National Aerospace

Class *Un*-Restricted

No. of Copies 6

Title: Studies in Signal Processing Techniques for Speech Enhancement: A comparative study

Author/s *M.Shivamurti, Dr.S.V.Narasimhan*

Division ALD

NAL Project No. SIP-04

Document No. PD AL 0919

Date of issue: August 2009

Contents 40 Pages 11 Figures 3 Tables 23 References

External Participation *Nil*

Sponsor *NAL*

Approval Head, ALD

Remarks

Key Words: **Spectral subtraction, Harmonic Wavelet Transform, Wiener Filter, MMSE**.

ABSTRACT: Speech enhancement is very essential to suppress the background noise and to increase speech intelligibility and reduce fatigue in hearing. There exist many simple speech enhancement algorithms like spectral subtraction to complex algorithms like Bayesian Magnitude estimators based on Minimum Mean Square Error (MMSE) and its variants. A continuous research is going and new algorithms are emerging to enhance speech signal recorded in the background of environment such as industries, vehicles and aircraft cockpit. In aviation industries speech enhancement plays a vital role to bring crucial information from pilot's conversation in case of an incident or accident by suppressing engine and other cockpit instrument noises. In this work proposed is a new approach to speech enhancement making use harmonic wavelet transform and Bayesian estimators. The performance indicators, SNR and listening confirms to the fact that newly modified algorithms using harmonic wavelet transform indeed show better results than currently existing methods. Further, the Harmonic Wavelet Transform is computationally efficient and simple to implement due to its inbuilt decimation-interpolation operations compared to those of filter-bank approach to realize sub-bands.

Contents

Abstract

Speech enhancement is very essential to suppress the background noise and to increase speech intelligibility and reduce fatigue in hearing. There exist many simple speech enhancement algorithms like spectral subtraction to complex algorithms like Bayesian Magnitude estimators based on Minimum Mean Square Error (MMSE) and its variants. A continuous research is going and new algorithms are emerging to enhance speech signal recorded in the background of environment such as industries, vehicles and aircraft cockpit. In aviation industries speech enhancement plays a vital role to bring crucial information from pilot's conversation in case of an incident or accident by suppressing engine and other cockpit instrument noises. In this work proposed is a new approach to speech enhancement making use harmonic wavelet transform and Bayesian estimators. The performance indicators, SNR and listening confirms to the fact that newly modified algorithms using harmonic wavelet transform indeed show better results than currently existing methods. Further, the Harmonic Wavelet Transform is computationally efficient and simple to implement due to its inbuilt decimation-interpolation operations compared to those of filter-bank approach to realize sub-bands.

Key Words: Spectral subtraction, Harmonic Wavelet Transform, Wiener Filter, MMSE.

Introduction:

Speech enhancement is concerned with improving some perceptual aspects of speech that has been degraded by additive noise. In most applications, the aim of speech enhancement is to improve the quality and intelligibility of degraded speech. The improvement in quality is highly desirable as it can reduce listener fatigue, particularly in situations in which the listener is exposed to high levels of noise for long period's time (e.g. manufacturing). Speech enhancement algorithms reduce or suppress the background noise to some degree and are sometimes referred to as noise suppression algorithms.

Speech enhancement is concerned with the processing of corrupted or noisy speech in order to improve the quality or intelligibility of the signal. Applications range from front-ends for speech recognition systems, to enhancement of telecommunications for aviation, military, teleconferencing, and cellular environments.

In airplane communication system, the interference of aviation noise makes speech enhancement necessary. As amplitude of aviation noise is very strong, it can do great harm to audition Further it may also affect speech coding and impair speech quality. Strong noise produced by the airplane engine may seriously affect the performance of airborne communication system. Under such circumstances, enhancement of SNR, especially improving the articulation and intelligibility of speech is very important. The aviation noise is one kind of wide-band noise. Because this noise and the speech signal overlap strongly in the frequency range, the traditional method to eliminate the noise doesn't work well. Now there are many ways for speech enhancement, but most of them have their flaws [1].

In order to realize these algorithms in real-time applications, their efficient implementation in terms of computational load, simplicity and performance is of main concern. In this direction sub band approach has been used over decades to meet the real-time specific requirements.

In the multi-rate scenario, wavelet transform is an improvement over Short time Fourier transform (STFT), as it enables good time localization in detecting fast events like transients and good frequency resolution for low frequency slow processes. The speech enhancement in multi-rate domain using spectral Subtraction and MMSE show that a non-uniform frequency resolution like in DWT does improve the quality of speech. However, the different sampling rates for the subband complicate the selection of smoothing factor which becomes function of subband sampling frequencies and requires experimental effort to fix constants used for speech uncertainty detection in order to reduce background and musical noise [2]. The nature of algorithms used in multi-rate processing should not complicate the subband process in realizing a simple, computationally efficient speech enhancement method.

The continuous wavelet transform (CWT) is basically a correlation of the signal with a wavelet of appropriate scale at desired shifts or translations. This provides the shift invariant nature in a limited sense as the scale and the shifts can be selected. However, from the implementation point of view compared to CWT, the discrete wavelet transform (DWT) realized by a dyadic structure using a perfect reconstruction filter bank, is generally used due to its computational efficiency. In DWT, the WT coefficients are computed at predetermined translations and scales. Though this is computationally efficient, in statistical applications, denoising, signal analysis, pattern recognition, WT computed at

smaller translations similar to CWT is preferred. This is due to the fact that the time resolution of DWT is very coarse and a better one is desirable especially in detecting time of occurrence of an event/ transient in individual higher scales.

The discrete wavelet transform may be used as a signal-processing tool for visualization and analysis of non-stationary, time-sampled waveforms. The highly desirable property of shift invariance can be obtained at the cost of a moderate increase in computational complexity, and accepting a least-squares inverse (pseudo-inverse) in place of a true inverse. A new algorithm for the pseudo-inverse of the shift-invariant transform that is easier to implement in array-oriented scripting languages than existing algorithm was presented [3] together with self-contained proofs. Its application to speech preserved original pitch and formant frequencies also informal listening tests found clear and understandable.

The Harmonic wavelet transform (HWT) is attractive from computational point of view, as it has built in decimation and easier interpolation operation. This is based on grouping of DFT coefficients in a dyadic fashion. The inverse transform of each group, gives the WT coefficients for that scale. For reconstruction, these groups can be concatenated to get the complete FT and its inverse FT gives the signal. The DFTHWT and DCTHWT have been explored for speech enhancement in conjunction with MMSE [4].

Though the DFTHWT is computationally efficient, it suffers from the problems of DFT, mainly the leakage and complex WT coefficients due to lack of DFT symmetry in the grouping process. This further limits processing wavelet coefficients, which are complex. These are solved by the discrete cosine harmonic wavelet transform (DCHWT). The symmetrical signal extension effectively reduces the leakage. Further, the DCT being real and its built in symmetry, the WT coefficients are assured to be real.

2. Speech algorithms descriptions:

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 assumption made is that noise is stationary or slowly varying process, and that the noise spectrum does not change significantly between the updating periods. The enhanced signal is obtained by computing the inverse discrete Fourier transform of the estimated signal spectrum using the phase of the noisy signal.

2.1. MAGNITUDE SPECTRAL SUBTRACTION (MSS)

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

$$y(n) = x(n) + d(n) \tag{1}$$

Taking the discrete-time Fourier transform of both sides gives

$$Y(\omega) = X(\omega) + D(\omega) \tag{2}$$

We can express $Y(\omega) = |Y(\omega)|e^{j\theta_y(\omega)}$ (3)

Where $|Y(\omega)|$ is the magnitude spectrum and $\theta_y(\omega)$ is the phase of the corrupted noisy signal. The noise spectrum $D(\omega)$ can also be expressed in terms of its magnitude and phase spectra as $D(\omega) = |D(\omega)|e^{j\theta_d(\omega)}$. The magnitude noise spectrum $|D(\omega)|$ is unknown but can be replaced by its average value computed during nonspeech activity. Similarly, the noise phase $\theta_d(\omega)$ can be replaced by the noisy speech phase $\theta_y(\omega)$. This is partly motivated by the fact that phase that does not affect speech intelligibility may affect speech quality to some degree. After making these substitutions to eqn. (2), we can obtain an estimate of the clean signal spectrum:

$$\hat{X}(\omega) = \left[|Y(\omega)| - \left|\widehat{D}(\omega)\right|\right]e^{j\theta_y(\omega)} \tag{4}$$

Note that the magnitude spectrum of the enhanced signal $\left|\hat{X}(\omega)\right|$ can be negative owing to inaccuracies in estimating the noise spectrum. The magnitude spectra, however, cannot be negative. One solution to this is to half-wave-rectify the difference spectra i.e., set the negative spectral components to zero as follows:

$$\left|\hat{X}(\omega)\right| = \begin{cases} |Y(\omega)| - \left|\widehat{D}(\omega)\right| & if\ |Y(\omega)| > \left|\widehat{D}(\omega)\right| \\ 0 & else \end{cases} \tag{5}$$

2.2. Power Spectral Subtraction (PSS):

The preceding derivation of the magnitude spectral subtraction algorithm can be easily extended to the power spectrum domain. In some cases, it might be best to work with power spectra rather than magnitude spectra. To obtain the short-time power spectrum of the noisy speech, we multiply $Y(\omega)$ in eqn. (2) by its conjugate$Y^*(\omega)$. In doing so, eqn. (2) becomes:

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega)|^2 + |D(\omega)|^2 + X(\omega)\cdot D^*(\omega) + X^*(\omega)\cdot D(\omega) \\ &= |X(\omega)|^2 + |D(\omega)|^2 + 2Re\{X(\omega)D^*(\omega)\} \end{aligned} \tag{6}$$

The terms $|D(\omega)|^2$, $X(\omega)\cdot D^*(\omega)$, and $X^*(\omega)\cdot D(\omega)$ cannot be obtained directly and are approximated as $E[|D(\omega)|^2]$, $E[|X^*(\omega)\cdot D(\omega)|]$, and $E[|X(\omega)\cdot D^*(\omega)|]$, where $E[\cdot]$ denotes the expectation operator. Typically, $E[|D(\omega)|^2]$ is estimated during nonspeech activity and is denoted by $\left|\widehat{D}(\omega)\right|^2$. If we assume that $d(n)$ is zeros mean and uncorrelated with the clean signal $x(n)$, then the terms $E[|X^*(\omega)\cdot D(\omega)|]$ and $E[|X(\omega)\cdot D^*(\omega)|]$ reduce to zero. Thus, after using the preceding assumptions, the estimate of the clean speech power spectrum can be obtained as follows:

$$\left|\hat{X}(\omega)\right|^2 = |Y(\omega)|^2 - \left|\widehat{D}(\omega)\right|^2 \tag{(7}$$

The preceding equation describes the power spectrum subtraction algorithm. As before, the estimated power spectrum in eqn. (7) is not guaranteed to be positive, but can be half-wave rectified as shown in eqn. (5). The enhanced signal is finally obtained by computing the inverse Fourier transform of $\left|\hat{X}(\omega)\right|$ and adding phase of the noisy speech signal. Note that if we take the inverse Fourier transform of both sides in eqn. (7) we get similar equation in the autocorrelation domain, i.e.,

$$r_{\hat{x}x}(m) = r_{\hat{y}y}(m) - r_{\hat{d}d}(m) \tag{8}$$

Where $r_{\hat{x}x}(m)$, $r_{\hat{y}y}(m)$, and $r_{\hat{d}d}(m)$ are the autocorrelation sequences of the estimated clean signal, the noisy speech signal, and the estimated noise signals, respectively. Hence, the subtraction could in principle be performed in the autocorrelation domain.

Equation (7) can also be written in the following form:

$$\left|\hat{X}(\omega)\right|^2 = H^2(\omega)|Y(\omega)|^2 \tag{9}$$

Where

$$H(\omega) = \sqrt{1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}} \tag{10}$$

In the context of linear systems theory, $H(\omega)$ is known as the system's transfer function. In speech enhancement, we refer to $H(\omega)$ as the gain function, or suppression function. Note that $H(\omega)$ in equation is real, a zero phase filter and in principle, is always positive, taking values in the range of $0 \leq H(\omega) \leq 1$. Negative values are sometimes obtained owing to inaccurate estimates of the noise spectrum. $H(\omega)$ is called the suppression function or SNR-dependent attenuator because it provides the amount of suppression applied to the noisy power spectrum $|Y(\omega)|^2$ at a given frequency to obtain the enhanced power spectrum $\left|\hat{X}(\omega)\right|^2$. The attenuation at each frequency increases with the decreasing SNR, and conversely decreases with the increasing SNR eqn. (10).

A more generalized version of the spectral subtraction algorithm is given by

$$\left|\hat{X}(\omega)\right|^p = |Y(\omega)|^p - \left|\hat{D}(\omega)\right|^p \tag{11}$$

Where *p* is the power exponent, with p=1 yielding the original magnitude spectral subtraction [5], and p=2 yielding the power spectral subtraction algorithm. The general form of the spectral subtraction algorithm is shown in figure (1).

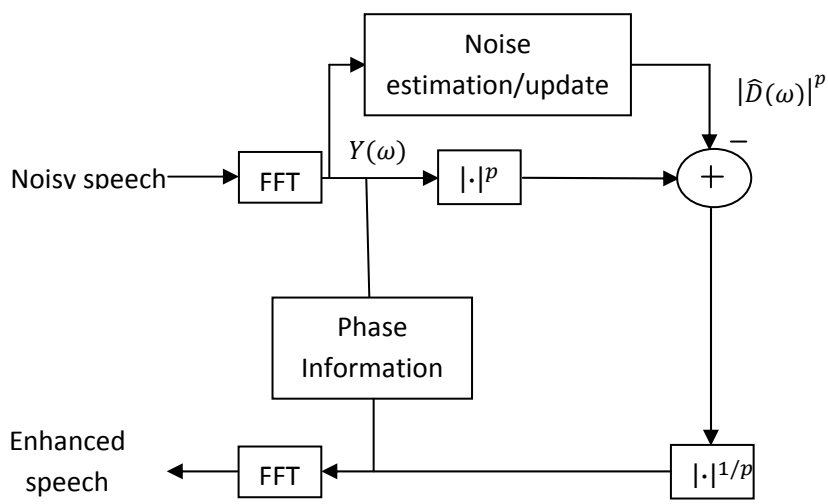


Fig.1. General Form of the spectral subtraction

2.3. Wiener Filtering (WF):

The spectral-subtractive algorithms are based largely on intuitive and heuristically based principles. More specifically, these algorithms exploited the fact that noise is additive and one can obtain an estimate of the clean signal spectrum simply by subtracting the noise spectrum from the noisy speech spectrum. The enhanced signal spectrum was not derived in an optimal way. Considering the statistical filtering problem, Fig.2. The input signal goes through a linear and time-invariant system to produce an output signal$y(n)$. We are to design the system in such a way that the output signal $\hat{d}(n)$ is as close to the desired signal,$d(n)$ as possible. This can be done by computing the estimation error, $e(n)$, and making it as small as possible. The optimal filter that minimizes the estimation error is called the *wiener filter*, named after the author.

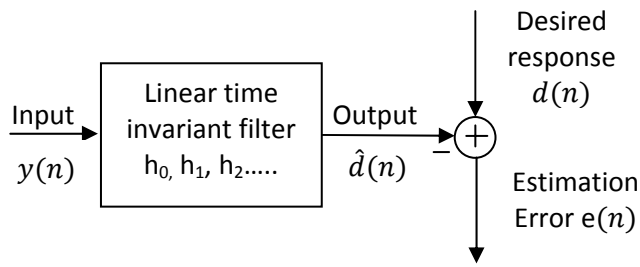


Fig.2. Block diagram of the statistical filtering problem

The mean square of the estimation error is commonly used as a criterion for minimization, and the optimal filter coefficients can be derived in the time or frequency domain. In this work frequency domain is considered.

