NOISE REDUCTION USING *mel*-SCALE SPECTRAL SUBTRACTION WITH PERCEPTUALLY DEFINED SUBTRACTION PARAMETERS-A NEW SCHEME

Ch.V.Rama Rao[1*], M.B.Rama Murthy[2] and K.Srinivasa Rao[3]

[1] Faculty Member, dept.of ECE, Gudlavalleru Engineering College, Gudlavalleru-521356, AP, India.
[2] Professor, Jayaprakash Narayan College of Engineering, Dharmapur, Mahabubnagar-509001, AP, India.
[3] Principal, TRR College of Engineering, Pathancheru-502319, AP, India
[1*] chvramaraogec@gmail.com

ABSTRACT

The noise signal does not affect uniformly the speech signal over the whole spectrum isn the case of colored noise. In order to deal with speech improvement in such situations a new spectral subtraction algorithm is proposed for reducing colored noise from noise corrupted speech. The spectrum is divided into frequency sub-bands based on a nonlinear multiband bark scale. For each sub-band, the noise corrupted speech power in past and present time frames is compared to statistics of the noise power to improve the determination of the presence or absence of speech. During the subtraction process, a larger proportion of noise is removed from sub-bands that do not contain speech. For sub-bands that contain speech, a function is developed which allows for the removal of less noise during relatively low amplitude speech and more noise during relatively high amplitude speech .Further the performance of the spectral subtraction is improved by formulating process without neglecting the cross correlation between the speech signal and background noise. Residual noise can be masked by exploiting the masking properties of the human auditory system. In the proposed method subtraction parameters are adaptively adjusted using noise masking threshold. A psychoacoustically motivated weighting filter was included to eliminate residual musical noise. Experimental results show that the algorithm removes more colored noise without removing the relatively low amplitude speech at the beginning and ending of words.

Keywords

Speech enhancement, Spectral subtraction, Cross correlation, Sub-bands

1. INTRODUCTION

In speech communication, the speech signal is always accompanied by some noise. In most cases the background noise of the environment where the source of speech lies, is the main component of noise that adds to the speech signal. Though the obvious effect of this noise addition is to make the listening task difficult for a direct listener, a related problem is processing degraded speech in preparation for coding by a bandwidth compression system. Hence speech enhancement not only involves processing speech signals for human listening but also for further processing prior to listening. The interfering noise generally degrades the quality and intelligibility of speech. While the term intelligibility refers to the recognisability of actual content of speech, quality refers to the

DOI : 10.5121/sipij.2011.2110

aspect of the speech that determines the ease with which one can understand the speech. This degradation of speech by noise creates problems not only for just interpersonal communication but more serious problems in applications in which decision or control is made on the basis of speech signal. Hands free voice control is one such application. Thus the main objective of speech enhancement is ultimately to improve one or more perceptual aspects of speech such as overall quality and intelligibility.

Speech enhancement techniques can be classified as single channel and dual channel or multi channel enhancement techniques. In single channel enhancement techniques only noise affected signal is available where as in dual channel enhancement techniques, an additional reference signal for the noise is available. For single channel systems, several schemes [1-3] are available in literature for enhancing noise corrupted speech signal, the family of spectral subtraction algorithms can effectively reduce the background noise. The base of most spectral subtraction algorithms find their origin from works of S.Boll [4] and Berouti etal [5]. Since in a real world environment, noise spectrum is colored, its spectrum is not uniform for all the frequencies. A nonlinear frequency spacing approach for sub-band is proposed in this paper based on the fact that human ear sensibility varies nonlinearly with frequency. Spectral subtraction has been improved by dividing the spectrum into frequency sub-bands [6], [7]. To account for the hearing characteristics of the human ear, the Bark scale [8] has been used as a guide to determine possible ranges of frequencies in each sub-band [6]. Speech signal quality can be improved in perceptual sense using a nonlinear Bark scaled frequency spacing. This concept is based on the fact that human ear sensibility varies nonlinearly in frequency spectrum. In this paper, a modified spectral subtraction algorithm including cross correlation between speech and noise signal, to reduce colored noise from noise corrupted speech, is proposed.

The principle of speech enhancement technique using spectral subtraction method is given in the next section.

