SUPPRESSION OF ADDITIVE BACKGROUND NOISE FROM DEGRADED SPEECH SIGNALS BY SPECTRAL PROCESSING

M. Khandelwal,
Student,
Electronics and Telecommunication Engineering,
AISSMS's Institute of Information Technology,
Pune.

Abstract: Speech enhancement with an objective to suppress noise can be a complex process and cause distortions or loss of information while processing the degraded speech signal. Enhancement with the help of Power spectrums (such as SNR) is used for processing speech signals in Spectral domain. The basic and the oldest algorithms proposed in this domain were the Spectral Subtraction and Mean Minimum Square Error. The present paper addresses detailed process of Spectral Subtraction and MMSE algorithms with a mathematical and practical approach. Spectral Subtraction is simple for implementation but suffers from distortions which are sometimes generated due to uncontrolled processing of the signals. However, this distortion can be reduced by making different approach to eliminate the noise. The MMSE algorithms are basically statistical estimators in which different assumptions and estimation are made required to suppress the noise present in the noisy speech signals. The consistency, drawbacks and results are too analyzed in the paper.

Index Terms – Speech Enhancement, Spectral Processing, Spectral Subtraction, Mean Minimum Square Error, Maximum a posteriori, Zero Cross-terms.

1. INTRODUCTION

Speech signals are basically acknowledged by its quality and intelligibility scale. These two characteristics can get affected due to various factors (such as additive background noise) and hence degrade the original speech signal. The reception of such degraded speech signals can cause misinterpretation or loss of the information which was carried in the original speech signal. Enhancement of the degraded speech signals and trying to recover the original signal can be achieved by suppression of noise. Speech Enhancement is basically improving the quality and maintaining the intelligibility of the speech signals which is degraded by additive noise [1]. It is possible to remove the noise completely but at the cost of introducing speech distortions in the clean speech signal which may impair speech intelligibility. The noise can be correlated or uncorrelated with the clean speech signal. Various solutions for solving the general problems of Speech Enhancement depends on the characteristics of the noise source or interference, the relationship of noise with the clean speech signal and the hardware used. Yet, due to complexities there is a challenge to produce a completely clean speech signal. Speech enhancement has a tradeoff between reduction of degrading components and distortions introduced. Thus, there is need to develop algorithms which would suppress the noise without introducing any distortions in the signal. Spectral Processing involves estimation and elimination of components causing degradation. Spectral Subtraction is one of the earliest algorithms developed for noise suppression. It is simple to implement and can be most effective if the assumptions made are accurate. Assuming that there is an addition of noise, the estimation clean speech spectrum is done by subtracting the noise spectrum from the noisy spectrum. The noise spectrum has to be estimated and updated when there is no speech signal but only noise.

The MMSE algorithm uses models for the distribution of spectral components of the speech and noise signals [3]. In the earlier method discussed no specific assumptions for the distribution of the power components of either speech or noise. The MMSE - Short Time Spectral Amplitude estimator minimizes the mean square error from the degraded speech signals by comparison with clean signal and the noise signals. Also, it is assumed that each of the Fourier expansion coefficients if the speech and of the noise process can be modeled as independent, zero-mean, Gaussian random variables.

2. BACKGROUND & LITERATURE SURVEY

Let y(n) = x(n) + d(n), where the x(n) is the original desired(clean) signal and d(n) is the noise signal and hence y(n) is the degraded signal. The short-time Fourier Transform of the above equation can be given as,

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{1.1}$$

For $\omega_k = 2\Pi\, k/N$ and k = 0,1, 2, …., N-1, where N is the frame length in samples. We multiply the equation with its conjugate and we get,

$$\begin{aligned} |Y(\omega_k)|^2 &= |X(\omega_k)|^2 + |D(\omega_k)|^2 + X(\omega_k).D^*(\omega_k) + X^*(\omega_k).D(\omega_k) \\ &= |X(\omega_k)| + |D(\omega_k)| + 2|X(\omega_k)| + |D(\omega_k)|\cos(\theta_X(k) - \theta_D(k)) \end{aligned} \tag{1.2}$$

