ADAPTIVE NOISE ESTIMATION AND REDUCTION BASED ON TWO-STAGE WIENER FILTERING IN MCLT DOMAIN

Mahwash Ahmed, Zahid Hasan Bawar

National University of Sciences and Technology (NUST), Islamabad, Pakistan
mahwash_a@yahoo.com, Zahid_h_Bawar@yahoo.com

ABSTRACT

We propose an adaptive noise estimation and reduction algorithm which is capable of reducing additive noise from the noisy speech signals with low SNR values. The algorithm uses Modulated Complex Lapped Transform (MCLT) to estimate the power spectrum of input signal. The noise is estimated continuously from the spectrum using time-frequency dependent smoothing factor and tracking spectral minima. The gain function is then estimated using the smoothed *a priori* SNR value for the current frame instead of the previous frame using two-stage wiener filters. This method is simple to implement and greatly suppresses the residual musical noise as well as delay, providing consistent speech quality improvement across all SNRs and on average, nearly 0.13 Perceptual Evaluation of Speech Quality (PESQ) improvements.

Index Terms— noise estimation, noise reduction, wiener filter, spectral minimum, modulated complex lapped transform (MCLT)

1. INTRODUCTION

Noise reduction is a useful pre-processing step in many applications such as speech communication, automatic speech recognition, speaker recognition and speech coding systems. The unwanted random addition of noise not only degrades the speech quality and intelligibility, but also hinders system's performance. Hence noise reduction methods are used to reduce/suppress noise; without distorting the signal.

A number of noise reduction and speech enhancement algorithms, when only noisy speech is available, have been widely studied in the past. Classical approaches proposed by Boll [1], Berouti et al. [2], Lim and Oppenheim [3] are an intuitive and effective speech enhancement methods for the removal of additive noise. These methods have been experimentally optimized on the basis of the Signal-to-Noise Ratios (SNR) estimation of the input noisy signal ([4]-[8]). However, in these enhancement methods *a priori* SNR depends on speech spectrum estimation in the previous frame and as a result the gain function matches the previous frame rather than the current one. Moreover, these approaches require Voice Activity Detector (VAD) to estimate and update noise spectrum whose performance degrades considerably in low SNR conditions and highly non-stationary noise case. These limitations degrade the noise estimation and reduction performance, causing residual noise and speech distortion. The objective of this paper is to overcome these limitations to reduce/ suppress the noise, without introducing any perceptible distortion in the signal.

In this paper, we propose an adaptive noise spectrum estimation and reduction algorithm based on a two-stage wiener filtering capable of operating at a very low SNR (<10dB) in real time nonstationary environment. The spectral analysis - synthesis is performed using Modulated Complex Lapped Transform (MCLT). The noise spectrum is estimated continuously by using time-frequency dependent smoothing factor and then tracking the spectral minimum of noisy speech power without any distinction between speech activity and pause. The two stages of wiener filtering produce a smoothed estimate of *a priori* SNR close to the gain function for current frame rather than previous one; thus it is able to suppress residual musical noise as well as delay.

This paper is organized as follows: Section 2 discusses the proposed adaptive noise estimation method and two-stage wiener filtering to filter the estimated noise spectrum from noisy speech spectrum in MCLT domain. The experimental evaluation and results are presented in Section 3; moving to the conclusive remarks in Section 4.

2. PROPOSED ALGORITHM

In this paper we are basically concerned with the additive noise which is a major source to degrade signal SNR. The spectral components of both clean speech and noise signal are assumed to be zero mean statistically independent Gaussian random variables; this underlies the design of many speech enhancement systems.

$$y(n) = x(n) + d(n) \quad (1)$$

where y (n) is noise-corrupted input signal composed of the clean original speech signal x(n) and the uncorrelated (additive) noise signal d (n).