2.3.1. Wiener filters in the frequency domain:

Considering eqn. (1) and (2) and assuming a filter is used to get an output denoted by $\hat{\mathrm{X}}(\mathrm{k})$,

$$\hat{\mathrm{X}}(\mathrm{k}) = \mathrm{H}(\mathrm{k})\mathrm{Y}(\mathrm{k}) \tag{12}$$

The objective is to obtain an expression for $\mathrm{W}(\mathrm{k})$ which minimizes the least mean-square error, E_m, defined as follows:

$$E_m = E\left[\left(\hat{X}(k) - X(k)\right)^2\right]$$

$$E_m = E[(H(k)Y(k) - X(k))^2] \tag{13}$$

Expanding eqn. (13), we get

$$= H(k)^2 E[Y(k)^2] + E[X(k)^2] - 2H(k)E[Y(k)]E[X(k)] \tag{14}$$

Eqn. (2) in eqn. (14), we have

$$= H(k)^2 E[X(k) + D(k)^2] + E[X(k)^2] - 2H(k)E[X(k) + D(k)]E[X(k)]$$

Now setting derivative $\frac{dE_m}{dW(k)}$ in eqn. (14) to zero, we get well known wiener filter as follows, assuming $E[X(k)D(k)] = 0$.

$$H(k) = \frac{\xi_k}{\xi_k + 1} \tag{15}$$

where $\xi_k = \frac{E[X(k)^2]}{E[D(k)^2]}$

2.4. Statistical-Model-Based Methods:

In the previous section we described the wiener filter approach to speech enhancement. This approach derives in the mean-square sense the optimal complex discrete Fourier transform (DFT) coefficients of the clean signal. The wiener filter approach yields a linear estimator of the complex spectrum of the signal and is optimal in the MMSE sense when both the noise and speech DFT coefficients are assumed to be independent Gaussian random variables.

In this section, nonlinear estimators of magnitude rather than the complex spectrum of the signal (as done by the wiener filter), using various statistical models and optimization criteria. These nonlinear estimators take the probability density function (PDF) of the noise and the speech DFY coefficients explicitly into account and use, in some cases, non-Gaussian prior distributions. These estimators are often combined with soft-decision gain modifications that take the probability of speech presence into account.

The speech enhancement problem is posed in a statistical estimation frame-work [4]. Given a set of measurements that depend on an unknown parameter, we wish to find a nonlinear estimator of the parameter of interest. In our application, the measurements correspond to the set of DFT coefficients of the noisy signal and the parameters of interest are the set of DFT coefficients of the clean signal. Various techniques exist in the estimation theory literature for deriving these nonlinear estimators and include the maximum-likelihood estimators and the Bayesian estimators. These estimators differ primarily in the assumptions made about the parameter of interest (e.g. deterministic but unknown, random) and the form of optimization criteria used.

2.4.1. Maximum-Likelihood estimators:

The maximum-likelihood approach is perhaps the most popular approach in statistical estimation theory for deriving practical estimators, and is often used even for the most complicated estimation problems. It was first applied to speech enhancement by McAulay and Malpass [6].

Suppose that we are given an N-point data set $y = \{y(0), y(10, \ldots., y(N-1)\}$ that depends on an unknown parameterθ. In speech enhancement, y (the observed data set) might be the noisy speech magnitude spectrum, and the parameter of interest, θ, might be the clean speech magnitude spectrum. Furthermore, suppose that we know the pdf of y, which we denote by $p(y;\theta)$. The pdf of y is parameterized by the unknown parameter θ, and we denote that by the semicolon. As the parameter θ affects the probability of y, we should be able to infer the values of θ from the observed values of y. Mathematically, we can look for the value of θ that maximizes $p(y;\theta)$, that is

$$\widehat{\theta_{ML}} = \arg\max_{\theta} p(y;\theta) \tag{16}$$

The preceding estimate, $\widehat{\theta_{ML}}$, is called the maximum-likelihood estimate of θ. The pdf $p(y;\theta)$ is called the likelihood function as it can be viewed as a function of an unknown parameter. To find $\widehat{\theta_{ML}}$, we differentiate $p(y;\theta)$ with respect to θ, set the derivative equal to zero, and solve for θ. we have considered it is convenient to find $\widehat{\theta_{ML}}$ by differentiating instead the log of $p(y;\theta)$, which is called the log-likelihood function.

It is important to note that the parameter θ is assumed to be unknown but deterministic. This assumption differentiates the MLE approach from the Bayesian one, in which θ is assumed to be random.

Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean signal $x(n)$ and the noise signal $d(n)$. In the frequency domain, we have

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{17}$$

For $\omega_k = \frac{2\pi k}{N}$ $and\ k = 0,1,2,\dots,N-1$, where N is the frame length in samples.
The preceding equation can also be expressed in polar form as:

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{18}$$

Where $\{X_k, Y_k, D_k\}$ denote the magnitude and $\{\theta_y(k), \theta_x(k), \theta_d(k)\}$ denote the phases at frequency bin k of the noisy speech, clean speech, and noise, respectively.

In the maximum-likelihood approach proposed by McAulay and Malpass, the magnitude and phase spectra of the clean signal, i.e., X_k and $\theta_x(k)$ are assumed to be unknown but deterministic. The probability density functions of the noise Fourier transform coefficients $D(\omega_k)$ is assumed to be zero-mean, complex Gaussian. The real and imaginary parts of $D(\omega_k)$ are assumed to have variances $\frac{\lambda_d(k)}{2}$. Based on these two assumptions, we can form the probability density of the observed noisy-speech DFT coefficients, $Y(\omega_k)$. The probability density of $Y(\omega_k)$ is also Gaussian with variance $\lambda_d(k)$ and the mean $X_k e^{j\theta_x(k)}$:

$$p(Y(\omega_k); X_k, \theta_x(k)) = \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{\left|Y(\omega_k) - X_k e^{j\theta_x(k)}\right|^2}{\lambda_d(k)}\right]$$

$$= \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{Y_k^2 - 2X_k Re\{e^{-j\theta_x(k)} Y(\omega_k)\} + X_k^2}{\lambda_d(k)}\right] \tag{19}$$

To obtain the maximum-likelihood of estimate of X_k, we need to compute the maximum of $p(Y(\omega_k); X_k, \theta_x(k))$ with respect to X_k. This is not straight forward, however, because $p(Y(\omega_k); X_k, \theta_x(k))$ is a function of two unknown parameters: the magnitude and the phase. The phase parameter is considered to be a nuisance parameter, which can be easily eliminated by "integrating out" More specifically, we can eliminate the phase parameter by maximizing instead the following average likelihood function:

$$p(Y(\omega_k); X_k) = \int_0^{2\pi} p(Y(\omega_k); X_k, \theta_x)\, p(\theta_x) d\theta_x \tag{20}$$

Assuming a uniform distribution on $(0,2\pi)$ for the phase θ_x, i.e., assuming that $p(\theta_x) = \frac{1}{2\pi}$ for $\theta_x \in [0,2\pi]$, the likelihood function becomes:

$$p(Y(\omega_k); X_k) = \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{Y_k^2+X_k^2}{\lambda_d(k)}\right]\frac{1}{2\pi}\int_0^{2\pi} exp\left[\frac{2X_k Re(e^{-j\theta_x}Y(\omega_k))}{\lambda_d(k)}\right] d\theta_x \tag{21}$$

The integral in the preceding equation is known as the modified Bessel function of the first kind and is given by:

$$I_0(|x|) = \frac{1}{2\pi}\int_0^{2\pi} exp\left[Re(xe^{-j\theta_x})\right] d\theta_x \tag{22}$$

Where $x = 2X_k Y(\omega_k)/\lambda_d(k)$ (23)

As shown in Figure.1 for values of $|x| > 0.258$, the preceding Bessel function can be approximated as:

$$I_0(|x|) = \frac{1}{\sqrt{2\pi|x|}} \exp\,(|x|) \tag{24}$$

Substituting eqn. (23) in eqn. (24), we get

$$I_0(|2X_k Y(\omega_k)/\lambda_d(k)\,|) = \frac{1}{\sqrt{2\pi|2X_k Y(\omega_k)/\lambda_d(k)\,|}} \exp\,(|2X_k Y(\omega_k)/\lambda_d(k)\,|) \tag{25}$$

Again now using eqn. (25) for Bessel function substitution in eqn. (21) and reordering exponentials we get,

$$p(Y(\omega_k); X_k) = \frac{1}{\pi\lambda_d(k)}\frac{1}{\sqrt{2\pi\frac{2X_k Y_k}{\lambda_d(k)}}} exp\left[-\frac{Y_k^2+X_k^2-2X_k Y_k}{\lambda_d(k)}\right] \tag{26}$$

After differentiating the log-likelihood function $\log p(Y(\omega_k); X_k)$ with respect to the unknown X_k and setting the derivative to zero, we get the maximum-likelihood estimate of the magnitude spectrum:

$$\hat{X}_k = \frac{1}{2}\left[Y_k + \sqrt{Y_k^2 - \lambda_d(k)}\right] \tag{27}$$

Using the noisy phase θ_y in place of θ_x, we can express the estimate of the clean signal spectrum as:

$$\hat{X}(\omega_k) = \hat{X}_k e^{j\theta_y} = \hat{X}_k \frac{Y(\omega_k)}{Y_k}$$

$$= \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{Y_k^2-\lambda_d(k)}{Y_k^2}}\right] Y(\omega_k) \tag{28}$$

Letting $\gamma_k = \frac{Y_k^2}{\lambda_d(k)}$ denote the *a posteriori* or measured signal-to-noise ratio (SNR) based on the observed data, the preceding equation can also written as:

$$\hat{X}(\omega_k) = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{\gamma_k - 1}{\gamma_k}}\right] Y(\omega_k)$$
$$= G_{ML}(\gamma_k) Y(\omega_k) \tag{29}$$

Where $G_{ML}(\gamma_k)$ denotes the gain function of the maximum-likelihood estimator. It has been seen that the maximum-likelihood suppression rule provides considerably smaller attenuation compared to the power subtraction and wiener suppression rules.

In the preceding derivation, we assumed that the signal magnitude and phase (X_k and θ_x) were unknown but deterministic. If we now assume that both signal and speech DFT coefficients are modeled as independent, zeros-mean Gaussian random processes, but it is the signal variance, $\lambda_x(k)$, that is unknown and deterministic, we get a different likelihood function. As the signal and noise are assumed to be independent, the variance of $Y(\omega_k)$, denoted as $\lambda_y(k)$, is given by: $\lambda_y(k) = \lambda_x(k) + \lambda_d(k)$. Hence, the probability density of $Y(\omega_k)$ is given by:

$$p\big(Y(\omega_k); \lambda_x(k)\big) = \frac{1}{\pi[\lambda_x(k)+\lambda_d(k)]} exp\left[-\frac{Y_k^2}{\lambda_x(k)+\lambda_d(k)}\right] \tag{30}$$

Maximizing the likelihood function $p\big(Y(\omega_k); \lambda_x(k)\big)$ with respect to $\lambda_x(k)$, we get:

$$\hat{\lambda}_x(k) = Y_k^2 - \lambda_d(k) \tag{31}$$

Assuming that $X_k^2 \approx \lambda_x(k)$ and $D_k^2 \approx \lambda_d(k)$ $(and\ Y_k^2 - \lambda_d(k) > 0)$, we get estimate of the signal magnitude spectrum:

$$\hat{X}_k = \sqrt{Y_k^2 - D_k^2} \tag{32}$$

Note that this estimator of X_k is nothing but the power spectrum subtraction estimator. Hence, the original power spectrum subtraction approach can be derived using maximum-likelihood principles by assuming that the signal and noise Fourier transform coefficients are modeled as independent Gaussian random processes and the signal variance, $\lambda_x(k)$, is unknown but deterministic.

As in eqn. (29), we can compute the estimate of the clean signal spectrum obtained by power spectrum subtraction as:

$$\begin{aligned}\hat{X}(\omega_k) &= \hat{X}_k e^{j\theta_y} = \hat{X}_k \frac{Y(\omega_k)}{Y_k} \\ &= \sqrt{\frac{Y_k^2 - D_k^2}{Y_k^2}}\, Y(\omega_k)\end{aligned} \tag{33}$$

In terms of γ_k, the preceding equation can be written as:

$$\begin{aligned}\hat{X}(\omega_k) &= \sqrt{\frac{\gamma_k - 1}{\gamma_k}}\, Y(\omega_k) \\ &= G_{ps}(\gamma_k) Y(\omega_k)\end{aligned} \tag{34}$$

Where $G_{ps}(\gamma_k)$ is the gain function of the power spectrum subtraction method. Finally, it is worth noting that if we substitute the maximum-likelihood estimate of $\lambda_x(k)$ in the wiener filter eqn. (35):

$$\hat{X}(\omega_k) = \frac{\lambda_x(k)}{\lambda_x(k)+\lambda_d(k)} Y(\omega_k) \tag{35}$$

$$\begin{aligned}\hat{X}(\omega_k) &= \frac{Y_k^2-\lambda_d(k)}{Y_k^2}\, Y(\omega_k) \\ &= \frac{\gamma_k-1}{\gamma_k} Y(\omega_k) \\ &= G_{ps}^2(\gamma_k)\, Y(\omega_k)\end{aligned} \tag{36}$$

Comparing eqn. (34) with eqn. (36), we see that the wiener estimator is the square of the power spectrum subtraction estimator. Consequently, the wiener estimator provides more spectral attenuation than the power spectrum subtraction estimator, for a fixed value of γ_k.

Finally, it should be pointed out that the maximum-likelihood suppression rule is never used by itself, because it does not provide enough attenuation.

2.4.2. Bayesian Estimators:

In the maximum-likelihood approach for parameter estimation, in which we assumed that the parameter of interest, θ, was deterministic but unknown. Now, we assume that θ is a random variable, and we therefore need to estimate the realization of that random variable. This approach is called the Bayesian approach because its implementation is based on Bayes' theorem. The main motivation behind the Bayesian approach is the fact that if we have available *a priori* knowledge about θ, i.e., if we know $p(\theta)$, we should incorporate that knowledge in the estimator to improve estimation accuracy. The Bayesian estimators typically perform better than the MLE estimators, as they make use of prior knowledge.

2.4.2.1. MMSE Estimator:

Acknowledging the importance of the short-time spectral amplitude (STSA) on speech intelligibility and quality, several authors have proposed optimal methods for obtaining the spectral amplitudes from noisy observations. In particular, optimal estimators were sought that minimized the mean-square error between the estimated and true magnitudes:

$$e = E\left\{\left(\hat{X}_k - X_k\right)^2\right\} \tag{37}$$

Where $\hat{X}_k$is the estimate spectral magnitude at frequency ω_k, and X_k is the true magnitude of the clean signal.

The minimization of eqn. (37) can be done in two ways, depending on how we perform the expectation. In the classical mean-square error (MSE) approach, the expectation is done with respect to $p(Y; X_k)$, where Y denotes the observed noisy speech spectrum, $Y = [Y(\omega_0)\, Y(\omega_1) \ldots Y(\omega_{N-1})\,]$. In the Bayesian MSE approach, the expectation is done with respect to the joint pdf $p(Y, X_k)$, and the Bayesian MSE is given by:

$$Bmse(\hat{X}_k) = \iint \left(X_k - \hat{X}_k\right)^2\, p(Y, X_k) dY dX_k \tag{38}$$

Minimization of the Bayesian MSE with respect to $\hat{X}_k$ leads to the optimal MMSE estimator given by [7]:

$$\hat{X}_k = \int X_k \, p(X_k|Y) dX_k$$

$$= E[X_k|Y] \tag{39}$$

$$= E[X_k|Y(\omega_0)Y(\omega_1)\cdots Y(\omega_{N-1})]$$

Which is the mean of the a posteriori probability density function of X_k. The posteriori pdf of the clean spectral amplitudes, i.e., $p(X_k|Y)$, is the pdf of the amplitudes after all the data are observed. In contrast, the a priori pdf of X_k, i.e., $p(X_k)$, refers to the pdf of the clean amplitudes before the data are observed.

Note that there are two fundamental differences between the wiener estimator and the MMSE estimator given in eqn. (39). First, in the wiener filter derivation, we assumed that $\hat{X}(\omega_k) = H_k Y(\omega_k)$ for some unknown filter H_k that is, we assumed that there is a linear relationship between $Y(\omega_k)$ and $\hat{X}(\omega_k)$. Second, the wiener filter is obtained by evaluating the mean of the posterior pdf of $X(\omega_k)$ rather than X_k that is, it is given by $E[X(\omega_k)|Y(\omega_k)]$. The wiener filter is therefore the optimal complex spectrum estimator and not the optimal magnitude spectrum estimator under the assumed model.

The MMSE estimator given in eqn. (40), unlike the wiener estimator does not assume the existence of a linear relationship between the observed data and the estimator, but it does require knowledge about the probability distribution of the speech and noise DFT coefficients. Assuming that we do have prior knowledge about the distributions of the speech and noise DFT coefficients, we can evaluate the mean of the posterior probability density function of X_k, that is, the mean of $p(X_k|Y)$.