The terms $X(\omega_k)$. $D^*(\omega_k)$ and $X^*(\omega_k)$. $D(\omega_k)$ are approximated as $E|D(\omega_k)|^2$, $E\{X*(\omega_k).\ D(\omega_k)\}$ and $E\{X(\omega_k).\ D*(\omega_k)\}$, as they cannot be generated directly. E[.] denotes the expectation operator. Estimation of $E\{|D(\omega_k)|^2\}$ is carried out during the non-speech activity and denoted by $|\hat{D}(\omega_k)|^2$. Assuming that d(n) is zero mean and uncorrelated with the original signal then the terms $X(\omega_k)$. $D^*(\omega_k)$ and $X^*(\omega_k)$. $D(\omega_k)$ are eliminated. making these changes the above equation changes to

$$|X(\omega_k)|^2 = |Y(\omega_k)|^2 - |D(\omega_k)|^2 \quad (1.3)$$

This is so called as power subtraction equation.

2.1 SPECTRAL SUBTRACTION

In Spectral Subtraction, subtraction of average magnitude of noise from the spectrum of noisy speech is done to estimate the magnitude of the enhanced spectrum [2]. The Noise is assumed to be addition to the speech signal and is not correlated. Also, assumption is to be made that the noise source is locally stationary, by which the noise characteristics computed during the speech pauses are a good approximation to the noise characteristics.

$$|\hat{X}(\omega)|^p = |Y(\omega)|^p - |\hat{D}(\omega)|^p \quad (1.4)$$

where p is the power exponent, with p=1, provides the original magnitude spectral subtraction, and p= 2 holds for the power spectral subtraction algorithm. In the equation 2.1 some negative values are present in the obtained enhanced spectrum because of errors in estimating the noise spectrum. Hence to ensure a non-negative magnitude spectrum these values are half-wave rectified. It creates small, isolated peaks in the spectrum which are negative values created while processing occurring at random frequency locations in each of the frames. When the spectrum is converted in the time domain, these peaks sound similar to the tones with the frequencies that change randomly from frame to frame, i.e. tones that that are keep turning on and off simultaneously while the analysis. This type of noise is referred as musical noise [6, 7, 8]. Compared to the original distortion caused by the background noise, the musical noise is more annoying to the listeners. Boll [4] proposed modifications such as magnitude averaging, residual noise reduction and additional signal attenuation during speech activity to reduce the effect of musical noise. Berouti [6] suggested a method to reduce the musical noise by an estimation subtracting from the noise spectrum, while preventing the resultant spectral components from going below a minimum value.

$$|X(\omega)|^2 = |Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2, \quad \text{if } |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)| \quad (1.5)$$

also,

$$|X(\omega)|^2 = \beta|\hat{D}(k)|^2, \quad \text{otherwise.} \quad (1.6)$$

Where α is the over subtraction factor ($\alpha \geq 1$) and β ($0 < \beta << 1$) is a spectral floor parameter. The subtraction of noise spectrum from the noisy speech spectrum, some peaks are observed in the spectrum. Some are broadband whereas others are narrow-band, seen as spikes in the spectrum. If there is an over-subtraction i.e. $\alpha > 1$, the amplitude of the broadband peaks is reduced and, in some cases, eliminates them. this not sufficient because the deep valleys surrounding the peaks remain in the spectrum. Spectral Flooring is used to fill in the spectral valleys and mask the remaining peaks by the neighboring spectral components. The parameter β determines the effect of residual noise and the amount of perceived musical noise. For large values of β the residual noise gets audible but the musical noise is absent. Conversely, for a small value the musical noise will be more annoying and residual noise will be reduced. The parameter α controls the amount of the speech spectrum distortion. If α is too large the distortion is so severe that the intelligibility gets affected. Berouti [4] suggested that α should vary in every frame according to:

$$\alpha = \alpha_0 - 3/20(SNR), \quad -5dB \leq SNR \leq 20dB \quad (1.7)$$

where α_0 is the desired value of α at 0 dB SNR. Here, SNR is computed as the ratio of the noisy speech power to the estimated noise power. In general, higher the amount of over subtraction factor is, the stronger components with a low SNR get attenuated. This prevents musical noise. But too strong over subtraction factor will suppress too many components. Therefore, the value of α has to be carefully chosen in order to prevent both the musical noise and signal distortion [8]. The introduction of spectral floor β prevents the subtraction of spectral components of the enhanced speech spectrum falling below the predefined value.