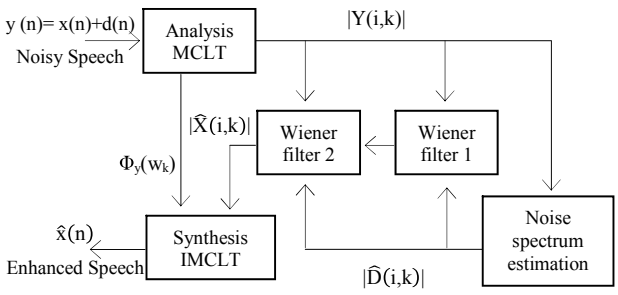

Figure 1. Basic layout of proposed solution

A simplified block diagram of our proposed solution is shown in Fig. 1. The input noisy speech spectrum is analyzed using MCLT, followed by background noise spectrum estimation which is filtered from the noisy speech signal using wiener filters. Finally the filtered signal is then used to synthesize the enhanced signal using the inverse MCLT.

2.1. Modulated Complex Lapped Transform (MCLT)

MCLT is used to calculate the spectra of input time domain speech signal. MCLT is similar to a windowed Fourier transform, but with slightly different center frequencies. It is structured as a cosine-/sine-modulated filter bank that maps overlapping blocks of a real-valued signal into complex-valued blocks of transform coefficients. Thus for a given block length the MCLT has twice the frequency resolution of the discrete Fourier transform. This greatly reduced the time-domain aliasing and the warbling artifacts.

The input noisy spectrum Y(i,k) of frame i at frequency bin k computed via MCLT is given below:

$$Y(i,k)=\sum_{n=0}^{2M-1} y(iM+n)p(n,k) \quad \text{where } k=0,1,\ldots,M-1 \quad (2)$$

where frame length N = 2M is typically chosen in the range of 20 – 40ms with 50 % overlap for speech signals analysis. $p(n,k)$ is the MCLT analysis basis function defined by H. S. Malvar [9] as

$$p(n,k)= p_c(n,k)- j p_s(n,k) \quad \text{where } j=\sqrt{-1}, \quad (3)$$

$$p_c(n,k)= h(n)\sqrt{2/M}\cos\left[\left(n+\tfrac{M+1}{2}\right)\left(k+\tfrac{1}{2}\right)\tfrac{\pi}{M}\right] \quad \text{and}$$

$$p_s(n,k)= h(n)\sqrt{2/M}\sin\left[\left(n+\tfrac{M+1}{2}\right)\left(k+\tfrac{1}{2}\right)\tfrac{\pi}{M}\right]$$

where h(n) is the window function commonly defined as $h(n)= -\sin\left[\left(n+\tfrac{1}{2}\right)\pi/2M\right]$. We can also write MCLT coefficients for a frame as $Y(k) = Y_c(k) – jY_s(k)$, with

$$Y_c(k)=\sum_{n=0}^{2M-1} y(n)p_c(n,k) \quad Y_s(k)=\sum_{n=0}^{2M-1} y(n)p_s(n,k) \quad (4)$$

The inverse MCLT of processed/ enhanced coefficients $\hat{X}(k)$ is computed via the reconstruction formula:

$$\hat{x}(n)=\beta_c\sum_{k=0}^{M-1}\hat{X}_c(k)p_c(n,k)+\beta_s\sum_{k=0}^{M-1}\hat{X}_s(k)p_s(n,k) \quad (5)$$

where the coefficients β_c and β_s can be chosen such that $\beta_c + \beta_s = 1$. For our proposed algorithm, we choose cosine only reconstruction for inverse MCLT ($\beta_c = 1$ and $\beta_s = 0$). This is followed by the overlap-add synthesis to reconstruct the final enhanced output signal $\hat{x}(n)$.

2.2. Noise Estimation

We use ‘minimum statistics’ approach for continuous estimation of the noise spectrum using time-frequency dependent smoothing factor and then tracking the spectral minimum of noisy speech power in each frequency band. This method does not need VAD and is not constrained by a specified time window for updating the estimate of noise spectrum. Fig.2 shows the flow diagram for noise estimation.