2. SPECTRAL SUBTRACTION METHODS

Spectral subtraction assumes that noise corrupted speech is composed of speech plus noise and noise is uncorrelated with the speech signal.

$$y(k) = s(k) + n(k) \qquad (1)$$

Taking the Fourier transform of $x(n)$ gives:

$$Y(e^{jw}) = S(e^{jw}) + N(e^{jw}) \qquad (2)$$

Short time power spectrum of the noisy speech is given by:

$$\begin{aligned} \left|Y(e^{jw})\right|^2 &= \left|S(e^{jw})\right|^2 + \left|N(e^{jw})\right|^2 \\ &+ S(e^{jw})N^*(e^{jw}) + S^*(e^{jw})N(e^{jw}) \end{aligned} \qquad (3)$$

Where $N^*(e^{jw}), S^*(e^{jw})$ represent complex conjugates of $N(e^{jw}), S(e^{jw})$. The terms $\left|N(e^{jw})\right|^2, S\left(e^{jw}\right)N^*(e^{jw})$ and $S^*(e^{jw})N(e^{jw})$ cannot be obtained directly and hence are approximated by average values given by $E\left\{\left|N(e^{jw})\right|^2\right\}, E\left\{S(e^{jw})N^*(e^{jw})\right\}$ and $E\left\{S^*(e^{jw})N(e^{jw})\right\}$, respectively. If noise and speech signal are uncorrelated then the terms $E\left\{S(e^{jw})N^*(e^{jw})\right\}$ and $E\left\{S^*(e^{jw})N(e^{jw})\right\}$ can be neglected.

A generalized form of the basic spectral is given by [4-5]. The estimate of the enhanced speech $\left|\hat{S}(e^{jw})\right|$ is given by the equation:

$$\left|\hat{S}(e^{jw})\right|^{\gamma} = \left|Y(e^{jw})\right|^{\gamma} - \alpha\left|\hat{N}(e^{jw})\right| \qquad (4)$$

$\left|\hat{N}(e^{jw})\right|$ is an estimate obtained by averaging of $\left|Y(e^{jw})\right|^{\gamma}$ during non-speech activity. For $\alpha = 1$ and $\gamma = 2$ we obtain the power spectral subtraction estimator by replacing noise square-magnitude $\left|\hat{N}(e^{jw})\right|^{2}$, with its average value taken during non-speech activity period. An important variation of spectral subtraction was proposed by Berouti etal [5] to minimize the presence of residual noise and musical noise in the processed speech. The proposed algorithm is

$$\left|\hat{S}(e^{jw})\right|^{2} = \left|Y(e^{jw})\right|^{2} - \alpha\left|\hat{N}(e^{jw})\right|^{2} \qquad (5)$$

$$\left|\hat{S}(e^{jw})\right|^{2} = \begin{cases} \left|\hat{S}(e^{jw})\right|^{2}, if \left|\hat{S}(e^{jw})\right|^{2} > \beta\left|\hat{N}(e^{jw})\right|^{2} \\ \beta\left|\hat{N}(e^{jw})\right|, otherwise \end{cases} \qquad (6)$$

where β is the spectral floor parameter and it is taken as 0.02.

3. PROPOSED SPECTRAL SUBTRACTION ALGORITHM

The assumption that speech signal and noise are completely uncorrelated limits the performance of the spectral subtraction algorithm and is not a valid assumption in real world environments. Therefore, we cannot neglect the terms $S(e^{jw})N(e^{jw})$ and $S^{*}(e^{jw})N(e^{jw})$. By accounting for the cross-terms it would be possible to reduce the residual noise in the enhanced speech and provide a better estimate of the clean speech. Unfortunately, we do not have access to the clean speech $s(k)$. Therefore, in an attempt to approximate the cross-terms, we use the corrupted signal $y(k)$ and estimates $Y(e^{jw})N^{*}(e^{jw})$ and $Y^{*}(e^{jw})N(e^{jw})$ instead of $S(e^{jw})N^{*}(e^{jw})$ and $S^{*}(e^{jw})N(e^{jw})$.

Further in real environments, noise spectrum is not uniform for all the frequencies. For example, in the case of engine noise the most of noise energy is concentrated in low frequency. To take into account the fact that colored noise affects the speech spectrum differently at various frequencies, a multi-band linear frequency spacing approach to spectral subtraction was proposed by S.Kamath etal., [4]. However they have not considered any cross correlation terms.