2.2 MINIMUM MEAN SQUARE ERROR ESTIMATOR

The MMSE technique uses models for the distribution of spectral components of the speech and noise signals [3]. The MMSE - Short Time Spectral Amplitude estimator minimizes the mean square error from the degraded speech signals by comparison with clean signal and the noise signals. As proposed in Ephraim and Malah [3], in MMSE estimator an assumption is made that real and the imaginary parts of the clean DFT coefficients can be modeled by a Gaussian distribution. This assumption might only apply for the noise DFT coefficients, it does not hold for the speech DFT coefficients. All methods based on MMSE require the estimate of the priori SNR. A decision-direct estimation is the approach taken to compute a priori SNR [3]. Equation one can also be written in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \quad (1.8)$$

where Y_k, X_k, D_k and $\theta_y(k)$, $\theta_x(k)$, $\theta_d(k)$ are denoting the magnitudes and the phases of degraded speech, desired(clean) speech and noise, respectively. As proposed in [3] the following is the MMSE-short-time power spectrum

$$\begin{aligned} \hat{X}_k^2 &= E\{X_k^2 \mid Y(\omega_k)\} \\ &= \int_0^\infty X_k^2 f_{\{X_k\}}(X_k \mid Y(\omega_k))\, dX \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k}\right)\left(+\frac{\xi_k}{1+\xi_k}\right) Y_k^2 \end{aligned} \quad (1.9)$$

Where

$$\xi_k = \frac{\sigma_x^2(k)}{\sigma_d^2(k)}\ ,\ \gamma_k = \frac{Y_k^2}{\sigma_d^2(k)}$$

(1.10)

$$\sigma_x^2(k) = E\{X_k^2\},\ \sigma_d^2(k) = E\{D_k^2\} \quad (1.11)$$

where ξ_k and γ_k are the assumptions made for the estimator called as a priori and a posteriori SNRs, respectively. The Maximum a posteriori estimator and the MMSE estimator were based on the following density functions:

$$f_{\{x_k\}}\left(X_k\,|Y(\omega_k)\right) = \frac{X_k}{\sigma_k^2}\exp\left(-\frac{X_k^2+s_k^2}{2\sigma_k^2}\right)I_0\left(\frac{X_k s_k}{\sigma_k^2}\right) \quad (1.12)$$

where

$$\frac{1}{\lambda\prime(k)} = \frac{1}{\sigma_k^2(k)} + \frac{1}{\sigma_d^2(k)} \quad (1.13)$$

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \quad (1.14)$$

Where

$$\sigma_k^2 = \frac{\lambda\prime(k)}{2},\ s_k^2 = \frac{v_k}{\lambda\prime(k)} \quad (1.15)$$

and the $I_0(.)$ is the first kind of modified Bessel function of zeroth order. The suppression is low for a priori SNR. Loizou [2] proposed short-time power-spectrum estimators comparative to the above derived one. The assumptions made are in statistical sense but can give simple noise reduction methods. x(n) and d(n) are uncorrelated stationary random processes, the power spectrum of the noise-corrupt signal, $P_y(\omega)$ is given as,

$$P_y(\omega) = P_x(\omega) + P_d(\omega) \quad (1.16)$$

Two assumptions are made for the proposed derivation of estimators in Loizou [2]. From the above equation the first assumption is made by approximations of the power spectrum, a sample estimate and can be written as,

$$Y_k^2 \approx X_k^2 + D_k^2. \quad (1.17)$$

The real and imaginary parts the discrete Fourier transform (DFT) are designed with equal variance and independent Gaussian random variables Ephraim [3]. The probability density function of X_k^2 is exponential and is given by

$$f_{\{X_k^2\}}(X_k^2) = \frac{1}{\sigma_x^2(k)}\ e^{\left\{--\frac{X_k^2}{\sigma_X^2(k)}\right\}} \quad (1.18)$$

$$f_{\{D_k^2\}}(D_k^2) = \frac{1}{\sigma_d^2(k)}\ e^{\left\{--\frac{D_k^2}{\sigma_d^2(k)}\right\}} \quad (1.19)$$