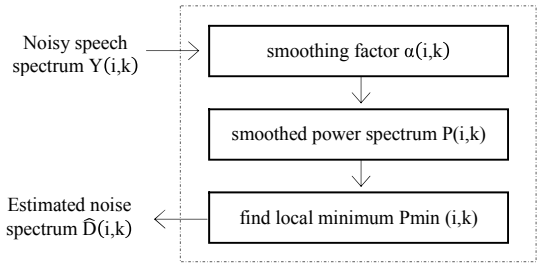

Figure 2. Flow diagram of proposed noise-estimation algorithm

A short-time smoothed version of the periodogram of noisy speech is computed as:

$$P(i,k) = \alpha(i,k)P(i-1,k) + (1- \alpha(i,k))|Y(i,k)|^2 \quad (6)$$

where P(i,k) is the smoothed power spectrum, $|Y(i,k)|^2$ is the short time power spectrum of noisy speech and α(i,k) is a time and frequency dependent smoothing constant.

To derive an optimal smoothing constant, we require smoothed power spectrum P (i,k) to be close to the true noise spectrum $|D(i,k)|^2$ during speech pauses. Thus, our objective is to minimize the conditional mean square error

$$E\{ (P(i,k) - |D(i,k)|^2)^2 \mid P(i-1,k) \}$$

After substituting P(i,k) from (6) in above equation and using $E\{|Y(i,k)|^2\}=|D(i,k)|^2$ and $E\{|Y(i,k)|^4\}=2|D(i,k)|^4$ during speech pauses, the mean square error is given by

$$E\{ (P(i,k) - |D(i,k)|^2)^2 \mid P(i-1,k) \}$$
$$=\alpha^2(i,k)\left(P(i\text{-}1,k)\text{-}|D(i,k)|^2\right)^2+|D(i,k)|^4(1-\alpha(i,k))^2 \quad (7)$$

Setting the first derivative with respect to (i,k) to zero yields

$$\alpha_{opt}(i,k) = 1/(1 + (P(i-1,k)/|D(i,k)|^2 - 1)^2) \quad (8)$$
$$= 1/(1 + (\bar{\gamma} - 1)^2)$$

where $\bar{\gamma} \triangleq P(i\text{-}1,k)/|D(i,k)|^2$ is smoothed *a posteriori* SNR and $0< \alpha_{opt}(i,k)<1$. Ideally, for better noise tracking we would like α_{opt} smoothing constant to be close to zero when

speech is present (i.e., for large $\overline{\gamma}$). In practice, however, we replace $|D(i,k)|^2$ with the latest estimate $\left|\hat{D}(i\text{-}1,k)\right|^2$, that in general lags the true noise spectrum.

To detect the deviations of the short term psd estimate from the actual averaged periodogram, we compare the average short-term psd estimate of the previous frame $1/M \sum_{k=0}^{M-1} P(i-1,k)$ of the average periodogram $1/M \sum_{k=0}^{M-1} |Y(i,k)|^2$ as proposed by Martin [11] and adjust the smoothing parameter accordingly.

$$\tilde{\alpha}_c(i) = 1/\left(1+\left(\sum_{k=0}^{M-1} P(i\text{-}1,k)/\sum_{k=0}^{M-1}|Y(i,k)|^2 -1\right)^2\right) \quad (9)$$

The resulting correction factor is limited to values larger than 0.7 and smoothed over time

$$\alpha_c(i) = 0.7\alpha_c(i-1) + 0.3\max(\tilde{\alpha}_c(i), 0.7) \quad (10)$$

The final smoothing parameter after including the preceding correction factor $\boldsymbol{\alpha}_c(i)$ gives:

$$\alpha(i,k) = \alpha_{max}\alpha_c(i)/\left(1+\left(P(i\text{-}1,k)/\left|\hat{D}(i\text{-}1,k)\right|^2 -1\right)^2\right) \quad (11)$$

where $\boldsymbol{\alpha}_{max} = 0.96$ in [11] to avoid deadlock when $\overline{\gamma} = 1$.