Measuring the true probability distributions of the speech Fourier transform coefficients, however, has been difficult, largely because the speech signal is neither a stationary nor an ergodic process. Several have attempted to measure the probability distributions by examining the long-time behavior of the processes [8-10]. As argued in [4], however, it is questionable whether histograms of the Fourier coefficients, obtained using a large amount of data, measure the relative frequency of the Fourier transform coefficients rather than the true probability density of the Fourier transform coefficients.

To circumvent these problems, Ephraim and Malah [4] proposed a statistical model that utilizes the asymptotic statistical properties of the Fourier transform coefficients [11]. This model makes two assumptions:

1. The Fourier transform coefficients (real and imaginary parts) have a Gaussian probability distribution. The mean of the coefficients is zero, and the variances of the coefficients are time-varying owing to the nonstationarity of speech.
2. The Fourier transform coefficients are statistically independent and hence uncorrelated.

The Gaussian assumptions are motivated by the central limit theorem, as the Fourier transform coefficients are computed as a sum of N random variables. Consider, for instance, the computation of the noisy speech Fourier transform coefficients, $Y(\omega_k)$:

$$Y(\omega_k) = \sum_{n=0}^{N-1} y(n)e^{-j\omega_k n} = y(0) + a_1 y(1) + a_2 y(2) + \cdots + a_{N-1} y(N-1) \tag{40}$$

Where $a_m = \exp(-j\omega_k m)$ are constants, and $y(n)$ is the time-domain samples of the noisy speech signal. According to the central limit theorem [12], if the random variables $\{y(n)\}_{n=0}^{N-1}$ are statistically independent, the density of $Y(\omega_k)$ will be Gaussian. The central limit theorem also holds when sufficiently separated samples are weakly dependent as is the case with the speech signal.

The uncorrelated assumption is motivated by the fact that the correlation between different Fourier coefficients approaches zero as the analysis frame length N approaches infinity [11,13]. In speech applications, however, we are constrained by the nonstationarity of the speech signal to use analysis frame lengths on the order of 20-40msec. This may cause the Fourier transform coefficients to be correlated to some degree [14]. Despite that, overlapping analysis windows are typically used in practice. Although such "window overlap" clearly violates the assumption of uncorrelatedness, the resultant models have proved simple, tractable, and useful in practice. Models that take this correlation into account have also been proposed [15].

2.5. MMSE Magnitude Estimator:

To determine the MMSE estimator we first need to compute the posterior pdf of X_k, i.e., $p(X_k|Y(\omega_k))$. We can use Bayes' rule to determine it as:

$$p(X_k|Y(\omega_k)) = \frac{p(Y(\omega_k)|X_k)p(X_k)}{p(Y(\omega_k))}$$

$$= \frac{p(Y(\omega_k)|X_k)p(X_k)}{\int_0^\infty p(Y(\omega_k)|x_k)p(x_k)dx_k} \tag{41}$$

Where $x_k$is a realization of the random variable X_k. Note that $p(Y(\omega_k))$ is a normalization factor required to ensure that $p(X_k|Y(\omega_k))$ integrates to 1. Assuming statistical independence between the Fourier transform coefficients, i.e.,

$$E[X_k|Y(\omega_0)\,Y(\omega_1)\,Y(\omega_2)\cdots Y(\omega_{N-1})] = E[X_k|Y(\omega_k)] \tag{42}$$

And using the preceding expression for $p(X_k|Y(\omega_k))$, the estimator in eqn. (39) simplifies to:

$$\hat{X}_k = E[X_k|Y(\omega_k)]$$

$$= \int_0^\infty x_k p(x_k|Y(\omega_k))\,dx_k \tag{43}$$

$$= \frac{\int_0^\infty x_k p(Y(\omega_k)|x_k)p(x_k)dx_k}{\int_0^\infty p(Y(\omega_k)|x_k)p(x_k)dx_k}$$

Since

$$p(Y(\omega_k)|X_k)p(X_k) = \int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x)p(x_{k,}\theta_x)\,d\theta_x \tag{44}$$

Where θ_x is the realization of the phase random variable of $X(\omega_k)$, we get

$$\hat{X}_k = \frac{\int_0^\infty \int_0^{2\pi} x_k p(Y(\omega_k)|x_k,\theta_x) p(x_k,\theta_x) d\theta_x dx_k}{\int_0^\infty \int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x) p(x_k,\theta_x) d\theta_x dx_k} \tag{45}$$

Next, we need to estimate $p(Y(\omega_k)|x_k,\theta_x)$ and $p(x_k,\theta_x)$. From the assumed statistical model, we know that $Y(\omega_k)$ is the sum of two zero-mean complex Gaussian random variables. Therefore, the conditional pdf $p(Y(\omega_k)|x_k,\theta_x)$ will also be Gaussian:

$$p(Y(\omega_k)|x_k,\theta_x) = p_D(Y(\omega_k) - X(\omega_k)) \tag{46}$$

Where $p_D(\cdot)$ is the pdf of the noise Fourier transform coefficients, $D(\omega_k)$. The preceding equation then becomes:

$$p(Y(\omega_k)|x_k,\theta_x) = \frac{1}{\pi\lambda_d(k)} exp\left\{-\frac{1}{\lambda_d(k)}|Y(\omega_k) - X(\omega_k)|^2\right\} \tag{47}$$

Where $\lambda_d(k) = E\{|D(\omega_k)|^2\}$ is the variance of the kth spectral component of the noise. For complex Gaussian random variables, we know that the magnitude (X_k) and the phase ($\theta_x(k)$) random variables of $X(\omega_k)$ are independent, and can therefore evaluate the joint pdf $p(x_k,\theta_x)$ as the product of the individual pdf's, i.e., $p(x_k,\theta_x) = p(x_k)p(\theta_x)$. The pdf of X_k is Rayleigh since $X_k = \sqrt{r(k)^2 + i(k)^2}$, where $r(k) = Re\{X(\omega_k)\}$ and $i(k) = Im\{X(\omega_k)\}$ are Gaussian random variables. The pdf of $\theta_x(k)$ is uniform in $(-\pi,\pi)$ and therefore the joint probability $p(x_k,\theta_x)$ is given by:

$$p(x_k,\theta_x) = \frac{x_k}{\pi\lambda_x(k)} exp\left\{-\frac{x_k^2}{\lambda_x(k)}\right\} \tag{48}$$

Where $\lambda_x(k) = E\{|X(\omega_k)|^2\}$ is the variance of the kth spectral components of the clean signal. Substituting eqn. (47) and eqn. (48) into eqn. (45), we get:

$$\hat{X}_k = \frac{\int_0^\infty \int_0^{2\pi} x_k^2 exp\left[-\frac{Y_k^2 - 2x_k Re\left(e^{-j\theta_x} Y(\omega_k)\right) + x_k^2}{\lambda_d(k)} - \frac{x_k^2}{\lambda_x(k)}\right] d\theta_x dx_k}{\int_0^\infty \int_0^{2\pi} x_k exp\left[-\frac{Y_k^2 - 2x_k Re\left(e^{-j\theta_x} Y(\omega_k)\right) + x_k^2}{\lambda_d(k)} - \frac{x_k^2}{\lambda_x(k)}\right] d\theta_x dx_k}$$

$$= \frac{\int_0^\infty x_k^2 exp\left[-\frac{x_k^2}{\lambda_k}\right] \int_0^\infty exp[2x_k Re\{e^{-j\theta_x} Y(\omega_k)\}] d\theta_x dx_k}{\int_0^\infty x_k exp\left[-\frac{x_k^2}{\lambda_k}\right] \int_0^\infty exp[2x_k Re\{e^{-j\theta_x} Y(\omega_k)\}] d\theta_x dx_k} \tag{49}$$

Where $\frac{1}{\lambda_k} = \frac{1}{\lambda_d(k)} + \frac{1}{\lambda_x(k)}$ (50)

Note that λ_k can also be expressed as:

$$\lambda_k = \frac{\lambda_x(k)\lambda_d(k)}{\lambda_x(k) + \lambda_d(k)} = \frac{\lambda_x(k)}{1 + \xi_k} \tag{51}$$

The inner integral in eqn. (49) is the modified Bessel function of the first kind, and has the following form:

$$I_0(|z|) = \frac{1}{2\pi}\int_0^{2\pi} exp\left[Re(ze^{-j\theta_x})\right]d\theta_x \tag{52}$$

Where $z = 2x_k Y(\omega_k)/\lambda_d(k)$. Using the preceding integral relationship in eqn. (49) We get:

$$\hat{X}_k = \frac{\int_0^\infty x_k^2 exp\left[-\frac{x_k^2}{\lambda_k}\right] I_0(2x_k Y(\omega_k)/\lambda_d(k))dx_k}{\int_0^\infty x_k exp\left[-\frac{x_k^2}{\lambda_k}\right] I_0(2x_k Y(\omega_k)/\lambda_d(k))dx_k} \tag{53}$$

The ratio $Y_k/\lambda_d(k)$ in the preceding eqn. can be expressed in terms of λ_k(eqn. (51)) as follows:

$$\begin{aligned}\frac{Y_k}{\lambda_{d(k)}} &= \sqrt{\frac{Y_k^2}{\lambda_d(k)}\frac{\lambda_x(k)}{\lambda_d(k)}\frac{1}{\lambda_x(k)}} \\ &= \sqrt{\frac{\gamma_k \xi_k}{\lambda_x(k)}} = \sqrt{\frac{\frac{\gamma_k}{\xi_k+1}\xi_k}{\frac{\lambda_x(k)}{\xi_k+1}}} \\ &= \sqrt{\frac{\nu_k}{\lambda_k}}\end{aligned} \tag{54}$$

And eqn. (49) reduces to:

$$\hat{X}_k = \frac{\int_0^\infty x_k^2 exp\left[-\frac{x_k^2}{\lambda_k}\right] I_0(2x_k\sqrt{\frac{\nu_k}{\lambda_k}})dx_k}{\int_0^\infty x_k exp\left[-\frac{x_k^2}{\lambda_k}\right] I_0(2x_k\sqrt{\frac{\nu_k}{\lambda_k}})dx_k} \tag{55}$$

We can evaluate the preceding integral using confluent hypergeometric functions as follows:

$$\begin{aligned}\hat{X}_k &= \frac{\Gamma(1.5)(\lambda_k)^{\frac{3}{2}}\phi(\frac{3}{2},1;\nu_k)}{\lambda_x\phi(1,1;\nu_k)} \\ &= \Gamma(1.5)\sqrt{\lambda_x}\frac{\phi(\frac{3}{2},1;\nu_k)}{\phi(1,1;\nu_k)}\end{aligned} \tag{56}$$

Where $\phi(a,b;z)$ is the confluent hypergeometric function and $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$. Using the hypergeometric relationship for the numerator and denominator, the preceding estimator finally simplifies to:

$$\begin{aligned}\hat{X}_k &= \frac{\Gamma(1.5)\sqrt{\lambda_k}\, e^{\nu_k}\phi(-0.5;1;-\nu_k)}{e^{\nu_k}} \\ &= \Gamma(1.5)\sqrt{\lambda_k}\phi(-0.5,1;-\nu_k)\end{aligned} \tag{57}$$

The above equation can be further simplified using some mathematical relationships and we get:

$$\hat{X}_k = \frac{\sqrt{\pi}}{2}\sqrt{\lambda_k}\exp\left(-\frac{\nu_k}{2}\right)\left[(1+\nu_k)I_o\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] \tag{58a}$$

Where $I_o(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively.

The eqn. (58a) is a function of two parameters: the *a priori SNR* $\xi_k$and the *a posteriori SNR* value γ_k. we can express the estimated magnitude in terms of a gain function, i.e., $\hat{X}_k = G(\xi_k, \gamma_k)Y_k$. The spectral gain function $G(\xi_k, \gamma_k)$:

$$G(\xi_k, \gamma_k) = \frac{\hat{X}_k}{Y_k} = \frac{\sqrt{\pi}}{2}\sqrt{\lambda_k}\exp\left(-\frac{\nu_k}{2}\right)\left[(1+\nu_k)\mathrm{I}_o\left(\frac{\nu_k}{2}\right) + \nu_k \mathrm{I}_1\left(\frac{\nu_k}{2}\right)\right] \tag{58b}$$

To examine the dependency of $\xi_k$and $\gamma_k$on the gain function, we can plot $G(\xi_k, \gamma_k)$ as a function of the *a priori SNR* for a fixed *a posteriori SNR* values.

Note that for large values (~20dB) of the instantaneous SNR, the MMSE gain function is similar to the Wiener gain function, which is given by

$$G_w(\xi_k) = \frac{\xi_k}{\xi_k + 1} \tag{59}$$

In other words, the MMSE estimator behaves like the Wiener estimator when ξ_k is large. Note that unlike the MMSE gain function, the Wiener gain function does not depend on the *a posteriori SNR* γ_k. The fact that $G(\xi_k, \gamma_k)$ depends on both ξ_k and γ_k will prove to be important for reducing the musical noise.

The spectral gain function can alternatively be plotted as a function of the *a posteriori SNR* $(\gamma_k - 1)$ for a fixed values of ξ_k. The fact that the suppression curve is relatively flat for a wide range of $\gamma_k$when $\xi_k \geq -10dB$ suggests that the *a posteriori SNR* γ_k has a small effect on suppression. This suggests that the *a priori SNR* $\xi_k$is the main parameter influencing suppression. The effect of γ_k on suppression is only evident for extremely low values of ξ_k(i.e., $\xi_k = -15dB$). The behavior of $\gamma_k$is counterintuitive in that more suppression is applied when $\gamma_k$is low.

2.5.1. Estimating the A priori SNR:

The MMSE amplitude estimator eqn. (58) was derived under the assumption that the *a priori SNR* $\xi_k$and the noise variance $\lambda_d(k)$ are known. However, we only have access to the noisy speech signal. The noise variance can be estimated easily assuming noise stationarity, and can in principle be computed during nonspeech activity with the aid of a voice activity detector or a noise estimation algorithm. Estimating ξ_k, however, is considerably more difficult.

Ephraim and Malah[4] first examined the sensitivity of the amplitude estimator to inaccuracies of the *a priori SNR* ξ_k. They found the MMSE estimator to be relatively insensitive to small perturbations of the ξ_k value. More interesting was the finding that the MMSE estimator was more sensitive to underestimate rather than overestimates of the *a priori SNR* ξ_k.

Several methods were proposed for estimating the *a priori SNR* $\xi_k$but the method used in this project activity is explained in detail.

2.5.2. Decision-Directed Approach:

The approach is based on the definition of $\xi_k$and its relationship with the *a posteriori SNR* γ_k. We know that $\xi_k$is given by:

$$\xi_k(m) = \frac{E\{X_k^2(m)\}}{\lambda_d(k,m)} \tag{60}$$

We also know that $\xi_k$is related to $\gamma_k$by

$$\begin{aligned}\xi_k(m) &= \frac{E\{Y_k^2(m)-D_k^2(m)\}}{\lambda_d(k,m)} \\ &= \frac{E\{Y_k^2(m)\}}{\lambda_d(k,m)} - \frac{E\{D_k^2(m)\}}{\lambda_d(k,m)} \\ &= E\{\gamma_k(m)\} - 1\end{aligned} \tag{61}$$

Combining the two expressions for ξ_k i.e.,eqn. (60) and eqn. (61), we get:

$$\xi_k(m) = E\left\{\frac{1}{2}\frac{X_k^2(m)}{\lambda_d(k,m)} + \frac{1}{2}[\gamma_k(m) - 1]\right\} \tag{62}$$

The final estimator for $\xi_k$is derived by making the preceding eqn. recursive:

$$\hat{\xi}_k(m) = a\frac{\hat{X}_k^2(m-1)}{\lambda_d(k,m-1)} + (1-a)max[\gamma_k(m) - 1,0] \tag{63}$$

Where $0 < a < 1$ is the weighting factor replacing the ½ in eqn. (62), and $\hat{X}_k^2(m-1)$ is the amplitude estimator obtained in the past analysis frame. The $\max(\cdot)$ operator is used to ensure the positiveness of the estimator, as $\hat{\xi}_k(m)$ needs to be nonnegative.

