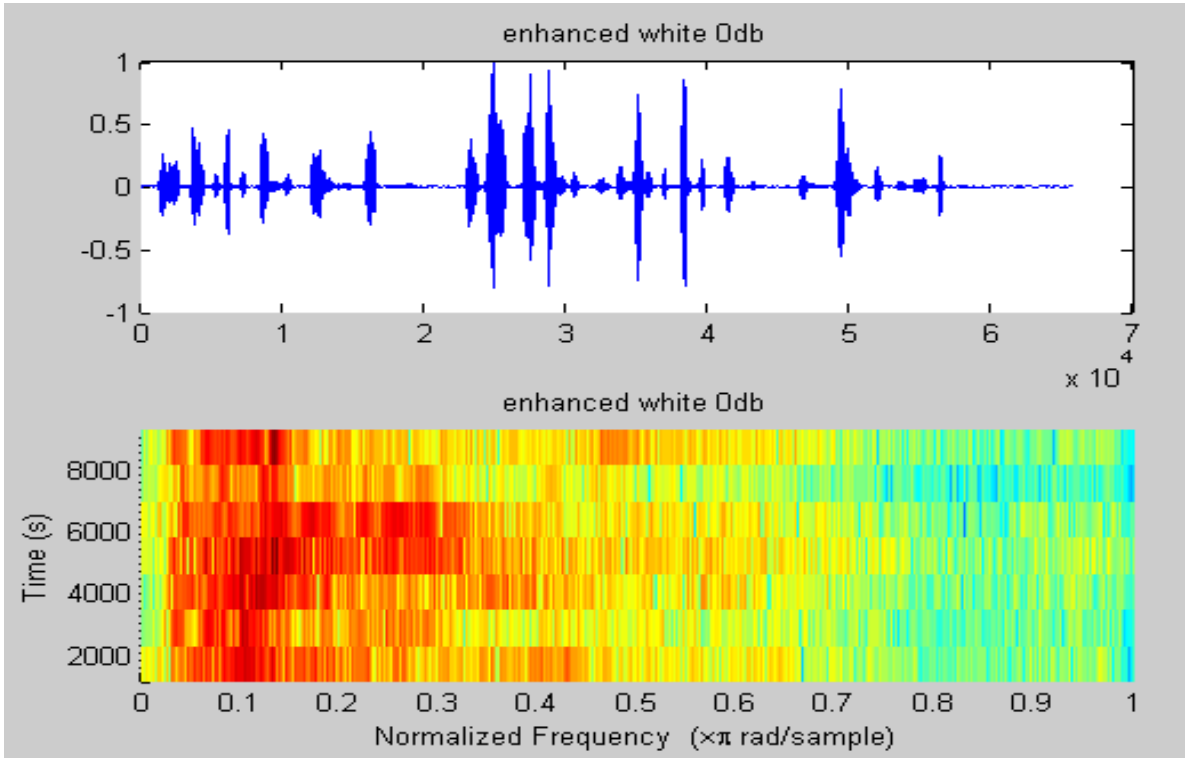


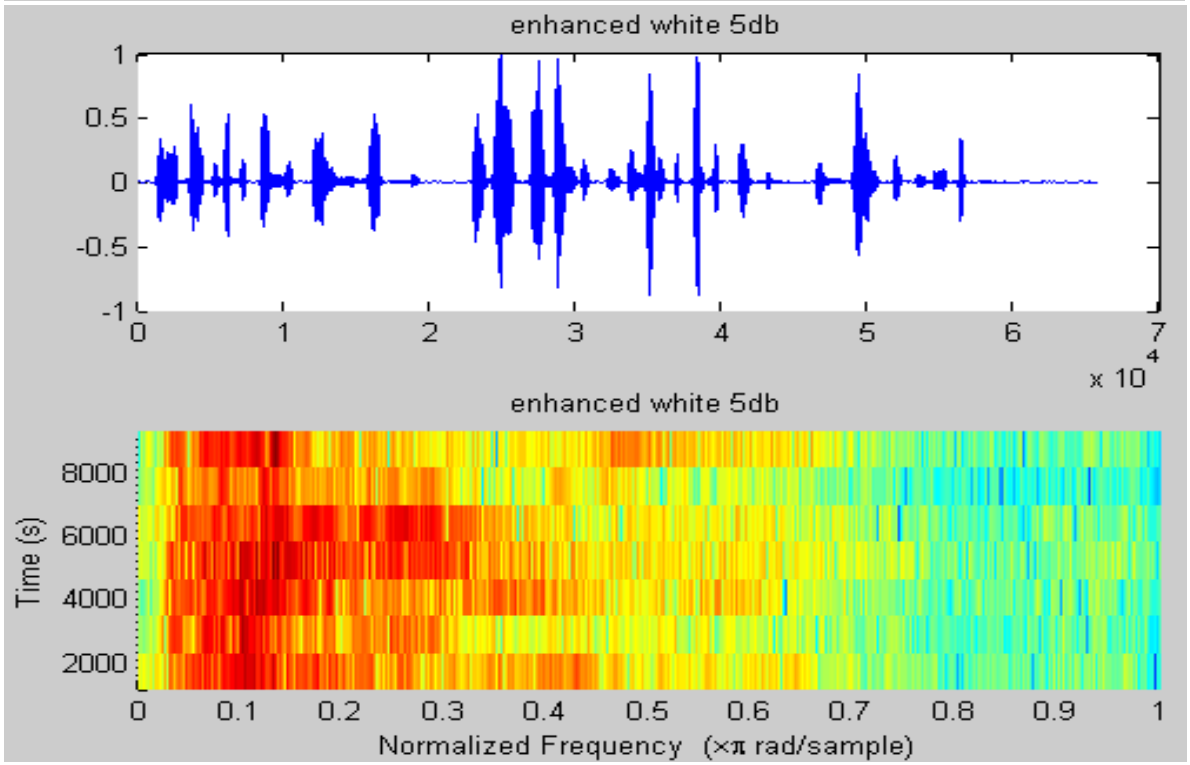
The performance changes as well, when facing the low level of babble noise, the noise level in enhanced speech is nearly eliminated, when