The local minimum of the noisy speech power spectrum is computed by averaging the past spectral values with a look-ahead factor (beta) as defined in [10]:

$$\text{if } P_{min}(i-1,k) < P(i,k) \quad (12)$$

$$P_{min}(i,k) = \gamma P_{min}(i-1,k) + \frac{1-\gamma}{1-\beta}P(i,k) - \beta P(i-1,k)$$

$$\text{else } P_{min}(i,k) = P(i,k)$$

where $P_{min}(i,k)$ denotes the local minimum of the noisy speech power spectrum and $\boldsymbol{\gamma}$=0.998 and $\boldsymbol{\beta}$=0.96 are constants.

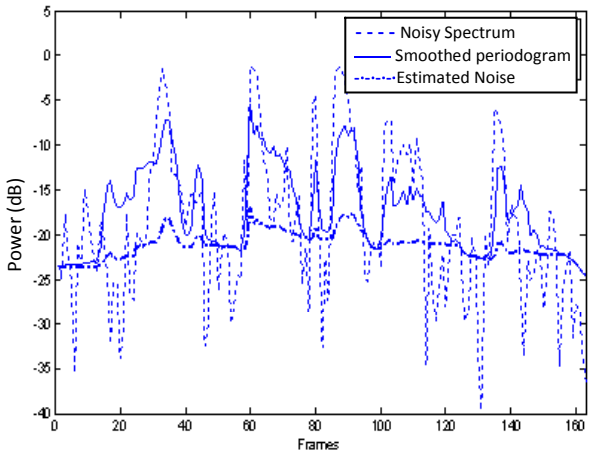

Figure 3. Dotted line (.) noisy speech power spectrum; solid line (-) smoothed periodogram; dash dotted line (-.) estimated noise spectrum using proposed method for a speech signal degraded by babble noise at 5dB global SNR at k=13 bin.

Figure.3 shows the power spectrum of noisy speech and the local minimum tracked using the above proposed method for a sentence degraded by babble noise at 5 dB global SNR. We consider the estimates at frequency bin as k=13. It can be easily seen that the local minimum estimation algorithm adapts very quickly to highly non-stationary noise environments.

2.3. Two Stage Wiener Filtering

The basic principle of the Wiener filter is to obtain an estimate of clean signal $\hat{x}(n)$ from corrupted signal. This estimate is obtained by minimizing the mean square error (MSE) between the desired signal x(n) and the estimated signal $\hat{x}(n)$ i.e. $E[\{x(n) - \hat{x}(n)\}^2]$.

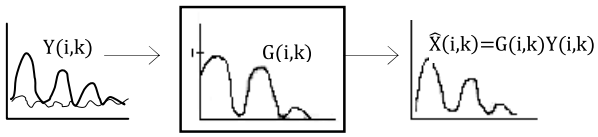

Figure 4. Filtering operation realized in frequency domain

Wiener filter is a zero-phase filter given by:

$$G(k) = \frac{|X(k)|^2}{|D(k)|^2 + |X(k)|^2} = \frac{\xi_k}{1+\xi_k} \quad (13)$$

$$\text{with } G(k) \approx \begin{cases} 1, & |X(k)|^2 \gg |D(k)|^2 \quad \text{i.e. } \xi k \to \infty \\ 0, & |X(k)|^2 \ll |D(k)|^2 \quad \text{i.e. } \xi k \to 0 \end{cases}$$

Wiener filter is an adaptive gain function G(k) that weighs spectrum according to SNR at different frequencies.

$$\xi_k = \frac{|X(k)|2}{|D(k)|^2} \qquad \gamma_k = \frac{|Y(k)|^2}{|D(k)|^2} \qquad \overline{\xi}_k = \gamma_k - 1$$

ξ_k is the a *priori* SNR and γ_k is the a *posteriori* SNR of the kth spectral component. The instantaneous SNR $\overline{\xi}_k$ can be interpreted as the un-smoothed estimate of *a priori* SNR.