This new estimator of ξ_k is a weighted average of the past *a priori SNR* (given by the first term) and the present *a priori SNR* estimate (given by the second term). Note that the present *a priori SNR* estimate is also the maximum-likelihood estimate of the SNR. Eqn. (63) was called the decision-directed estimator because $\hat{\xi}_k(m)$ is updated using information from the previous amplitude estimate. The decision-directed approach for estimating the a priori SNR was found not only important for MMSE-type algorithms but also in other algorithms

Eqn. (63) needs initial conditions for the first frame, i.e. for $m = 0$. The following initial conditions were recommended [6] for $\hat{\xi}_k(n)$:

$$\hat{\xi}_k(0) = a + (1-a)max[\gamma_k(0) - 1,0] \tag{64}$$

Good results were obtained with $a = 0.98$.

2.5.3. Elimination of MUSICAL NOISE

Ephraim and Malah [4] noted that when the *a priori* SNR was estimated using the decision-directed approach, the enhanced speech had no "musical noise."But when the ML (one of the method to estimate *a priori* SNR, this method should not be confused with speech enhancement method) approach was used to estimate the *a priori* SNR, the enhanced signal had musical noise. Yet, in both cases the same suppression rule was used. No explanation was given in [4] as to why that was the case. Cappe

[16] 10 years later, provided a detailed explanation of the mechanisms that countered the musical noise phenomenon.

Cappe noted that the effectiveness of the *a priori* SNR estimator is closely coupled to the suppression rule. The suppression rule in eqn. (58) is greatly affected by both *a priori* SNR ξ_k and *a posteriori SNR* γ_k parameters. Of the two parameters, the *a priori* SNR ξ_k is the dominant one in that it exerts the most influence of suppression. But what is the role of the *a posteriori SNR* γ_k?

a posteriori SNR $\gamma_k$acts as a correction parameter that influences attenuation only when ξ_k is low. The correction, however, is done in an intuitively opposite direction. As in figure, Strong attenuation is applied when γ_k is large, and not when γ_k as we could expect. This counterintuitive behavior is not an artifact of the algorithm, but it is actually useful when dealing with low-energy speech segments.

Understanding the dominant behavior of ξ_k on suppression is critical in understanding the mechanism responsible for eliminating musical noise. The underlying mechanism for eliminating the musical noise lies in the recursive calculation of the *a priori* SNR. The decision-directed estimator of ξ_k exhibits two different types of behaviors, depending on the value of γ_k. When γ_k stays below or close to 0dB, the ξ_k estimate corresponds to a smoothed version of γ_k. In fact, when $\gamma_k$is large ξ_k can be approximated as $\hat{\xi}_k(m) = (1-\alpha)\gamma_k(m-1)$-that is, ξ_k follows $\gamma_k$but with a delay of one frame. Increasing the value of α increases the time delay, and that might have an adverse effect when encountering short transient segments of speech.

As the attenuation in the MMSE algorithm is primarily influenced by the smoothed value of the *a priori* SNR, the attenuation itself will not change radically from frame to frame. Consequently, the musical noise will be reduced or eliminated altogether. In contrast, the spectral subtraction algorithm depends on the estimation of the *a posteriori SNR,* which can change radically from frame to frame. As a result, musical noise is produced. It is the smoothing behavior of the decision-directed approach in conjunction with the suppression rule that is responsible for reducing the musical noise effect in the MMSE algorithm.

2.6. LOG-MMSE ESTIMATOR

In the previous section, we derived the optimal MMSE spectral amplitude estimator, which minimized the error of the spectral magnitude spectra. Although a metric based on the squared error of the magnitude spectra is mathematically tractable, it may not be subjectively meaningful. In this section a metric based on the squared error of the log-magnitude spectra may be more suitable for speech processing. Derivation of an estimator that minimizes the mean-square error of the log-magnitude spectra is as follows:

$$E\{(\log X_k - \log \hat{X}_k)^2\} \tag{65}$$

The optimal log-MMSE estimator can be obtained by evaluating the conditional mean of the $\log X_k$, i.e.,

$$\log \hat{X}_k = E\{\log X_k | Y(\omega_k)\} \tag{66}$$

From which we can solve for $\hat{X}_k$:

$$\hat{X}_k = \exp\left(E\{\log X_k | Y(\omega_k)\}\right) \tag{67}$$

The evaluation of $E\{\log X_k | Y(\omega_k)\}$ is not straightforward but can be simplified if we use the moment-generating function of X_k conditioned on $Y(\omega_k)$.

Let $Z_k = \log X_k$, then the moment-generating function of Z_k conditioned on $Y(\omega_k)$ is given by:

$$\begin{aligned} \emptyset_{Z_k|Y(\omega_k)}(\mu) &= E\{exp[\mu Z_k]|Y(\omega_k)\} \\ &= E\{X_k^{\mu}|Y(\omega_k)\} \end{aligned} \tag{68}$$

The conditional mean of $\log X_k$ can then be obtained from the moment-generating function by evaluating the derivative of $\emptyset_{Z_k|Y(\omega_k)}(\mu)$ at $\mu = 0$, i.e.,

$$E\{\log X_k | Y(\omega_k)\} = \frac{d}{d\mu}\emptyset_{Z_k|Y(\omega_k)}(\mu)|\mu = 0 \tag{69}$$

We are then left with the task of evaluating the moment-generating function $\emptyset_{Z_k|Y(\omega_k)}(\mu)$. From eqn. (68) we see that we need to evaluate the term $E\{X_k^{\mu}|Y(\omega_k)\}$, which is very similar to (43) i.e.,

$$\begin{aligned} \emptyset_{Z_k|Y(\omega_k)}(\mu) &= E\{X_k^{\mu}|Y(\omega_k)\} \\ &= \frac{\int_0^{\infty}\int_0^{2\pi} x_k^{\mu} p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)d\theta_x dx_k}{\int_0^{\infty}\int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)d\theta_x dx_k} \end{aligned} \tag{70}$$

Using the same statistical model as in derivation of the MMSE estimator, and after substituting (47) and (48) in eqn. (71), we get:

$$\emptyset_{Z_k|Y(\omega_k)}(\mu) = \lambda_k^{\mu/2}\Gamma\left(\frac{\mu}{2}+1\right)\phi(-\mu/2,1;-\nu_k) \tag{71}$$

It is easy to see the similarities between eqn. (71) with eqn. (60) by simply putting $\mu = 1$.

Where $\Gamma(\cdot)$ is the gamma function, $\phi(a,b;x)$ is the confluent hyper geometric function, ν_k is defined in eqn. (54), and γ_k is defined in eqn. (51).

After taking the derivative of $\emptyset_{Z_k|Y(\omega_k)}(\mu)$ with respect to μ and evaluating it at $\mu = 0$, we get the conditional mean of the $\log X_k$:

$$E\{\log X_k | Y(\omega_k)\} = \frac{1}{2}\log\lambda_k + \frac{1}{2}\log\nu_k + \frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt \tag{72}$$

Finally, substituting the preceding equation into eqn. (73), we get the optimal log-MMSE estimator:

$$\begin{aligned} \hat{X}_k &= \frac{\xi_k}{\xi_k+1} exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right\}Y_k \\ &= G_{LSA}(\xi_k,\nu_k)Y_k \end{aligned} \tag{73}$$

Where $\xi_k$is the *a priori* SNR, and $G_{LSA}(\xi_k, \nu_k)$ is the gain function of the log-MMSE estimator. The integral in the preceding eqn. is known as the exponential integral and can be evaluated numerically. The exponential integral, $Ei(x)$, can be approximated as follows:

$$Ei(x) = \int_x^{\infty} \frac{e^{-x}}{x} dx \approx \frac{e^x}{x} \sum_k \frac{k!}{x^k} \tag{74}$$

It has been seen that the gain function of the log-MMSE estimator is shifted down for the most part by 3 dB relative to the gain function of the linear-MMSE estimator. This suggests that the log-MMSE estimator provides more attenuation than the linear-MMSE estimator for the same values of the *a posteriori* and *a priori* SNRs. This also confirmed with listening experiments. The log-MMSE estimator reduces the residual noise, and most importantly, without affecting the speech signal itself, i.e., without introducing much speech distortion. It is also clear from these spectrograms that the log-MMSE estimator reduces the residual noise considerably without affecting the speech signal.

2.7. INCORPORATING SPEECH ABSENCE PROBABILITY IN SPEECH ENHANCEMENT:

In the preceding methods, it was implicitly assumed that speech was present at all times. However, in reality speech contains a great deal of pauses, even during speech activity. The stop closures, for example, which are brief silent periods occurring before the burst of stop consonants, often appear in the middle of a sentence. Also, speech may not be present at a particular frequency even during voiced speech segments. This was something that was exploited in multiband speech coders [17], in which spectrum was divided into bands, and each band was declared as being voiced or unvoiced. The voiced bands were assumed to be generated by a periodic excitation, whereas the unvoiced bands were assumed to be generated by random noise. Such a mixed source excitation model was shown to produce better speech quality than the traditional voiced/unvoiced models [18]. It follows then that a better noise suppression rule may be produced if we assume a two-state model for speech events; that is, that either speech is present or it is not.

2.8. Incorporating speech-presence uncertainty in MMSE Estimators:

The linear-MMSE estimator that takes into account the uncertainty of signal presence can be derived with a new estimator given by:

$$\hat{X}_k = E\left(X_k \middle| Y(\omega_k), H_1^k\right) P(H_1^k | Y(\omega_k)) \tag{75}$$

Note that this estimator is uses a complex noisy speech $Y(\omega_k)$, rather than the noisy magnitude spectrum Y_k. To compute $P(H_1^k | Y(\omega_k))$, we use Bayes' rule:

$$P(H_1^k | Y(\omega_k)) = \frac{p\left(Y(\omega_k) | H_1^k\right) P(H_1)}{p\left(Y(\omega_k) | H_1^k\right) P(H_1) + p\left(Y(\omega_k) | H_0^k\right) P(H_0)} \tag{76}$$

$$= \frac{\Lambda(\mathrm{Y}(\omega_k), \mathrm{q}_k)}{1 + \Lambda(\mathrm{Y}(\omega_k), \mathrm{q}_k)}$$

Where $\Lambda(\mathrm{Y}(\omega_k), \mathrm{q}_k)$ is the generalized likelihood ratio defined by:

$$\Lambda(\mathrm{Y}(\omega_\mathrm{k}), \mathrm{q_k}) = \frac{1-\mathrm{q_k}}{\mathrm{q_k}} \frac{p(Y(\omega_k)|H_1)}{p(Y(\omega_k)|H_0)} \tag{77}$$

Where $q_k \cong P(H_0^k)$ denotes the *a priori* probability of speech absence for frequency bin k. The *a priori* probability of speech presence, i.e., $P(H_1^k)$, is given by $(1 - q_k)$.

Under hypothesis H_0, $Y(\omega_k) = D(\omega_k)$, and as the pdf of the noise Fourier transform coefficients, $D(\omega_k)$, is complex Gaussian with zero mean and variance $\lambda_d(k)$, it follows that $p(Y(\omega_k)|H_0^k)$ will also have a Gaussian distribution with the same variance, i.e.,

$$p\big(Y(\omega_k)|H_0^k\big) = \frac{1}{\pi\lambda_d(k)} exp\left(-\frac{Y_k^2}{\lambda_d(k)}\right) \tag{78}$$

Under hypothesis $H_1, Y(\omega_k) = X(\omega_k) + D(\omega_k)$, and because the pdfs of $D(\omega_k)$ and $X(\omega_k)$ are complex Gaussian with zero mean and variances $\lambda_d(k)$ and $\lambda_x(k)$, respectively, it follows that $Y(\omega_k)$ will also have a Gaussian distribution with variance $\lambda_d(k) + \lambda_x(k)$:

$$p\big(Y(\omega_k)|H_1^k\big) = \frac{1}{\pi[\lambda_d(k)+\lambda_x(k)]} exp\left(-\frac{Y_k^2}{\lambda_d(k)+\lambda_x(k)}\right) \tag{79}$$

Substituting eqn. (78) and eqn. (79) into eqn. (77), we get an expression for the likelihood ratio:

$$\Lambda(\mathrm{Y}(\omega_\mathrm{k}), \mathrm{q_k}, \xi'_\mathrm{k}) = \frac{1-\mathrm{q_k}}{\mathrm{q_k}} \frac{\exp\left[\frac{\xi'_\mathrm{k}}{1+\xi'_\mathrm{k}}\gamma_\mathrm{k}\right]}{1+\xi'_\mathrm{k}} \tag{80}$$

Where ξ'_k indicates the conditional *a priori* SNR:

$$\xi'_k = \frac{E\left[X_k^2|H_1^k\right]}{\lambda_d(k)} \tag{81}$$

Note that the original definition of ξ_k was unconditional, n that it gave the *a priori* SNR of the kth spectral component regardless of whether speech was present or absent at that frequency. In contrast, ξ'_k provides the conditional SNR of the kth spectral component, assuming that speech is present at that frequency. The conditional SNR is not easy to estimate, nut can be expressed in terms of the unconditional SNR ξ_k as follows:

$$\begin{aligned} \xi_k &= \frac{E\left[X_k^2\right]}{\lambda_d(k)} \\ &= P\big(H_1^k\big) \frac{E\left[X_k^2|H_1^k\right]}{\lambda_d(k)} \\ &= (1 - \mathrm{q_k})\xi'_k \end{aligned} \tag{82}$$

Therefore, the conditional SNR ξ'_k is related to the unconditional SNR ξ_k by:

$$\xi'_k = \frac{\xi_k}{1-\mathrm{q_k}} \tag{83}$$

Substituting eqn. (80) in eqn. (76) and after some algebraic manipulations, we express the *a posteriori* probability of speech presence as:

$$P(H_1^k|Y(\omega_k)) = \frac{1-\mathrm{q_k}}{1-\mathrm{q_k}+\mathrm{q_k}(1+\xi_k')\exp\,(-\nu_\mathrm{k}')} \tag{84}$$

Where $\nu_\mathrm{k}' = \frac{\xi_k'}{\xi_k'+1}\gamma_\mathrm{k}$ (85)

It is interesting to note that when ξ_k' is large, suggesting that speech is surely present, $P\left(H_1^k\middle|Y(\omega_k)\right) \approx 1$, as expected. On the other hand, when ξ_k' is extremely small, $P\left(H_1^k\middle|Y(\omega_k)\right) \approx 1 - \mathrm{q_k}$, i.e., it is equal to the a priori probability of speech presence, $P(H_1^k)$.

The final linear-MMSE estimator that incorporates signal presence uncertainty has the form:

$$\hat{X}_k = P\left(H_1^k|Y(\omega_k)G(\xi_k,\gamma_k)\right)|_{\xi_k=\xi_k'}Y_k$$

$$= \frac{1-\mathrm{q_k}}{1-\mathrm{q_k}+\mathrm{q_k}(1+\xi_k')\exp\,(-\nu_\mathrm{k}')}\,G(\xi_k',\gamma_k)Y_k \tag{86}$$

Where $G(\xi_k',\gamma_k)$ is the gain function defined in eqn. (64b) but with $\xi_k$replaced with ξ_k'. Note that if $\mathrm{q_k} = 0$, then $P\left(H_1^k\middle|Y(\omega_k)\right) = 1$ and the preceding estimator reduces to the original linear-MMSE estimator. A comparison between the MMSE estimators that incorporated signal-presence uncertainty with the original MMSE estimator indicated that the former estimator resulted in better speech quality and lower residual noise.

2.9. Incorporating speech-presence uncertainty in Log-MMSE Estimators:

Using a similar procedure, we can derive the log-MMSE estimator that takes into account signal-presence uncertainty. Following eqn. (75), we have:

$$\log\hat{X}_k = E\left(\log X_k\,\middle|Y(\omega_k),H_1^k\right)P(H_1^k|Y(\omega_k)) \tag{87}$$

And after solving for X_k, we obtain:

$$\hat{X}_k = \left(e^{E\left(\log X_k|Y(\omega_k),H_1^k\right)}\right)^{P(H_1^k|Y(\omega_k))} \tag{88}$$

The exponential term in the parenthesis is the log-MMSE estimator and can also be expressed using eqn. (73) as:

$$\hat{X}_k = [G_{LSA}(\xi_k,\nu_k)Y_k]^{P(H_1^k|Y(\omega_k))} \tag{89}$$

Note that the a posteriori probability term $P(H_1^k|Y(\omega_k))$ is no longer multiplicative as it was in eqn. (86). Simulation showed that the preceding estimator did not result in any significant improvements over the original log-MMSE estimator. For this reason the following multiplicatively modified estimator was suggested.

$$\hat{X}_k = [G_{LSA}(\xi'_k, \nu'_k)]P(H_1^k|Y(\omega_k))\, Y_k \tag{90}$$

Where $P(H_1^k|Y(\omega_k))$ is defined in eqn. (84) and $G_{LSA}(\xi'_k, \nu'_k)$ is given by eqn. (79)

$$G_{LSA}(\xi'_k, \nu'_k) = \frac{\xi'_k}{\xi'_k+1} exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{91}$$

And ξ'_k, ν'_k are given by eqn. (83) and eqn. (85), respectively.