facing the 0dB level of noise, the

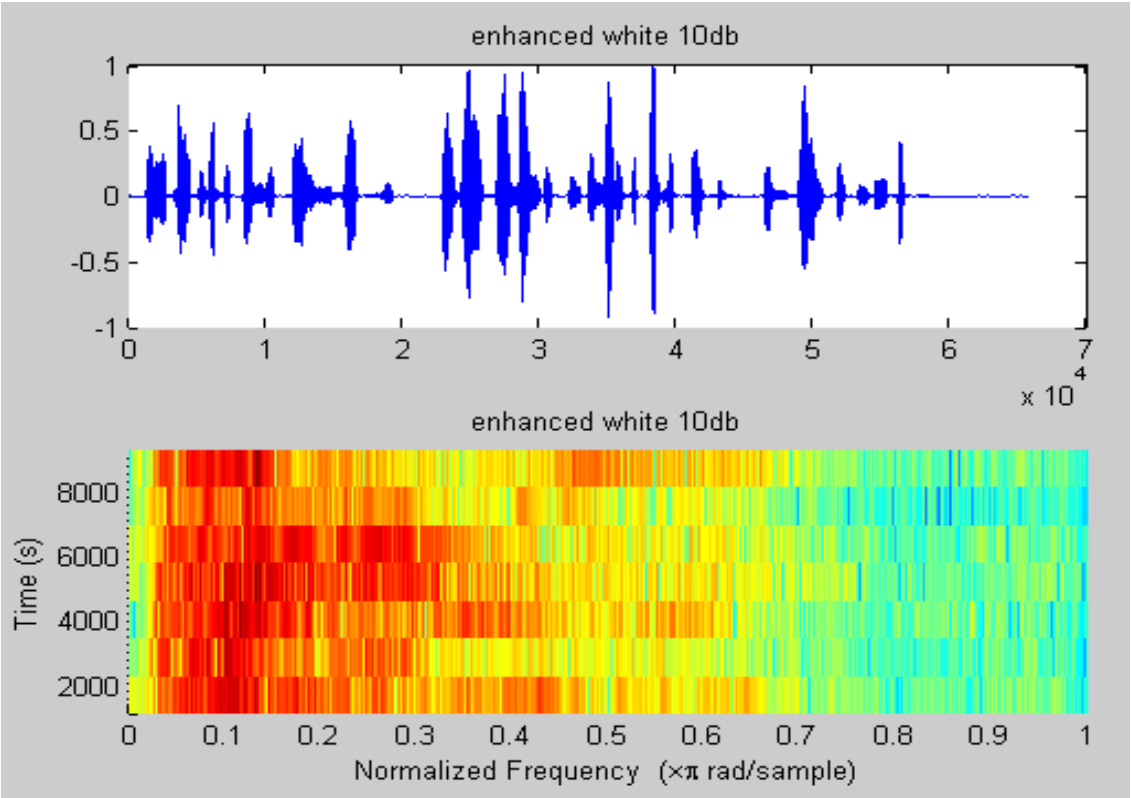
noise still can be heard at the half part. However, it does not show significant differences in both time domain and frequency domain.