The estimate of the short-time clean speech spectral magnitude is obtained as:

$$\left|\hat{X}(i,k)\right| = G(i,k)|Y(i,k)| \quad (14)$$

The focus is on getting low-variance estimate of the *a priori* SNR ξ_k as it can eliminate the musical noise [6]. On the contrary, when frame-adaptive spectral estimates are used to compute the wiener filter gain in (13), low-level speech frames can make G(k) fluctuate rapidly, generating annoying musical noise in the filtered signal.

In the first stage, the input signal is wiener filtered using SNR $\xi_{i,k}$. We use "Decision-Directed" (DD) method proposed by Ephraim and Malah [4] for the estimator $\hat{\xi}_{i,k}$ of $\xi_{i,k}$ as:

$$\hat{\xi}_{i,k} = \alpha_d \frac{\left|\hat{X}_1(i\text{-}1,k)\right|^2}{\left|\hat{D}(i\text{-}1,k)\right|^2} + (1-\alpha_d)\max\left(\gamma_{i,k}\text{-}1, 0\right) \quad (15)$$

where $\left|\hat{X}_1(i\text{-}1,k)\right|^2$ denotes the enhanced speech spectrum in the (i-1)th analysis frame. $\left|\hat{D}(i\text{-}1,k)\right|^2$ is the estimate of kth noise spectral component. $\boldsymbol{\alpha}_d$ is a weighting factor ($0 \le \alpha_d \le 1$)

that controls the trade-off between noise reduction and the transient distortion introduced into the signal. The large α_d value indicates low level of residual musical noise (typically equal to 0.98 [7]).

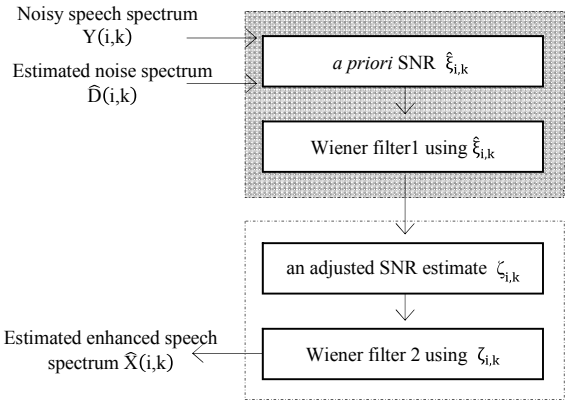

Figure 5. Flow chart of Wiener Filtering process

With the smoothed estimate $\hat{\bar{\xi}}_{i,k}$, we reduce variations in the wiener gain G(i,k) over time. This helps to suppress the residual musical noise. But the delay (frame duration) may lead to reverberation effects at the end of speech utterances. To avoid this distortion, second wiener filter is used. The flow chart of two stage wiener filtering is shown in Fig.5. This approach is able to suppress the delay while maintaining benefits of the DD algorithm.

In the second stage, output of first wiener filter is again wiener filtered using an adjusted SNR estimate:

$$\zeta_{i,k} = \alpha\hat{P}(i-1,k) + (1-\alpha)\frac{|\hat{X}_1(i,k)|^2}{|\hat{D}(i,k)|^2} \qquad (16)$$

with $\hat{X}_1(i,k)$ is the filtered signal from the first wiener filter. α is smoothing constant.

$$\hat{P}(i-1,k) = |\hat{X}(i-1,k)|^2 / |\hat{D}(i-1,k)|^2 \qquad (17)$$

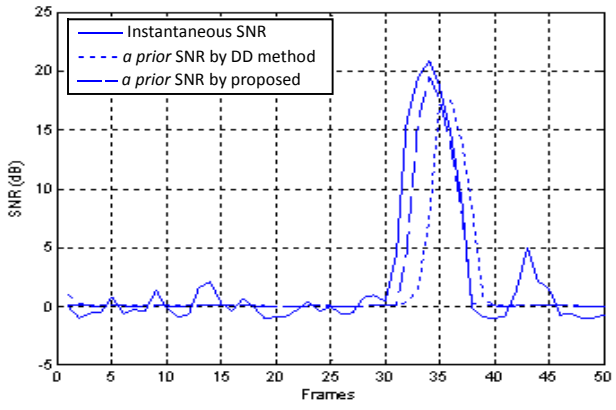

Figure 6. Solid line (-) instantaneous SNR $\bar{\xi}_k$; dotted line (.) a prior SNR $\hat{\xi}_{i,k}$ of DD algorithm (after smoothing); dashed line (--) final a prior SNR $\zeta_{i,k}$ estimated of proposed algorithm