The estimator given by eqn. (91) is suboptimal because the probability term $P(H_1^k|Y(\omega_k))$ was forced to be multiplicative. Starting from the original binary speech model given by:

$$\hat{X}_k = E\left(X_k|Y_k, H_1^k\right)P\left(H_1^k|Y_k\right) + E\left(X_k|Y_k, H_0^k\right)P\left(H_0^k|Y_k\right) \tag{92}$$

Taking log on both sides of eqn. (92), we have:

$$\log \hat{X}_k = E\left[\log X_k|Y(\omega_k)\,, H_1^k\right] P\left(H_1^k|Y(\omega_k)\right) + \; E\left[\log X_k|Y(\omega_k)\,, H_0^k\right]P\left(H_0^k|Y(\omega_k)\right) \tag{93}$$

Where $P\left(H_0^k|Y(\omega_k)\right) = 1 - P\left(H_1^k|Y(\omega_k)\right)$ denotes the a posteriori probability of speech absence. The second term $\left(E\left[\log X_k|Y(\omega_k)\,, H_0^k\right]\right)$ was previously assumed to be zero under hypothesis H_0^k. If we now assume that this term is not zero but very small, then we get:

$$\begin{aligned}\hat{X}_k &= e^{E\left[\log X_k|Y(\omega_k),H_1^k\right]P\left(H_1^k|Y(\omega_k)\right)}\; e^{E\left[\log X_k|Y(\omega_k),H_0^k\right]P\left(H_0^k|Y(\omega_k)\right)} \\ &= \left(e^{E\left[\log X_k|Y(\omega_k),H_1^k\right]}\right)^{P\left(H_1^k|Y(\omega_k)\right)} \left(e^{E\left[\log X_k|Y(\omega_k),H_0^k\right]}\right)^{P\left(H_0^k|Y(\omega_k)\right)}\end{aligned} \tag{94}$$

The first exponential in parenthesis is the original log-MMSE estimator and can be expressed as $G_{LSA}(\xi_k, \nu_k)Y_k$, and the second exponential in parenthesis is assumed to be small and is set to $G_{min}Y_k$, where G_{min} is a small value. The preceding estimator then becomes:

$$\begin{aligned}\hat{X}_k &= [G_{LSA}(\xi_k, \nu_k)Y_k]^{P\left(H_1^k|Y(\omega_k)\right)}[G_{min}Y_k]^{P\left(H_0^k|Y(\omega_k)\right)} \\ &= \left[G_{LSA}(\xi_k, \nu_k)^{P\left(H_1^k|Y(\omega_k)\right)}{G_{min}}^{1-P\left(H_1^k|Y(\omega_k)\right)}\right]{Y_k}^{P\left(H_1^k|Y(\omega_k)\right)}{Y_k}^{1-P\left(H_1^k|Y(\omega_k)\right)} \\ &= \left[G_{LSA}(\xi_k, \nu_k)^{P\left(H_1^k|Y(\omega_k)\right)}{G_{min}}^{1-P\left(H_1^k|Y(\omega_k)\right)}\right]Y_k \\ &= G_{OLSA}(\xi_k, \nu_k)Y_k\end{aligned} \tag{95}$$

Note that the new gain function, denoted by $G_{OLSA}(\xi_k, \nu_k)$, is now multiplicative. Comparisons between the preceding optimally modified log-spectrum amplitude (OLSA) estimator and the multiplicatively modified LSA estimator showed that the OLSA estimator yielded better performance in terms of objective segmental SNR measures. The advantage was more significant at low SNR levels.

2.10. Implementation Issues Regarding A priori SNR Estimation:

In section 2.52, the decision-directed approach for estimating the a priori SNR $\hat{\xi}_k(m)$ (eqn. 63). Under speech-presence uncertainty, $\hat{\xi}_k(m)$ is modified by dividing $1 - \mathrm{q_k}$ (eqn.83). Several studies have noted, however, that this division might degrade the performance. In [18], it was shown that it is always preferable to use $\hat{\xi}_k(m)$ rather than $\hat{\xi}_k(m)/1 - \mathrm{q_k}$. For that reason, the original estimate $\hat{\xi}_k(m)$ is often used in both the gain function (e.g., $G_{LSA}(\xi_k, \nu_k)$) and the probability term $P(H_1^k|Y(\omega_k))$ eqn. (84).

Alternatively, a different approach can be used to estimate ξ_k and γ_k under speech-presence uncertainty [26]. The a priori SNR estimate $\hat{\xi}_k(m)$ is first obtained using the decision-directed approach, and then weighted by $P(H_1^k|Y(\omega_k))$ as follows:

$$\hat{\xi}_k(m) = \left(P(H_1^k|Y(\omega_k))\right)\hat{\xi}_k(m) \tag{96}$$

Similarly, the a posteriori SNR estimate $\gamma_k(m)$ at frame m is weighted by $P(H_1^k|Y(\omega_k))$:

$$\hat{\gamma}_k(m) = \left(P(H_1^k|Y(\omega_k))\right)\gamma_k(m) \tag{97}$$

The new estimate $\hat{\xi}_k(m)$ and $\hat{\gamma}_k(m)$ are then used to evaluate the gain function e.g., $G\left(\hat{\xi}_k, \hat{\gamma}_k\right)$.

2.10.1. Methods for estimating the a priori probability of speech absence:

In the preceding methods, the *a priori* probability of speech absence, i.e., $q_k = P\left(H_0^k\right)$ was assumed to be fixed, and in most cases it was determined empirically. In [6], q was set to 0.5 to address the worst-case scenario in which speech and noise are equally likely to occur. In [4], q was empirically set to 0.2 based on listening tests. In running speech, however, we would expect q to vary with time and frequency, depending on the words spoken. Improvements are therefore expected if we could somehow estimate q from the noisy speech signal.

Two methods for estimating q were proposed in [19]. The first method was based on comparing the conditional probabilities of the noisy speech magnitude, assuming that speech is absent or present.

$$P\left(Y_k|H_1^k\right) = \frac{2Y_k}{\lambda_d(k)} exp\left(-\frac{Y_k^2 + X_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2X_kY_k}{\lambda_d(k)}\right) \tag{98}$$

$$P\left(Y_k|H_0^k\right) = \frac{2Y_k}{\lambda_d(k)} exp\left(-\frac{Y_k^2}{\lambda_d(k)}\right) \tag{99}$$

Using above conditional probabilities a binary decision b_k was made for frequency bin k according to:

If $P\left(\left(Y_k|H_1^k\right) > \left(Y_k|H_0^k\right)\right)$ then
    $b_k = 0$ (speech present)
Else
    $b_k = 1$ (speech absent)
End (100)

After making the approximation $\xi_k = X_k^2/\lambda_d(k)$ in eqn. (98), the preceding condition can be simplified and expressed in terms of ξ_k and γ_k alone. More precisely, eqn. (100) becomes:

If $exp(-\xi_k)I_0\left(2\sqrt{\gamma_k\xi_k}\right) > 1$ then
$b_k = 0$ (speech present)
Else (101)
$b_k = 1$ (speech absent)
End

The a priori probability of speech absence for frame m, denoted as $q_k(m)$, can then be obtained by smoothing the values of b_k over past frames:

$$q_k(m) = cb_k + (1 - \mathrm{c})q_k(m-1) \tag{102}$$

Where c is a smoothing constant which was set to 0.1 in [19]. This method for determining the probability of speech absence can be considered as a hard-decision approach, in that the condition in eqn. (101) yields a binary value-speech is either present or absent. It is shown that, the residual noise is reduced substantially after incorporating speech-presence uncertainty. This, however, may come at a price: speech distortion.

3. New proposed algorithm:

In this section a Harmonic wavelet transform is used to realize sub bands using DFT and DCT coefficients. Though the work based on DFT harmonic wavelet transform exists in literature but limited to MMSE2 methods [2]. However in this document MMSE1 and MMSE2 is considered and also work [2] does not mention about DCT harmonic wavelet transform and its implementation.

3.1. DFT-Harmonic Wavelet Transform (DFTWHT)

The continuous wavelet transformation (CWT) of a signal $x(t)$ is given by

$$W_x^{\psi}(b,a) = |a|^{-1/2}\int_{-\infty}^{\infty} x(t)\,\psi^*\left(\frac{t-b}{a}\right)dt \tag{103}$$

Parsevals theorem allows the formulation of eqn. (109) in the frequency domain as

$$W_x(b,a) = |a|^{+\frac{1}{2}}\frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\,\psi^*(a\omega)e^{j\omega b}d\omega \tag{104}$$

Therefore, the wavelet transformation can be calculated by windowing the spectrum $X(\omega)$ with $\psi^*(a\omega)$ and inverse transformation:

$$W_x(b,a) = |a|^{\frac{1}{2}}\boldsymbol{\mathcal{F}}^{-1}\{X(\omega)\psi^*(a\omega)\} \tag{105}$$

Furthermore, the Fourier transform $\psi(\omega)$ of the mother wavelet $\psi(t)$ is chosen to be constant in a limited frequency range and zero outside:

$$\psi(x\omega) = \begin{cases} 1, & \omega_0 - \omega_g < \omega < \omega_0 + \omega_g \\ 0, & ow \end{cases} \tag{106}$$

The corresponding wavelet in the time-domain becomes

$$\psi(t) = \frac{\omega_g}{\pi} \frac{\sin(\omega_g t)}{\omega_g t} e^{j\omega_0 t} \tag{107}$$

For the transformation of discrete signals eqn. (105) becomes

$$\widehat{W_x}(b,a) = |a|^{\frac{1}{2}} \boldsymbol{\mathcal{F}}^{-1}\{X(\Omega)\psi^*(a\Omega)\} \tag{108}$$

Where $\psi(\Omega)$ is the Fourier Transform of the sampled wavelet $\psi(kT_a)$.

In a practical realization of the wavelet-transformation by eqn. (108) for a finite block $x(k), k = 0, \dots N-1$, the DFT of length M=N is used. The windowing is carried out by setting those discrete spectral values to zero which are not in the passband of the corresponding wavelet. After that the modified spectrum will be transformed with the IDFT of length M. In such a way a redundant wavelet-transformation is generated, because for every frequency band M wavelet- coefficients are calculated.

For a wavelet-representation of $x(k)$ with reduced redundancy the sampling rates can be fitted to the bandwidths of the wavelets. A simple way is to transform only M_r non-zero values of the modified spectrum with an IDFT of length $M_r < M$. A block diagram of realizing the wavelet transformation is shown in Figure.1 with a block length of $M = 32$ and one frequency band per octave. Because of the Hermitean symmetry of the DFT for real signals only half of the spectral values have to be considered.

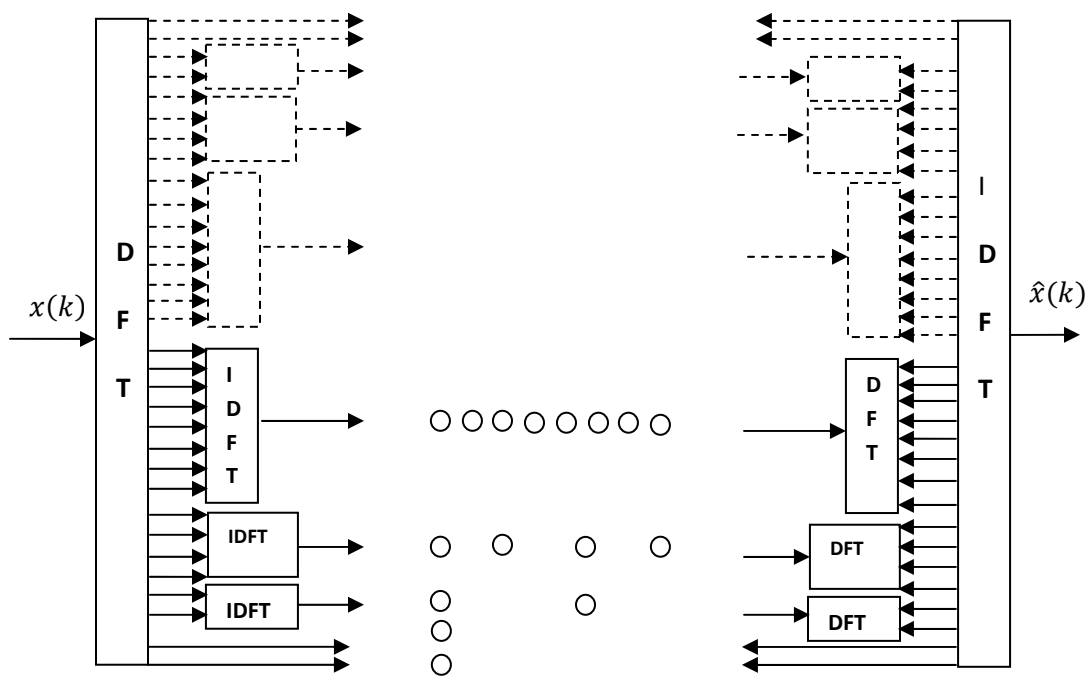


Fig.1. Realization of the harmonic wavelet transform by DFTs

3.2. DCT-Harmonic Wavelet Transform (DCTWHT)

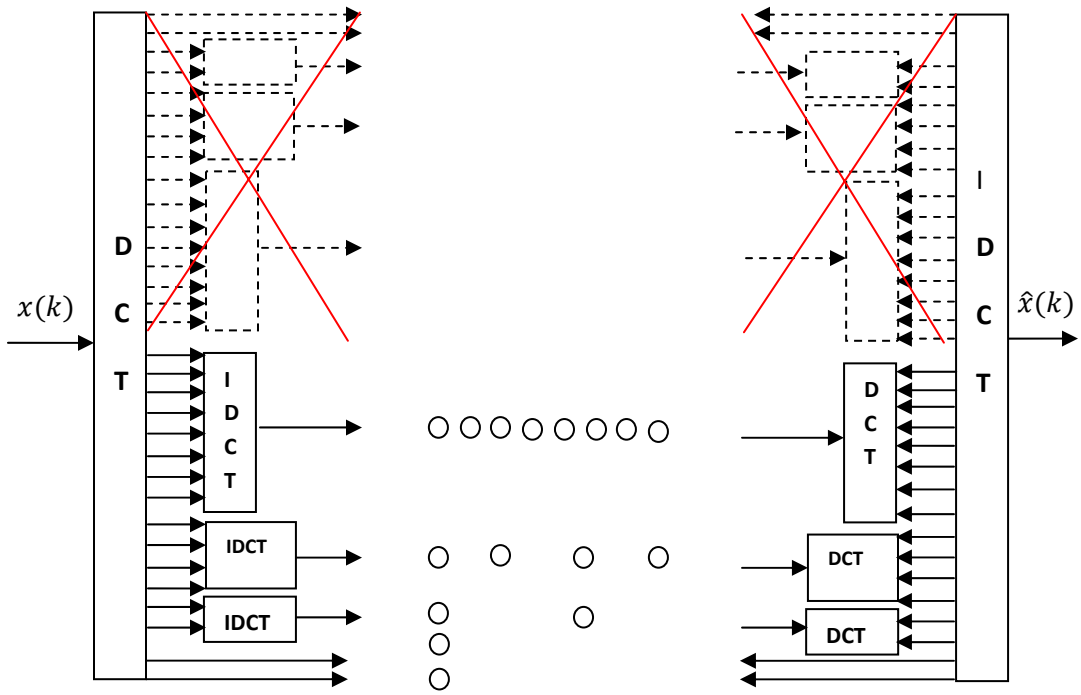


Fig.2. Realization of the harmonic wavelet transform by DCTs

It should be noted that in case of DCT based harmonic wavelet transform structure as shown in Figure.2. Hermitean symmetry does not exist as shown by cross mark, since DCT by itself gives a real transform making the structure computationally much simpler.

The reconstruction of the time signal is implemented with corresponding operations in inverse order.