Also, from Eq. (1.11),

$$f_{\{X_k^2\}}(X_k^2|Y_k^2) = \frac{f_{Y_k^2}\left(Y_k^2|X_k^2\right)f_{X_k^2}(X_k^2)}{f_{Y_k^2}\left(Y_k^2\right)}$$

$$= \psi_k\ e^{\left\{-\frac{x_k^2}{\lambda(k)}\right\}},\quad if\ \sigma_x^2(k) \neq \sigma_d^2(k)$$

$$= \frac{1}{Y_k^2}, if\ \sigma_x^2(k) = \sigma_d^2(k) \quad (1.20)$$

$$\frac{1}{\lambda(k)} = \frac{1}{\sigma_k^2} + \frac{1}{\sigma_d^2(k)}, if\ \sigma_x^2(k) \neq \sigma_d^2(k) \quad (1.21)$$

$$\psi_k \equiv \frac{1}{\lambda(k)\left\{1-exp\left(-\frac{Y_k^2}{\lambda(k)}\right)\right\}} \quad (1.22)$$

using the Eq. (1.17) - (1.20) a MMSE estimator can be obtained by computing the mean of the posteriori equation in Eq. (120)

$$\widehat{X_k^2} = E\{X_k^2\,|Y_k^2\}$$

$$= \int_0^{Y_k^2} X_k^2 f_{\{X_k\}}(X_k|Y_k^2)\,dX_k^2$$

$$= \left(\frac{1}{v_k^2}\right) - \frac{1}{\left\{e^{\{v_k\}}-1\right\}}, if\ \sigma_x^2(k) \neq \sigma_d^2(k)$$

$$= \frac{1}{2}Y_k^2, if\ \sigma_x^2(k) = \sigma_d^2(k) \quad (1.23)$$

Where v is given as,

$$v_k \equiv \frac{1-\xi_k}{\xi_k}\ \gamma_k \quad (1.24)$$

and the gain function is given as

$$G_{\{MMSE\}}(\xi_k,\gamma_k) = \left(\frac{1}{v_k^2}\right) - \frac{1}{e^{v_k}-1}, if\ \sigma_x^2(k) \neq \sigma_d^2(k)$$

$$= \frac{1^{\frac{1}{2}}}{2},\ if\ \sigma_x^2(k) = \sigma_d^2(k) \quad (1.25)$$

The MMSE estimator is computed by mean of the posteriori density given condition as the noise-corrupt magnitude is squared (Y_k^2), instead of ($Y(\omega_k)$) and we can observe that it is different from the one derived in Eq. (1.9). This estimator is called as MMSE-Spectrum Power estimator based on Zero Cross-terms assumptions. The gain function of the MMSE-SPZC relies on ξ_k and γ_k parameters. The MMSE-SPZC provides more suppression than MMSE-SP estimator derived in Eq. (1.9) for small values of ξ_k and for larger values of γ_k. Hence, we can expect that MMSE-SPZC can reduce the generated noise while speech processing in the MMSE-SP that is the Residual Noise.

In a posterior density function when ξ changes its sign while expressed in dB, the direction of density changes. This gives us the maximization of the equation and is given as follows:

$$\widehat{X_k^2} = \arg\max f_{\{X_k^2\}}(X_k^2|Y_k^2)$$
$$= Y_k^2,\ if\ \sigma_x^2(k) \geq \sigma_d^2(k)$$
$$= 0,\ if\ \sigma_x^2(k) < \sigma_d^2(k) \quad (1.26)$$

The gain function of this estimator is given by,

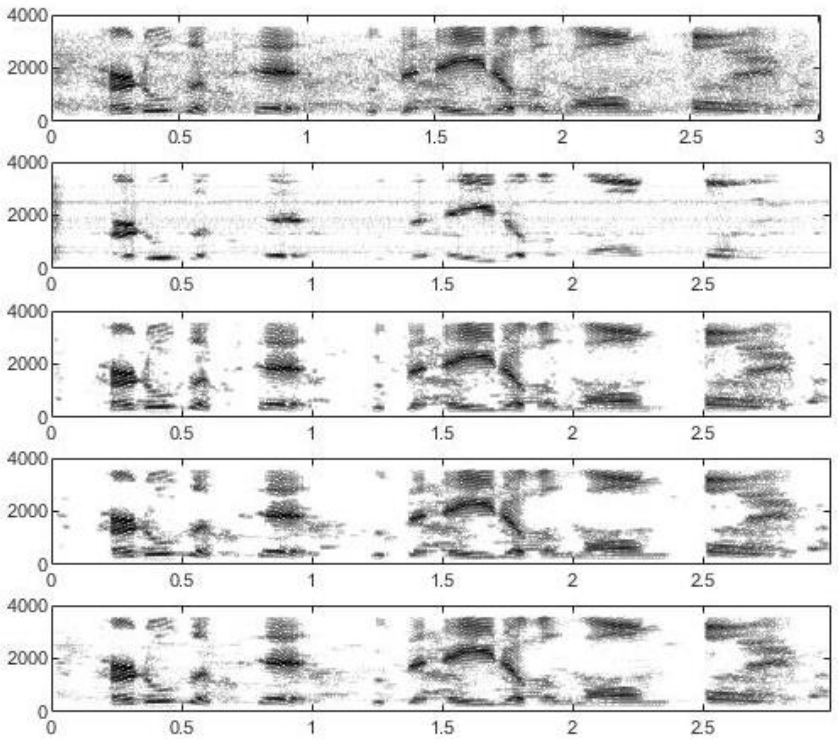
$$G_{\{MAP\}}(k) = 1,\ if\ \sigma_x^2(k) \geq \sigma_d^2(k)$$
$$= 0,\ if\ \sigma_x^2(k) < \sigma_d^2(k) \quad (1.27)$$