Figure 6 shows the comparative a priori SNRs estimation calculated by DD and proposed algorithm along with time varying instantaneous SNR. We consider the estimates using sentence corrupted by additive car noise at 375 Hz and 5 dB SNR. It can be seen that the isolated small magnitude pulses (corresponding directly to musical noise) are suppressed after the smoothing operation. In DD approach, the *a priori* SNR $\hat{\bar{\xi}}_k$ depends on speech spectrum estimation in the previous frame rather than the current one. As a consequence, $\hat{\bar{\xi}}_k$ is delayed with respect to $\bar{\xi}_k$. We note that newly estimated *a prior* SNR ζ_k is shifted back and synchronized with that of instantaneous SNR while suppressing the small magnitude pulses to avoid musical noise.

3. EVALUATION AND RESULTS

In this section, we assess the performance of proposed enhancement algorithm using Noizeus speech corpus [12]. Corpus is composed of 30 phonetically-balanced IEEE sentences belonging to six speakers (three males and three females), corrupted by eight real-world noise sources at various SNRs (0, 5, 10 & 15 dB). The corpus is sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets.

For objective speech quality evaluation, SNR and PESQ (Perceptual Evaluation of Speech Quality) are calculated for noisy and enhanced signal. The outcome of the PESQ measure is an estimate of the subjective mean opinion score (MOS), which has values between 0 (poor quality) and 4.5 (no perceptual distortion). Publicly available reference implementation of these methods (Loizou [12]) is employed in our study.

The evaluated results for SNR and PESQ are given in Table I and Table II respectively. We see that the proposed algorithm performs well, providing consistent improvements across all SNRs, with the average of about 0.13 PESQ improvement.

TABLE I. GLOBAL SNR VALUE OBTAINED

Noise Type	SNR (dB)			
	0	***5***	***10***	***15***
Airport	1.88	6.76	11.22	15.21
Babble	1.78	6.7	11.35	15.13
Car	3.04	7.63	11.95	15.57
Exhibition hall	1.96	6.74	11.31	15.16
Restaurant	1.41	6.25	10.93	14.86
Station	2.41	7.24	11.73	15.34
Street	2.24	7.10	11.51	15.26
Train	2.31	7.01	11.5	15.21

TABLE II. PESQ VALUE OF NOISY AND ENHANCED SIGNAL

Noise Type	PESQ				
	SNR	***0***	***5***	***10***	***15***
Airport	*Noisy*	1.73	2.02	2.34	2.63
	Enhanced	1.80	2.13	2.46	2.76
Babble	*Noisy*	1.71	2.01	2.32	2.65
	Enhanced	1.80	2.11	2.43	2.78
Car	*Noisy*	1.63	1.89	2.20	2.53
	Enhanced	1.77	2.05	2.39	2.72
Exhibition hall	*Noisy*	1.59	1.88	2.18	2.51
	Enhanced	1.68	2.02	2.34	2.67
Restaurant	*Noisy*	1.75	2.00	2.37	2.66
	Enhanced	1.82	2.09	2.46	2.76
Station	*Noisy*	1.67	1.96	2.25	2.58
	Enhanced	1.78	2.12	2.43	2.75
Street	*Noisy*	1.56	1.90	2.25	2.54
	Enhanced	1.76	2.07	2.41	2.71
Train	*Noisy*	1.60	1.86	2.16	2.49
	Enhanced	1.73	2.00	2.30	2.65

The spectrograms (having the dynamic range set to 50 dB) in Fig. 7 show a single clean speech signal corrupted by babble and street noise at 5dB SNR, as well as their corresponding enhanced versions. There is extensive improvement in the SNR with significant reduction of the musical noise artifacts.