4.3 Weiner Filter

4.3.1 White noise

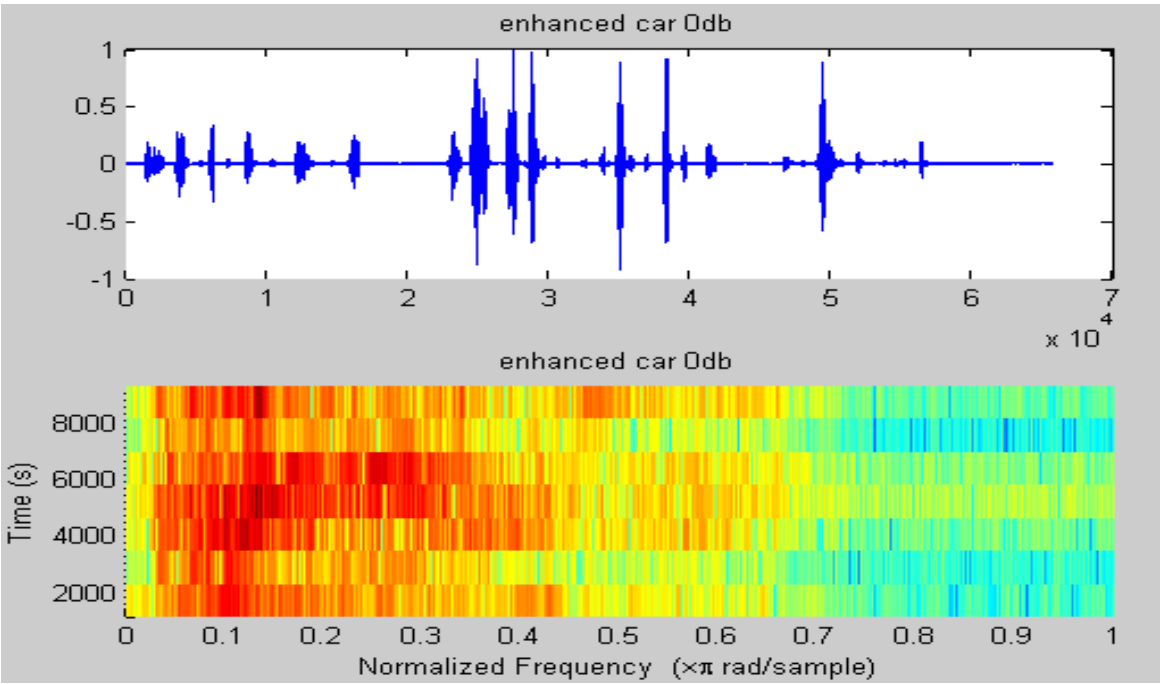


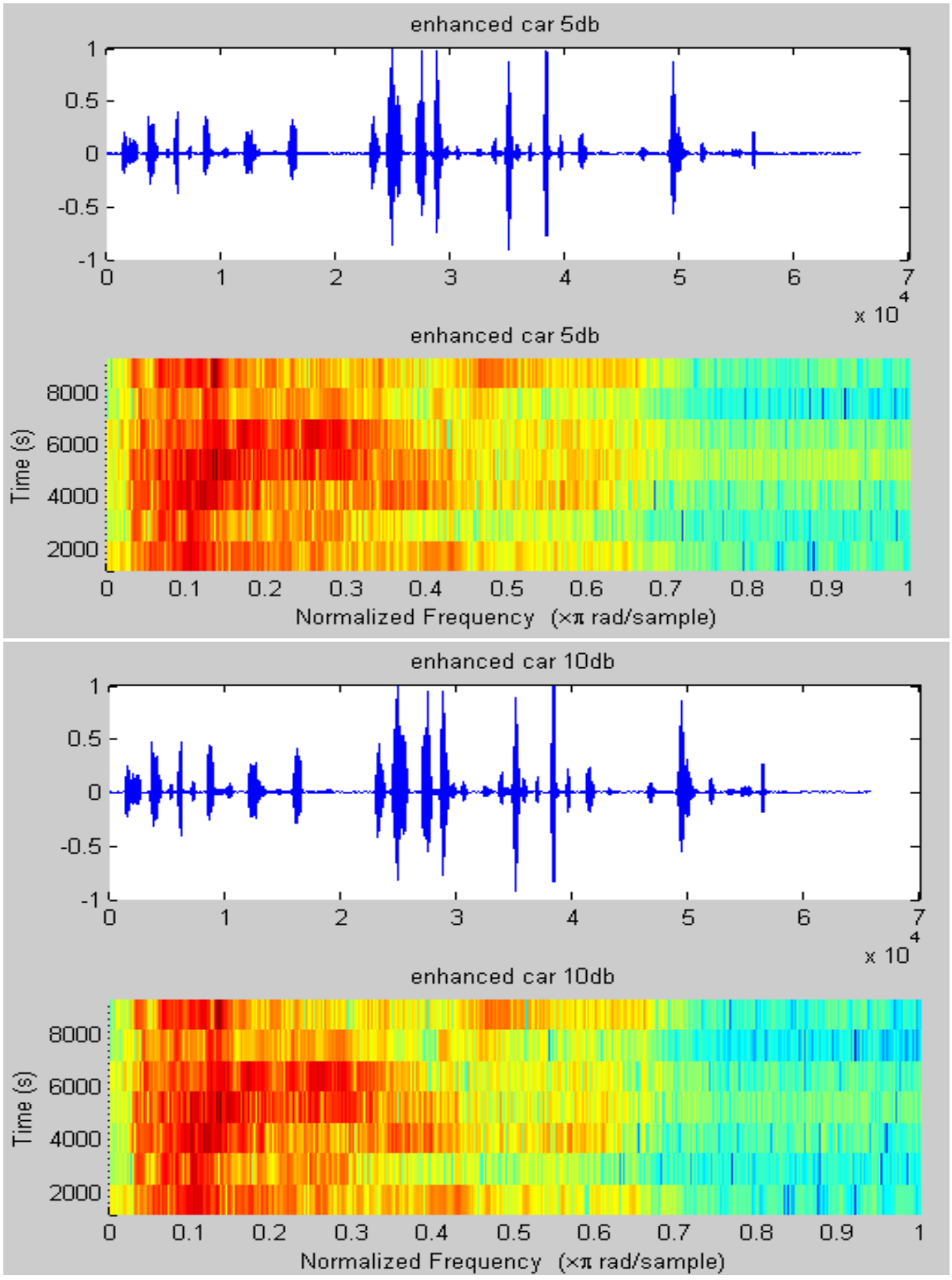




In this part, the result is quite different with other two method. Because of using normlisaton in code, the value becomes bigger than before. The SNR of three speeches are 81dB, 91dB, 97dB, respectively. The main difference among this method with other two method is that it has a good performance at low frequency. Value in each frame at low frequency is smaller speech in different level than that in original speech.