Simulation Results (White Noise)-DFT

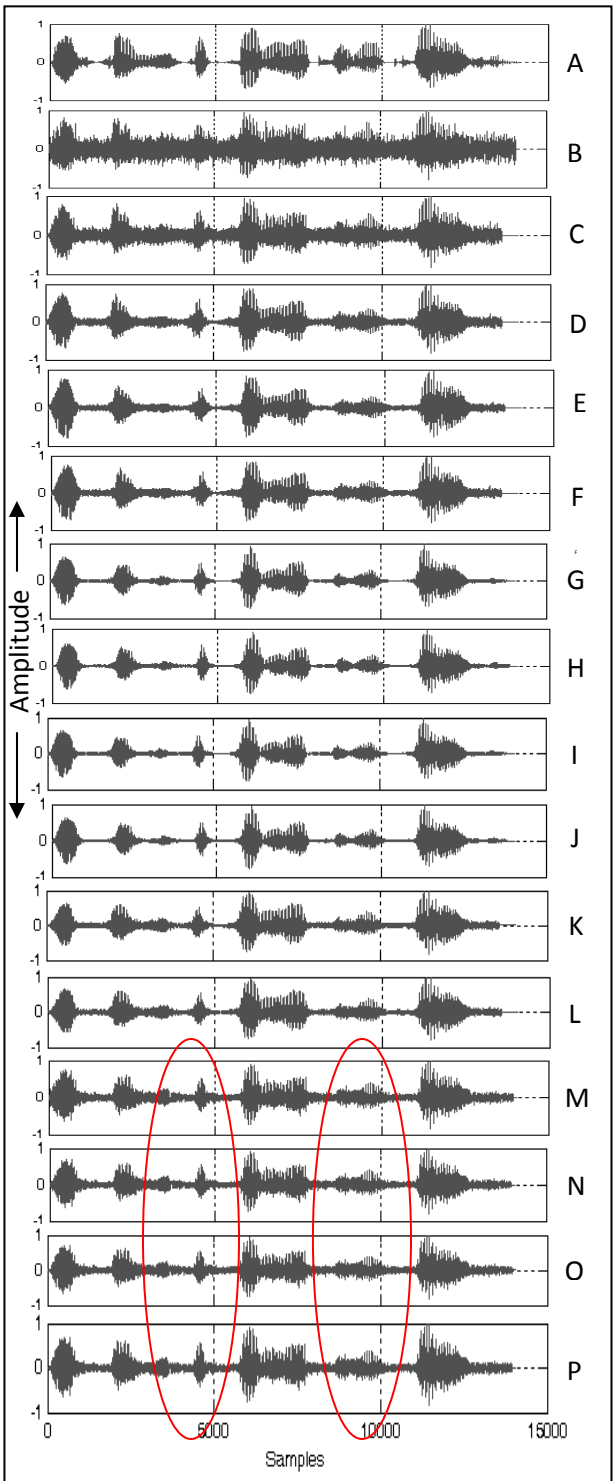


Fig.3. White noise superimposed results (DFT) (0dB)-Speech-1: A) Clean B) Noisy C) ML D) MSS E) PSS F) WF G) MMSE1 H) MMSE1 (SPU) I) MMSE-Log J) MMSE-Log (SPU) K) MMSE1-HWT L)MMSE1-HWT (SPU) M) MMSE2 N) MMSE2 (SPU) O) MMSE2-HWT P) MMSE2-HWT (SPU)

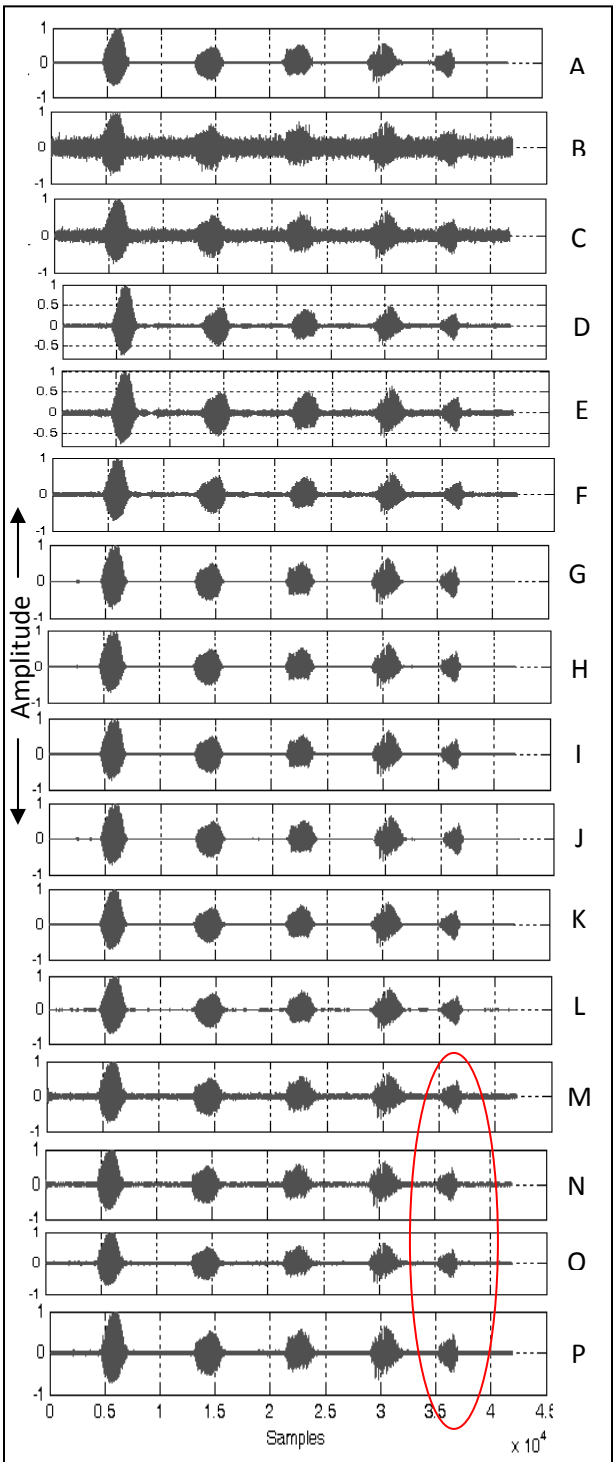


Fig.4. White noise superimposed results (DFT) (0dB)-Speech-1: A) Clean B) Noisy C) ML D) MSS E) PSS F) WF G) MMSE H) MMSE (SPU) I) MMSE-Log J) MMSE-Log (SPU) K) MMSE-HWT L) MMSE-HWT (SPU) M) MMSE2 N) MMSE2 (SPU) O) MMSE2-HWT P) MMSE2-HWT (SPU)

Simulation Results (Pink Noise)-DFT

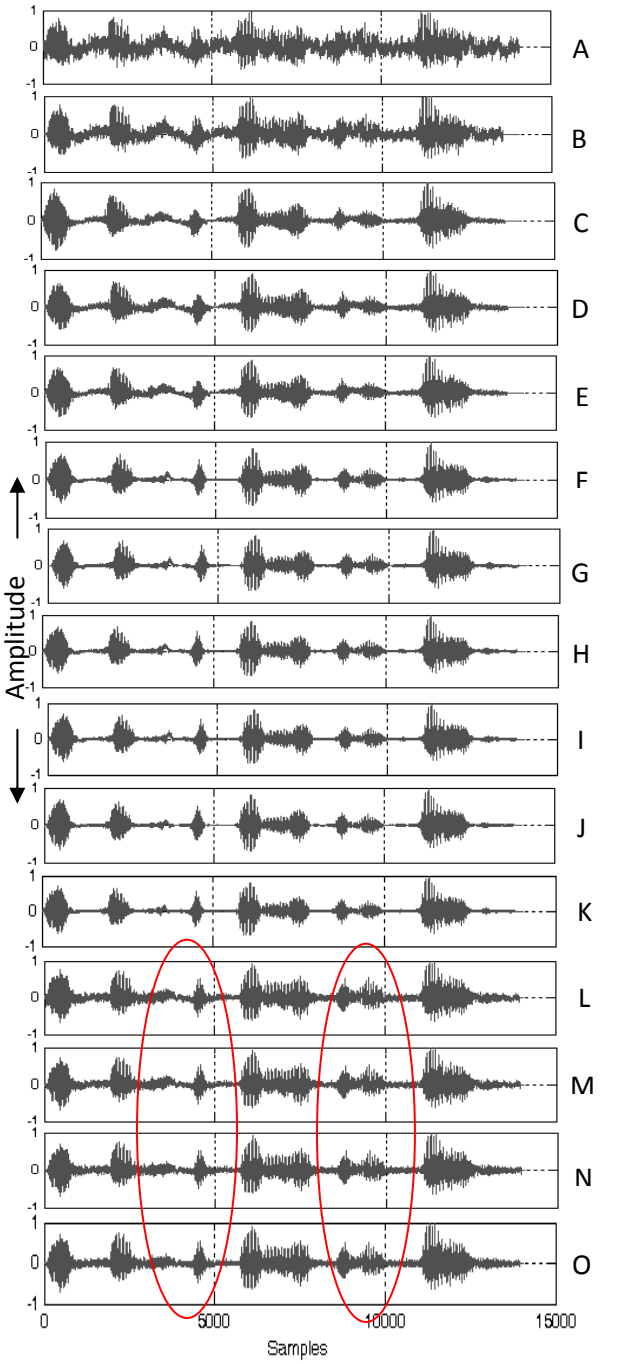


Fig.5. Pink noise superimposed results (DFT) (0dB)-Speech-1: A) Noisy B) ML C) MSS D) PSS E) WF F) MMSE1 G) MMSE1 (SPU) H) MMSE-Log I) MMSE-Log (SPU) J) MMSE1-HWT K) MMSE1-HWT (SPU) L) MMSE2 M) MMSE2 (SPU) N) MMSE2-HWT O) MMSE2-HWT (SPU)

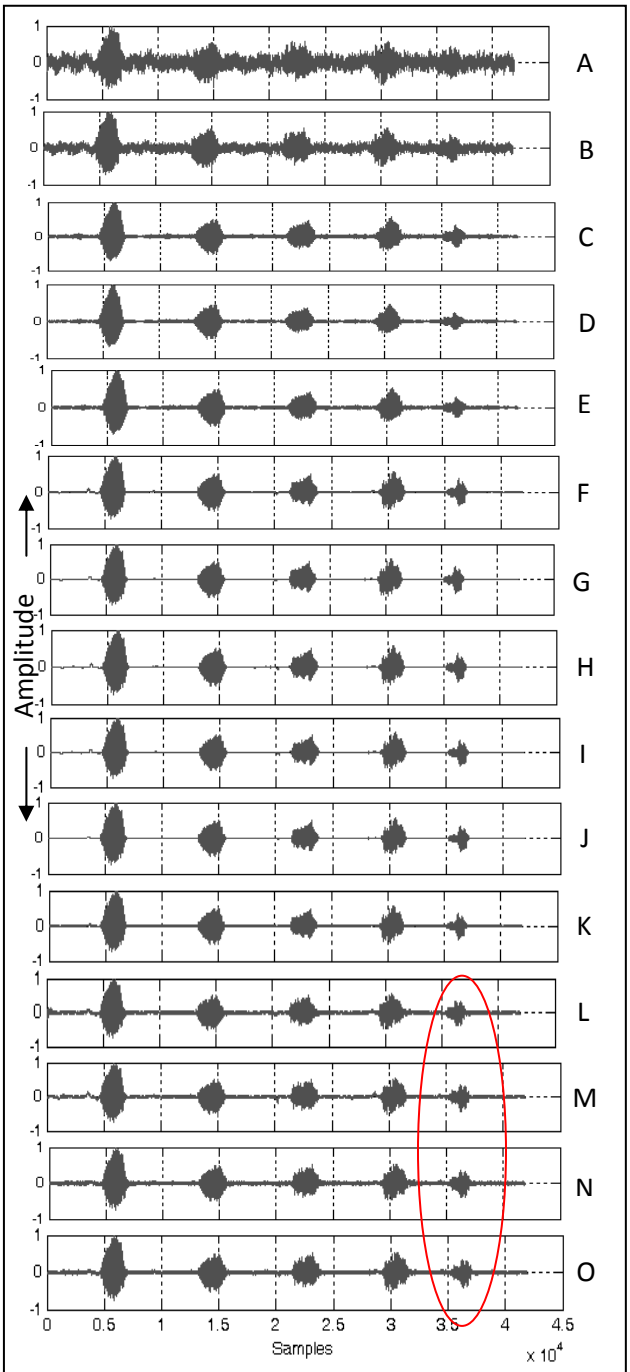


Fig.6. Pink noise superimposed results (DFT) (0dB)-Speech-2: A) Noisy B)ML C) MSS D) PSS E) WF F) MMSE1 G) MMSE 1(SPU) H) MMSE-Log I) MMSE-Log (SPU) J) MMSE1-HWT K) MMSE1-HWT (SPU) L) MMSE2 M) MMSE2 (SPU) N) MMSE2-HWT O) MMSE2-HWT (SPU)

Simulation Results (White Noise)-DCT

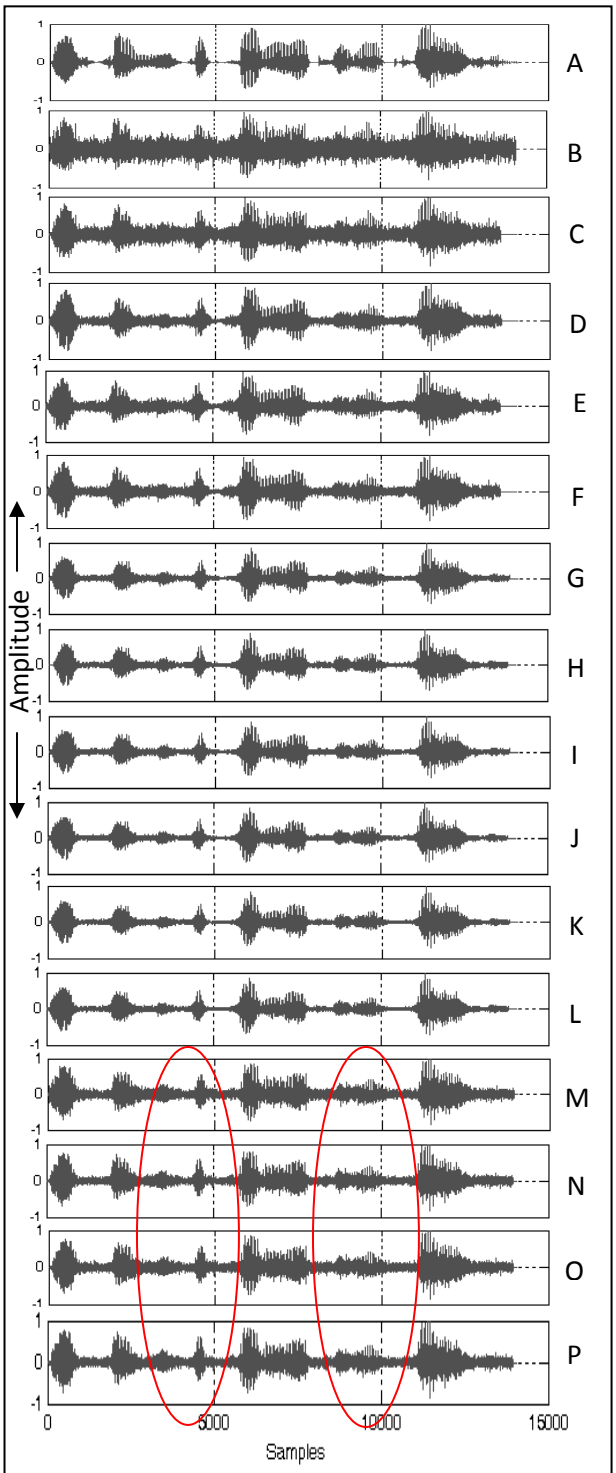


Fig.7. White noise superimposed results (DFT) (0dB)-Speech-1: A) Clean B) Noisy C) ML D) MSS E) PSS F) WF G) MMSE1 H) MMSE1 (SPU) I) MMSE-Log J) MMSE-Log (SPU) K) MMSE1-HWT L)MMSE1-HWT (SPU) M) MMSE2 N) MMSE2 (SPU) O) MMSE2-HWT P) MMSE2-HWT (SPU)