Using Eq. (1.10) the above equation can also be written as

$$G_{\{MAP\}}(\xi_k) = 1, \quad if\ \xi_k \geq 1$$
$$= 0,\ if\ \xi_k < 1 \quad (1.28)$$

from the above equation of MAP, we can observe that the original speech gets limited in the range of the noisy speech.

3. RESULTS



Wideband Spectrograms. (a)Degraded Noisy Speech. (b)Enhanced Speech from MMSE-SP. (c)Enhanced Speech from Spectral Subtraction. (d)Enhanced Speech from MAP. (e)Enhanced Speech from MMSE-SPZC

As seen in the Spectral Subtraction, we assume that the noise is additive and by subtracting the estimate of noise spectrum from the noisy speech spectrum we can obtain a clean speech spectrum which might stand true if the noise is uncorrelated and has the same phase. Also, the subtraction of the spectrum has to be done carefully, if too much gets subtracted, we might lose some information from the speech signal. And if less gets subtracted the existence of noise will have distortions. Also, while processing the speech signals, the spectrum is converted into the time domain, the negative values create small isolated peaks occurring randomly and sound similar as tones that are turned on and off, these are referred as musical noise. If we according to the method discussed in the paper the generation of residual noise in the Spectral Subtraction is less as proposed in method by Boll [4].

MMSE estimators are more consistent and perform well than the Spectral Subtraction algorithm. Among the estimators mentioned i.e. MMSE-SP, SPZC, MAP we can clearly observe the MAP estimator performs better than others as it depends on maximization a posterior function. MMSE-SPZC estimator makes the assumptions that the cross-terms are reduced to zero, but when the priori SNR becomes zero it provides an attenuation constantly of -3dB which is independent of the value of posteriori SNR. MMSE-SP estimator which is dependent on the posteriori density function, suffers from the residual noise as the estimation of noise is more approximated. All the MMSE estimators encounter with some distortions while processing speech referred as residual noise.

4. REFERENCES

1. P. Krishnamoorthy, S.R.M. Prasanna," Speech Enhancement by Temporal and Spectral Processing", IEEE Transactions on Audio, Speech, and Language Processing, 2009, Volume 17, Issue 2.
2. P. C. Loizou, Speech Enhancement: Theory and Practice, 1st ed. Boca Raton, FL.: CRC, 2007.
3. 3. Ephraim, Y. and Malah, D. (1984), Speech enhancement using a minimum mean square error short time spectral amplitude estimator, IEEE Trans. Acoustic. Speech Signal process.,32(6), pp.1109-1121
4. Boll, S.F. (1979), Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoustic. Speech Signal Process,27(2), pp.113-120.
5. Berouti, Schwartz, M and Makhoul. (1979), Enhancement of Speech corrupted by acoustic noise, Proc. IEEE Int. Conf. Acoustic. Speech Signal Process, pp.208-211.
6. S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction. John Wiley and Sons, 2006.
7. Y. Shao and C.-H. Chang," A generalized time frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," IEEE Trans. Systems, Man, Cybernetics, Part B, vol. 37, no. 4, pp. 877-889, Aug.
   2007.
8. B. Yegnanarayana, C. Avendano, H. Hermansky, and P. Satyanarayana Murthy," Speech enhancement using linear prediction residual," Speech Communication, vol. 28, pp. 2542, May 1999.