4. CONCLUSION

The evaluation of objective measures, as well as the spectrograms, confirms that the proposed algorithm works well across all SNRs providing consistent improvements and significant musical noise reduction. The average PESQ improvement is about 0.13, resulting in enhances speech quality.

The proposed method is capable of operating at a very low SNR (<10dB) in real time nonstationary noisy environment for a single-microphone system. It provides continuous estimation of the noise spectrum without using VAD and can immediately track nonstationary noise by using the spectral minima of smoothed noisy speech power. Moreover, the use of two- stage wiener filtering produces a smoothed estimate of a priori SNR close to the gain function for current frame rather than previous one; thus suppressing the residual musical noise as well as the delay caused by DD algorithm.

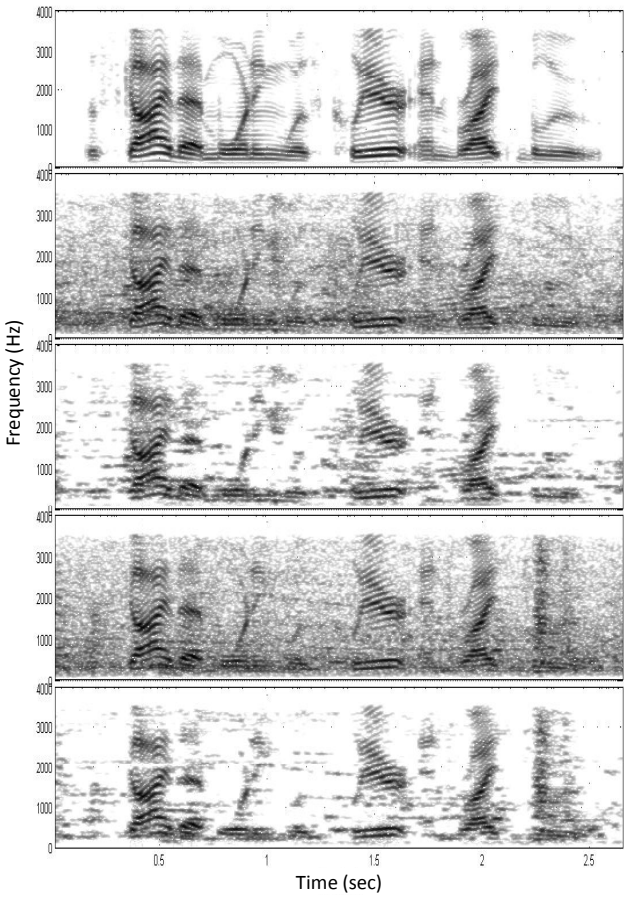

Figure 7. Spectrograms of a single speech utterance "The sky that morning was clear and bright blue." Belonging to a male speaker; (a) clean speech; (b,d) noisy speech at 5dB SNR for babble and street noise cases, and (c,e) corresponding enhanced speech

5. REFERENCE

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP,* vol.27, no.2, pp.113-120, 1979.

[2] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASSP*, pp. 208-211, 1979.

[3] J. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no.12, pp. 1586-1604, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum *mean*-square error short-time spectral amplitude estimator," *IEEE Trans. ASSP,* vol.32, no.6, pp. 1109–1121, 1984.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, 1992.

[6] O. *Cappe*, “Elimination of the musical noise phenomena with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Process.*, vol. 2, no.2 pp. 345–349, 1994.

[7] P. *Scalart* and J. Filho, “Speech enhancement based on a priori signal to noise estimation,” *Proc, IEEE ICASSP,* pp. 629–632, 1996.

[8] I. *Cohen*, “On the decision-directed estimation approach of Ephraim and Malah,” *Proc. IEEE ICASSP*, vol.1, pp. 293-296, 2004.

[9] H. S. Malvar, “A modulated complex lapped transform and its applications to audio processing,” *Proc. ICASSP.*, pp. 1421–1424, 1999.

[10] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. EUROSPEECH*, vol.2, pp. 1513-1516, 1995.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504-512, 2001.

[12] P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press LLC, 2007.