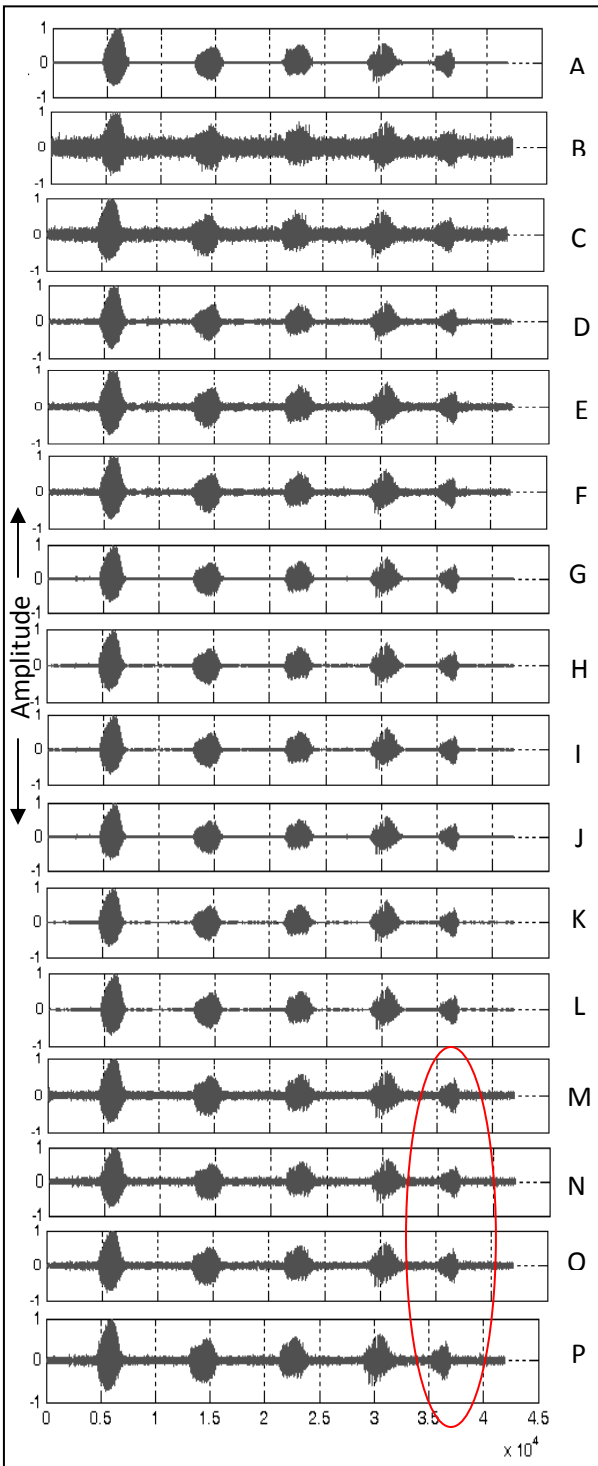


Fig.8. White noise superimposed results (DFT) (0dB)-Speech-1: A) Clean B) Noisy C) ML D) MSS E) PSS F) WF G) MMSE H) MMSE (SPU) I) MMSE-Log J) MMSE-Log (SPU) K) MMSE1-HWT L) MMSE1-HWT (SPU) M) MMSE2 N) MMSE2 (SPU) O) MMSE2-HWT P) MMSE2-HWT (SPU)

Simulation Results (Pink Noise)-DCT

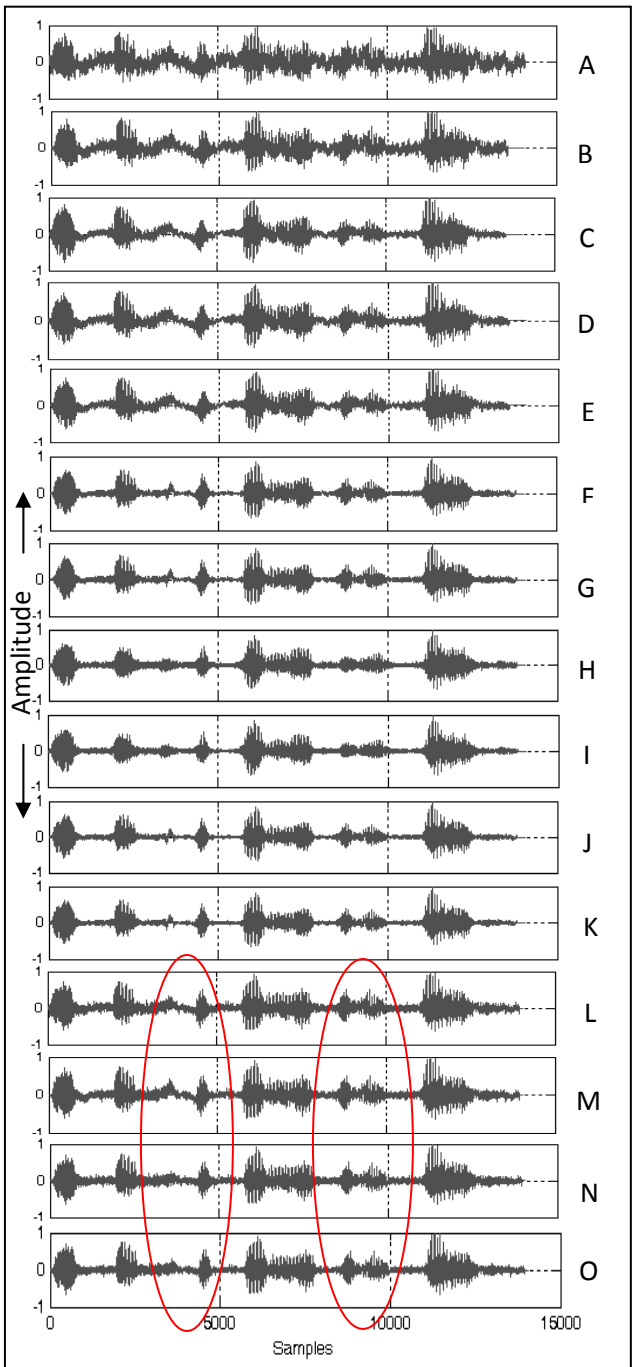


Fig.9. Pink noise superimposed results (DCT) (0dB)-Speech-1: A) Noisy B) ML C) MSS D) PSS E) WF F) MMSE1 G) MMSE1 (SPU) H) MMSE-Log I) MMSE-Log (SPU) J) MMSE1-HWT K) MMSE1-HWT (SPU) L) MMSE2 M) MMSE2 (SPU) N) MMSE2-HWT O) MMSE2-HWT (SPU)

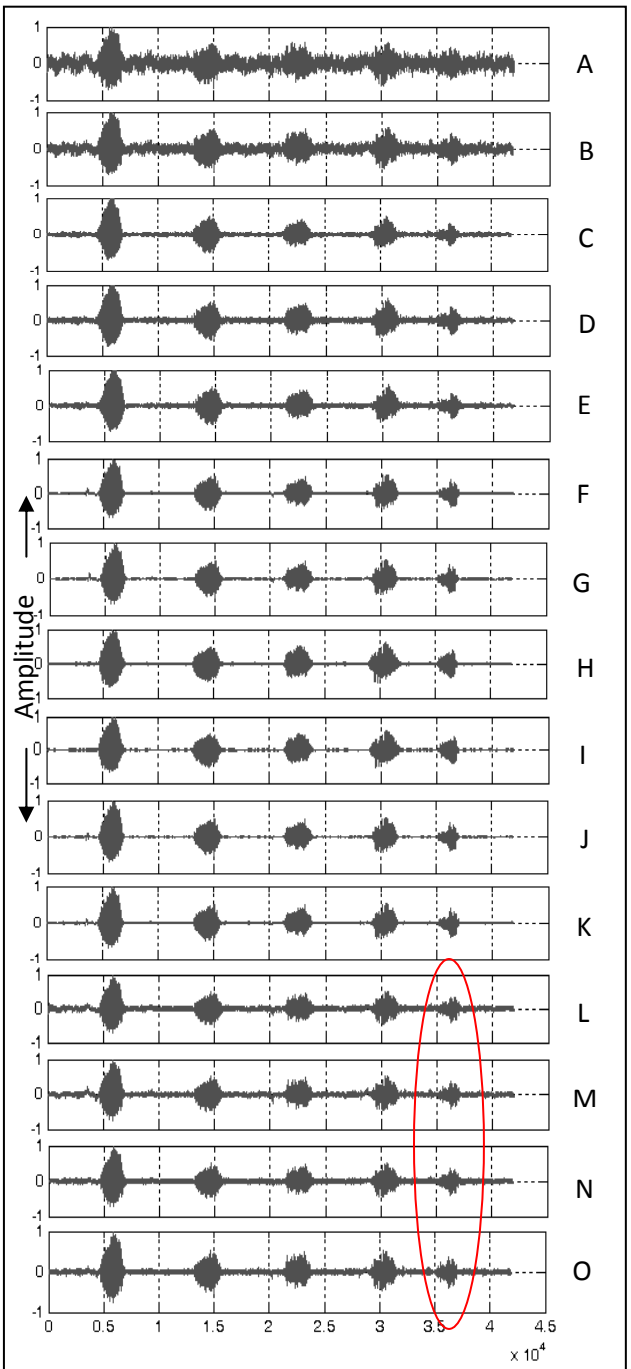


Fig.10. Pink noise superimposed results (DCT) (0dB)-Speech-2: A) Noisy B) ML C) MSS D) PSS E) WF F) MMSE1 G) MMSE 1(SPU) H) MMSE-Log I) MMSE-Log (SPU) J) MMSE1-HWT K) MMSE1-HWT (SPU) L) MMSE2 M) MMSE2 (SPU) N) MMSE2-HWT O) MMSE2-HWT (SPU)

Enhancement Methods using DFT transform	Speech1: “This changes formula to an Equation”				Speech2: “AEIOU”			
	White		Pink		White		Pink	
	MSE	SNR	MSE	SNR	MSE	SNR	MSE	SNR
ML	0.363	4.40	0.432	3.64	0.289	5.39	0.273	5.62
MSS	0.348	4.57	0.341	4.66	0.090	10.39	0.142	8.45
WF	0.316	4.99	0.301	5.21	0.081	10.85	0.122	9.10
PSS	0.302	5.19	0.283	5.48	0.082	10.47	0.111	9.52
MMSE1	0.275	**5.59**	0.260	**5.84**	0.048	**13.12**	0.096	**10.16**
MMSE1 (SPU)	0.276	**5.58**	0.263	**5.78**	0.047	**13.20**	0.095	**10.17**
MMSE-Log	0.253	5.96	0.269	5.69	0.047	13.20	0.095	10.19
MMSE-Log (SPU)	0.248	6.05	0.282	5.48	0.049	13.08	0.096	10.14
MMSE1-HWT	0.253	**5.67**	0.285	**5.45**	0.046	**13.37**	0.086	**10.61**
MMSE1HWT(SPU)	0.211	**5.97**	0.277	**5.57**	0.044	**13.52**	0.082	**10.82**
MMSE2	0.291	**5.35**	0.250	**6.01**	0.082	**10.82**	0.129	**8.89**
MMSE2 (SPU)	0.270	**5.67**	0.240	**6.28**	0.076	**11.47**	0.122	**9.11**
MMSE2-HWT	0.276	**5.58**	0.221	**6.55**	0.091	**11.45**	0.117	**9.31**
MMSE2HWT(SPU)	0.258	**5.87**	0.211	**6.75**	0.082	**10.84**	0.111	**9.53**

Table.1. Comparison of DFT results for various speech enhancement methods.

Enhancement Methods using DCT transform	Speech1: “This changes formula to an Equation”				Speech2: “AEIOU”			
	White		Pink		White		Pink	
	MSE	SNR	MSE	SNR	MSE	SNR	MSE	SNR
ML	0.469	3.28	0.489	3.10	0.328	4.83	0.306	5.13
MSS	0.356	4.48	0.362	4.40	0.096	10.17	0.130	8.85
WF	0.347	4.59	0.330	4.81	0.109	9.62	0.120	9.18
PSS	0.348	4.57	0.326	4.85	0.133	8.73	0.124	9.06
MMSE1	0.249	**6.02**	0.234	**6.30**	0.051	**12.85**	0.099	**10.03**
MMSE1 (SPU)	0.254	**5.95**	0.238	**6.22**	0.052	**12.80**	0.102	**9.89**
MMSE-Log	0.246	6.07	0.246	6.07	0.052	12.80	0.099	10.02
MMSE-Log (SPU)	0.256	5.91	0.256	5.91	0.060	12.21	0.113	9.44
MMSE1-HWT	0.238	**5.83**	0.240	**6.19**	0.050	**12.94**	0.099	**10.0**
MMSE1HWT(SPU)	0.229	**5.98**	0.240	**6.37**	0.050	**12.93**	0.101	**9.92**
MMSE2	0.348	**4.57**	0.270	**5.67**	0.145	**8.37**	0.172	**7.62**
MMSE2 (SPU)	0.301	**5.20**	0.274	**5.61**	0.137	**8.62**	0.184	**7.34**
MMSE2-HWT	0.299	**4.77**	0.230	**6.37**	0.136	**8.65**	0.150	**8.21**
MMSE2HWT(SPU)	0.248	**5.21**	0.220	**6.56**	0.125	**9.02**	0.153	**8.13**

Table.2. Comparison of DCT results for various speech enhancement methods.

<table>
<tr><th rowspan="2">Methods
Constant Parameters</th><th colspan="5">Speech1</th><th colspan="5">Speech2</th></tr>
<tr><th>Seg Length</th><th>Smooth Factor</th><th>Resln. Time (ms)</th><th>Sample freq.</th><th>BW</th><th>Seg Length</th><th>Smooth Factor</th><th>Resln. Time (ms)</th><th>Sample freq.</th><th>BW</th></tr>
<tr><td>ML</td><td rowspan="4">512</td><td rowspan="4">0.98</td><td rowspan="4">1.9</td><td rowspan="4">31.25</td><td rowspan="4">15.62</td><td rowspan="4">512</td><td rowspan="4">0.98</td><td rowspan="4">1.9</td><td rowspan="4">39.06</td><td rowspan="4">19.53</td></tr>
<tr><td>MSS</td></tr>
<tr><td>PSS</td></tr>
<tr><td>WF</td></tr>
<tr><td>MMSE1</td><td rowspan="3">256</td><td rowspan="3">0.98</td><td rowspan="3">3.9</td><td rowspan="3">62.5</td><td rowspan="3">31.25</td><td rowspan="3">256</td><td rowspan="3">0.98</td><td rowspan="3">3.9</td><td rowspan="3">78.12</td><td rowspan="3">39.06</td></tr>
<tr><td>MMSE-Log</td></tr>
<tr><td>MMSE1-HWT-DCT</td></tr>
<tr><td>MMSE1-HWT-DFT</td><td>512</td><td>0.98</td><td>1.9</td><td>31.25</td><td>15.62</td><td>512</td><td>0.98</td><td>1.9</td><td>39.06</td><td>19.53</td></tr>
<tr><td>MMSE2</td><td rowspan="2">128</td><td rowspan="2">0.98</td><td rowspan="2">7.8</td><td rowspan="2">125</td><td rowspan="2">62.5</td><td rowspan="2">128</td><td rowspan="2">0.98</td><td rowspan="2">7.8</td><td rowspan="2">156.2</td><td rowspan="2">78.12</td></tr>
<tr><td>MMSE2-HWT</td></tr>
</table>

Table.3. **Parameter constants used for simulation**

NOTE: A gain of 15 is set for noise estimate to be subtracted from noisy speech for methods ML, MSS, PSS, and WF.

4.0. Description of simulation results:

In this project work, we have carried out Research and MATLAB simulations of various speech enhancement algorithms. The methods used to study include following a) Maximum-Likelihood, Eqn.29 (ML), 2) Magnitude Spectral Subtraction, Eqn.11 (MSS), 3) Power Spectral Subtraction, Eqn.33 (PSS), 4) Wiener Filter (WF), Eqn.36 and 5) MMSE methods based on Ephraim-Mallah, MMSE1, (using Eqn.59 and Eqn. 63) and MMSE2, (using Eqn.58b and 63). In this work most of the theory, mathematical derivations and references is taken from the book *speech enhancement theory and Practice* by Philipos C. Loizou. Simulation work is carried out for two kinds of noise white Gaussian and Pink (filtered white noise, which is also characteristic of the aviation noise) noise using DFT/DCT transforms. The speech source files are taken from TIMIT database and noise files from NOISEX-92 database. The various constant parameters used for entire simulation for methods under discussion listed in Table.3. Here sub band are realized using harmonic wavelet transform.