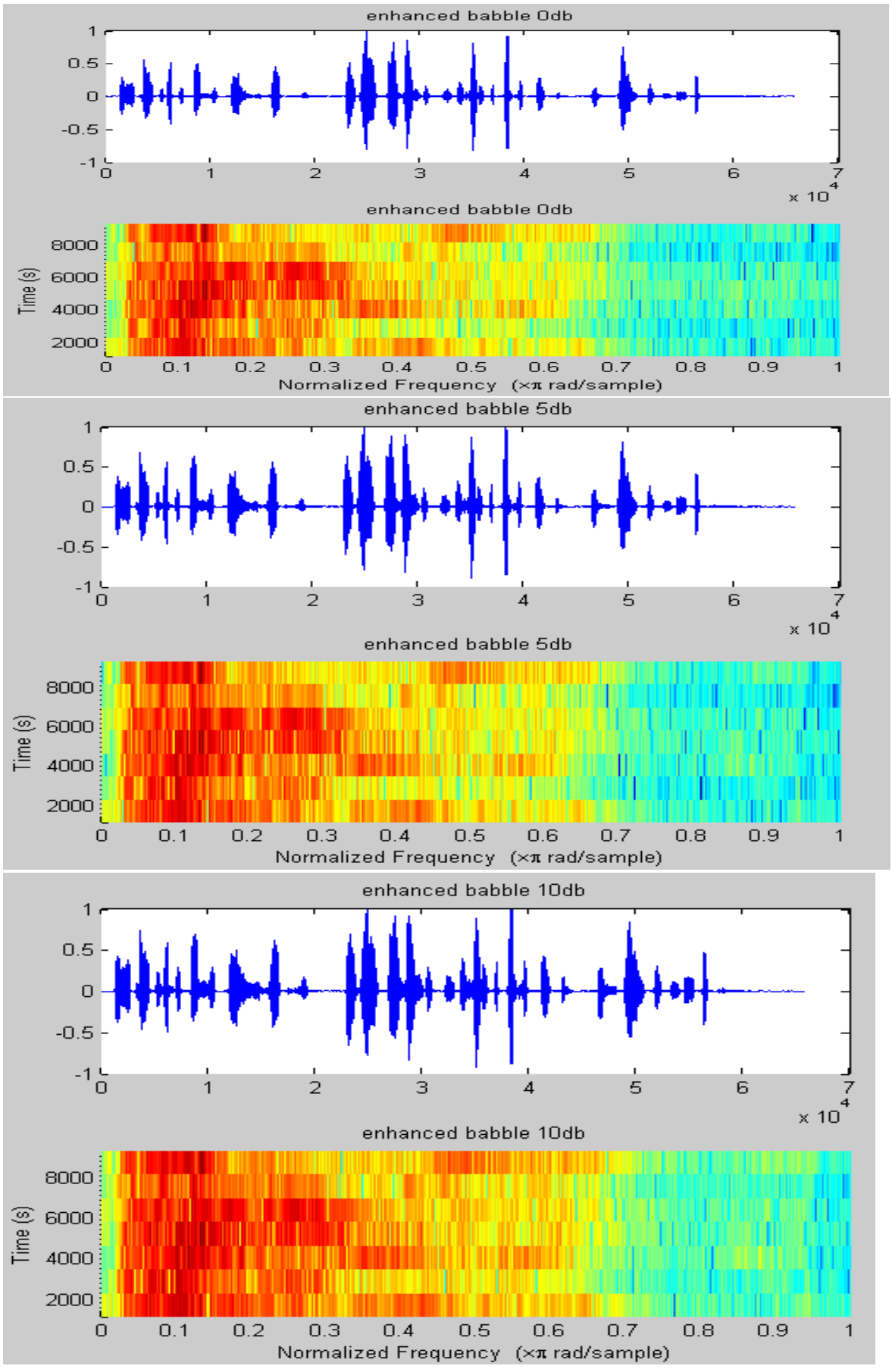
4.3.2 Car noise





SNRs are 58db,71db,82db.Although they looks similar in both time domain and frequency domain, the difference is siginificant. When compare 0db and 5db noise, value in 5db in each frame is bigger than that in 0db, which results in a significant difference when hearing them. Because of small value at some point in enhanced car 0db speech, some part cannot be hear clearly, while the same part in enhanced car 5dB speech can be hear clearly. It has similar result when comparing 5db speech and 10db speech. Weiner filter method will have a better proformance when level of car noise is lower.

4.3.3 Babble noise



The results show that it has a quite good performance when facing babble noise. Some noise can still be heard in enhanced babble 0dB speech at second half part, but in a small sound. The result of enhanced 10db babble nosie is quite similar with the clean speech, which shows a good performance. It is clear that Weiner filter method has a good performance when facing babble noise in different levels.

5.Analysis among different methods

Minimum statistic method has good performance on first two types of noise. All these results can be heard clearly, despite of the fact that some noise may not be eliminated but be weakened. This method may have a good performance at low level of non-stationary noise, however, when facing strong level of non-stationary noise the performance is limited. This may because of the method to determine the estimated noise power, when noise power changes rapidly this method cannot estimate the noise power correctly. And because of all the babble noise provided in file are become stronger at second half, the enhanced speech will have noise components at second half.

Suppression of Acoustic Noise method has good performance on eliminate the noise on stationary noise, however, may cause distortion in speech region. The distortion may result from misunderstood of the algorithm in the paper, or result from the method which used to determine the estimated noise. Based on the results, this method has the worst performance in three methods.

Weiner filter method can deal with both stationary and non-stationary noise, based on the results. However, when facing car noise which more concentrate on low frequency, the performance of using Weiner filter is quite limited because of it may eliminate some components in speech region.

6.Conclusion

In conclusion, minimum statistic method has advantage on stationary noise in both white noise and car noise in different level, and has good performance on low level of non-stationary noise. Suppression acoustic noise method has advantage on eliminate noise, but has the need for speech activity detector by exploiting the short time characteristics of speech signals, the distortion in this report may be biased because of misunderstood or incorrect method of estimated noise. Weiner filter method has advantage on non-stationary and white noise, but has limited performance on high level of car noise. In practice, an improved method based on minimum statistics method has been used, which using estimated power from D length window as the noise for the current frame and it will change with movement of the frame. However, the result only has a limited improvement, therefore, this method is abandoned.