In Fig.3, Fig.4, shown are the result of DFT based speech enhancement methods to noisy speech corrupted by white noise at 0dB level for speech-1 (sampled at 8KHz) and speech-2 (sampled at 10KHz) respectively. A gain of 15 (only for methods ML, MSS, WF, PSS) for noise estimate to be subtracted from noisy speech is fixed based on experimental observations. The MSS method suppression at low-SNR is slightly lesser than PSS and WF. PSS method produced better results in terms of O/P-SNR but with residual musical noise. WF on the other hand showed O/P-SNR slightly lesser than PSS but with reduced musical noise due to its high suppression factor.

MMSE1 methods, Fig.3, (G) and (H) (single and sub band) produced same audible effects compared to that of WF but with lesser speech distortion and musical noise, however produced highest O/P-SNR for both white and pink noise among all the methods as can be seen from table.1 and table.2. MMSE2 methods even though maintained certain flooring (single and sub band) removed musical noise to the maximum extent compared to all methods including MMSE1 also improved O/P-SNR but lesser than MMSE1. MMSE2-HWT under Speech Presence Uncertainty conditions further improved O/P-SNR and better speech intelligibility.

Fig.3, Methods (G), (H), (I), (J) waveforms though look similar but produced increase in O/P-SNR and improved audible characteristics. In (I) and (J) logarithmic MMSE method proved better in terms of O/P-SNR and further reduced back ground noise but with less musical noise in comparison (A)-(H).

Fig.3, Methods (K) and (L) are sub band realization of methods (G) and (H) and resulted best performance by making use of short time-frequency localizations, as tabulated in Table.1 and Table.2. Fig.3. (O) and (P) are the sub band realization of methods (M), (N) again shown better performance in terms of O/P-SNR and reduced musical noise and sound speech more or less natural.

It is worth noting that in figures, marked red indicates the O/P-SNR improvement in the case of MMSE2 methods. The regions marked clearly show the dense of speech energy with envelopes comparable to clean speech.

Fig.4. shows results for second speech under consideration at 10 KHz. Here again the methods behave similar and same explanation holds as for speech1 results.

In Fig.5, Fig.6, shown is the simulation carried out for pink noise. Table.1. and Table.2 though reflect good O/P SNR values especially for speech1 (8KHz) but distortion in speech is very noticeable and also suffered by more audible musical noise. The increase in O/P-SNR is observed due to lower O/P-noise, obtained by subtracting clean speech from enhanced speech also due to correlation of pink noise samples with speech samples than white noise makes this subtraction reduce O/P noise and result in increase of O/P SNR. Performance behavior of individual methods is similar to that of white noise as can be seen from Table.1 and Table.2.

In Fig.7, Fig.8, shown are the result of DCT based speech enhancement methods to noisy speech corrupted by white noise at 0dB level for speech-1 (sampled at 8KHz) and speech-2 (sampled at 10KHz) respectively. The results for pink noise are shown in Fig.9, Fig.10 for two speech signals. These results in comparison to DFT counterparts retained most of the speech envelope with less distortion except for the methods MMSE1-HWT that distorted some low amplitude signal. Further with DFT enhanced speech signal heard more natural with negligible musical noise, however DCT even though sound natural but had some musical noise but far less in comparison to simple method of speech enhancement (A-F). It is also noted that the O/P SNR of MMSE1 and MMSE2 methods in single and sub band behave almost similar with slight increase with sub band, however with DFT a noticeable difference in O/P SNR values with single and sub band approach.

To summarize, the overall simulation results for methods under consideration are shown Table.1 and Table.2 for DFT and DCT respectively. It can be observed that DFT performance is better than DCT. In comparison to MMSE1 and MMSE2, MMSE1 performed better than MMSE2 in terms of O/P-SNR values however with background musical noise but with MMSE2 residual background noise even though present was not modified and didn't produce musical noise effect, however for pink noise certain amount of musical noise can be heard in all methods. Further single grouping of DFT/DCT coefficients for MMSE1, MMSE2 doesn't produce good results in comparison to sub band coefficients for the same method using Harmonic Wavelet Transform (HWT) methods.

5.0. Conclusion:

In this work various speech enhancement algorithms performance superimposed with white and pink noise at 0dB level is considered to enrich CVR analysis capability. Here pink noise is considered because it characterizes aviation noise. Use of combination of present algorithms along with simple and computationally efficient harmonic wavelet transform, it is found that improvement in performance measures like O/P-SNR,MSE and audibility which are essential for aviation applications like CVR analysis. Further this project has scope for research and development for on-board and real-time speech communication enhancement for aviation industries.

References

[1]. Yuan Yao, Benshun Yi, Yongquong Yao, ”Speech Enhancement under aviation noise”, IEEE,2006.

[2]. T.Gulzow, T.Ludwig, U.heute. *“Spectral-subtraction speech enhancement in multirate systems with and without non-uniform and adaptive bandwidths.”* Signal Processing. Vol.83. 2003. Page. No. 1613-1631.

[3]. Jorg Enders, Weihua Geng, Peijun Li, and Michael W. Frazier, *“The shift-invariance discrete wavelet transform and application to speech waveform analysis.”* Acoustical Society of America. April 2005. Page. No. 2122-2133.

[4]. Y.Ephraim and D.Malah, “ Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP, no.6, pp. 1109-1121, Dec. 1984.

[5]. Boll, S.F. (1979), Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans.Acoust. Speech Signal Process, 27(2), 113-120.

[6]. McAulay, R.J. and Malpass, M.L. (1980), Speech enhancement using a soft-decision noise suppression filter, IEEE Trans.Acoust. Speech Signal Process., 28, 137-145.

[7]. Kay.S.(1993), Fundamentals of Statistical Signal Processing: Estimation Theory, Upper Saddle River,NJ: Prentice Hall.

[8]. Porter,J and Boll,S.F(1984), Optimal estimators for spectral restoration of noisy speech,proc. IEEE Int.Conf.Acoust. Speech Signal Process., pp.18A.2.1-18A.2.4.

[9]. Martin, R (2002), Speech enhancement using a MMSE short time spectral estimation with Gamma distributed speech priors, Int.Conf. Speech Acoust. Signal Process, PP.253-256.

[10]. Lotter,T. and Vary,P(2005), Speech enhancement by maximum a posteriori spectral amplitude estimation using a supergaussian speech model, EURASIP J. Appl.Signal Process, 2005(7),1110-1126.

[11]. Pearlman,W and Gray,R.(1978), Source coding of the discrete Fourier transform, IEEE Trans.Inform. Theory, 24(6),683-692.

[12]. Papoulis, A(1984), Probability, Random variables and Stochastic Processes, 2nd ed.New York: McGraw-Hill.

[13].Brillinger,D(2001), Time Series: Data Analysis and Theory, Philadelphia, PA:SIAM

[14]. Cohen,I(2005),Related statistical model for speech enhancement and a priori SNR estimation, IEEE Trans.Speech Audio Process., 13(5),870-881.

[15]. Wolfe,P., Godsill,S., and Ng, W.-J.(2004), Bayesian variable selection and regularization for time-frequency surface estimation, J.R.stat.Soc.B, 66,575-589.

[16]. Cappe,O(1994), Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans.Speech Audio Process.,2(2), 346-349.

[17]. Griffin,D and Lim,J (1988), Multiband excitation vocoder, IEEE Trans.Acoust.Speech Signal Process., 36(8),1223-1235.

[18]. Makhoul, J., Viswanathan, R., Schwartz,R., and Huggins,A.(1978), A mixed source excitation model for speech compression and synthesis, Proc. IEEE Int.Conf.Acoust.Speech Signal Process.,pp.163-166.

[19]. Soon,I., Koh,S., and Yeo,C.(1999), Improved noise suppression filter using self-adaptive estimator of probability of speech absence, Signal Process., 75, 151-159.

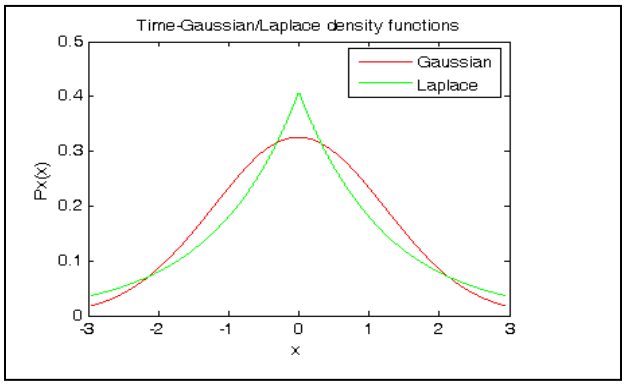
[20]. I.Y.Soon, S.N.Koh, Low distortion speech enhancement. IEE Proc. Vis. Image Signal Process, Vol.147, No.3, June 2000.

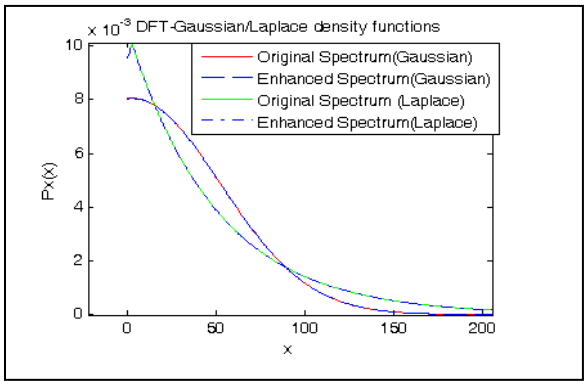
[21]. Arkady Bron, Wavelet based denoising of speech. Technion-Israel Institute of Technology.

[22]. Bin Chen, Philipos C.Loizou., A Laplacian-based MMSE estimator for speech enhancement, speech communication 49 (2007),2006. PP. 134-143.

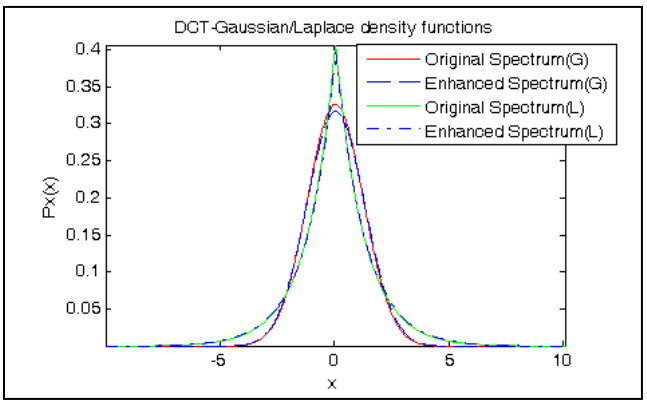
[23]. Zou Xia, Zhang Xiongwei., Speech enhancement using an MMSE short time DCT coefficients estimator with supergaussian speech modeling, Journal of Electronics (China), May 2007, Vol.24, No.3.

Appendix-I

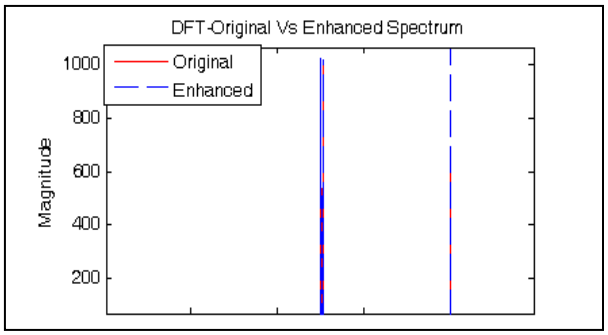


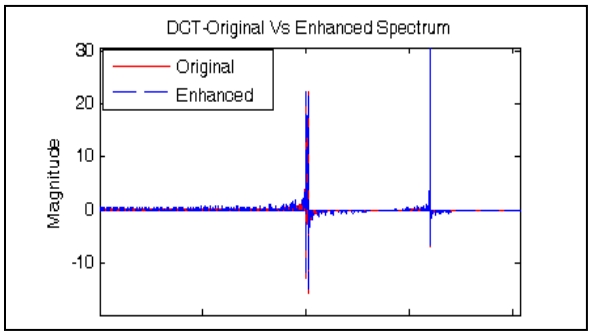


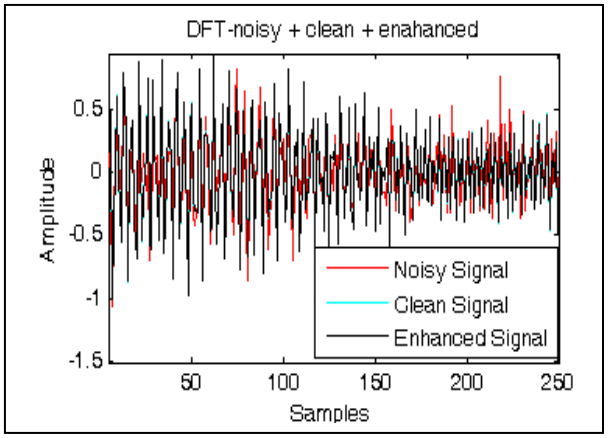
MSE Vs PDF Sinusoid Signal	DFT-MSE	DCT-MSE
Gaussian	**1.37e-6**	0.3316
Laplace	**8.12e-7**	0.4181



MSE Vs PDF Speech Signal	DFT-MSE	DCT-MSE
Gaussian	**0.1305**	0.2599
Laplace	**0.1875**	0.3386







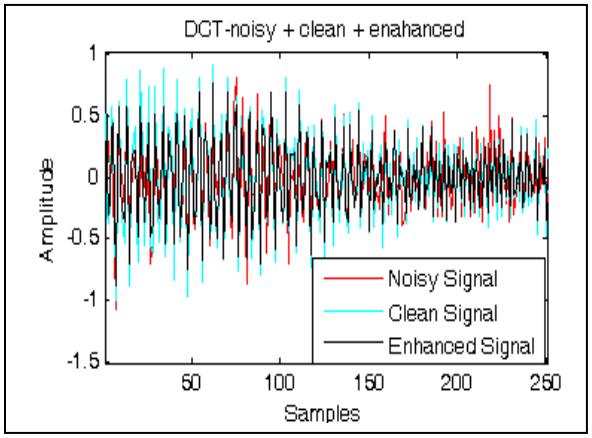


Fig.11. probability distribution comparison of DFT and DCT Coefficients

Description:

It can be argued that the DCT based performance is better than the DFT methods. In this context it is important to consider that simply changing transform from DFT to DCT for analysis is not sufficient. It is seen that performance of any speech enhancement algorithm depends on the kind of computation involved in transform domain using DFT and DCT. In case of spectral subtraction using DCT, the performance measure mean square error (MSE) when computed in transform domain over segments shows better than DFT this observation is in coincidence with the DCT properties like better transform domain resolution, speech spectral energy compaction, smooth truncations and non-correlated phase of the speech [20]. However the MSE between the overall enhanced signal and clean signal showed DFT performance is better than DCT. This contrasting behavior between DFT and DCT is further explored considering the distributions of transform coefficients and found that the DFT spectrum probability density function (pdf) have lesser MSE between original and enhanced spectrum than that of DCT as indicated in MSE VS PDF table in Fig.11, for sinusoid and speech signal. In the work [21] related to denoising using DFT/DCT have mentioned that given only the noisy observations and estimated noise squared-spectral components, the phase of clean speech cannot be anymore exactly reconstructed using real-valued transform. Further it is noted that with DCT noise estimation (which is very essential in any speech enhancement algorithm that characterizes the background noise) is poor in terms of much lower amplitude distribution of coefficients which do not subtract effectively from noisy coefficients effectively. Authors [22] have found that noisy speech and clean speech transform coefficients follow Gaussian and Laplace distribution respectively. These statistical based algorithms shows that Laplacian-distribution has yielded better results than that of Gaussian distribution. The work [22] has been further explored with DCT transform [23] to see better reduction in residual noise. These works do suggest that the kind of distribution of DFT/DCT matters in speech enhancement.

Distribution List

1. Director, NAL
2. Head, ALD (*for official records*)
3 Sponsor
4. M. Shivamurti (*for internal reference - 2 copies*)

National Aerospace Laboratories

Studies in Signal Processing Techniques for Speech Enhancement: A comparative study

M. Shivamurti and Dr.S.V.Narasimhan

Aerospace Electronics & Systems Division

Project Document AL 0919

August-2009

Bangaluru-560 017, India