A nonlinear frequency spacing approach for mulit-band subtraction factor estimation is proposed including the cross correlation terms. For dividing the audible frequency range of 8 KHz into 31 abutting bands *mel*-scale is used. Table.1 shows frequency ranges for sub-bands as suggested in [6]. An approximate analytical expression to achieve conversion from linear frequency f, into the *mel*-scale is:

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) \qquad (7)$$

The spectral subtraction algorithm is modified to obtain the power spectrum on a *mel*-scale m, for a critical band analysis, using the relation

$$Y_i(m) = \sum_{w(i)} \left|Y(e^{jw})\right|^2, i = 1,.., K \qquad (8)$$

Where i is the critical band number, K=31 is the total number of critical bands and w(i) is the frequency index depending on the lower and upper frequency boundary of the critical band i.

The speech spectrum is divided into N nonoverlapping bands, and spectral subtraction is performed independently in each band. The estimate of the clean speech spectrum in the i^th band is obtained by:

$$\left|\hat{S}(m)\right|^2 = \begin{cases} \left|Y_i(m)\right|^2 - \alpha_i\left|\hat{N}(m)\right|^2 - \delta\left|Y_i(m)\right| \cdot \left|\hat{N}(m)\right|, if \left|\hat{S}(m)\right|^2 > 0 \\ \beta\left|\hat{N}(m)\right|^2, otherwise \end{cases} \qquad (9)$$

for $w_i < m < w_{i+1}$ where w_i and w_{i+1} are the beginning and ending frequency bins of the i^th frequency band and α_i is the subtraction factor of the i^th band. The choice of the value of parameter α_i controls the amount of noise subtracted from the noisy signal. Whereas for over subtraction $\alpha_i > 1$. Over subtraction allows the time-frequency spectrum to be attenuated more than necessary. This factor should be appropriately selected to provide the best trade-off between residual noise peaks and audible distortion. The noise flooring factor β $(0 \le \beta < 1)$ makes use of the addition of background noise to mask the residual noise. It determines the minimum value of the gain function. If this factor is increased, parts of the residual noise can be masked, but the level of background noise retained in the enhanced speech increases. The cross correlation coefficient, δ for estimating the correlation between the noisy speech and noise signal in a frame is calculated from [7] as

$$\delta = \left|\frac{\gamma_{yd} - \mu_y.\mu_n}{\sigma_y.\sigma_n}\right|, 0 \le \delta \le 1 \qquad (10)$$

where

$$\gamma_{yn} = \frac{1}{N}\sum_{k=0}^{N-1} Y(k) \cdot \hat{N}(k); \qquad \mu_y = \frac{1}{N}\sum_{k=0}^{N-1} Y(k)$$

$$\mu_n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{N}(k); \; \sigma_y^2 = \frac{1}{N}\sum_{k=0}^{N-1} \left(Y(k) - \mu_y\right)^2; \; \sigma_n^2 = \frac{1}{N}\sum_{k=0}^{N-1} \left(\hat{N}(k) - \mu_d\right)^2$$

Table 1. Frequency ranges of the sub-bands

Sub band	Number of bins	Frequency(Hz)
0	1	0-31
1	1	31-62
….	….	….
12	1	375-406
13	2	406-469
14	2	469-531
15	2	531-594
16	2	594-656
17	2	656-719
18	2	719-781
19	2	781-844
20	3	844-938
21	3	938-1031
22	4	1031-1156
23	6	1156-1344
24	6	1344-1531
25	8	1531-1781
26	9	1781-2063
27	10	2063-2375
28	12	2375-2750
29	14	2750-3188
30	18	3188-3750
31	9	3750-4000

and N is the FFT frame length. The value is proportional to the degree of correlation between clean speech and noise.

Equation (9) shows that the value of instantaneous power spectrum of estimated speech signal converges to the noise free signal. However, it is observed that for nonstationary signals, such as speech, the objective is to recover the instantaneous or short-time signal, and only a relatively small amount of averaging can be applied. A gain function, $G(m)$ for use with the equation (9) can be defined such that its magnitude lies in the range between 0 and 1, i.e.,

$$\hat{S}(m) = G(m).Y(m);0 \le G(m) \le 1 \qquad (11)$$

and

$$G(m) = \sqrt{1 - \frac{\alpha}{SNR_{post}} - \delta \frac{\left|\hat{N}(m)\right|}{\left|Y_i(m)\right|}} \qquad (12)$$

Where $$SNR_{post} = \frac{\sum_{m=w_i}^{w_{i+1}} \left|Y_i(m)\right|^2}{\sum_{m=w_i}^{w_{i+1}} \left|\hat{N}_i(m)\right|^2}$$

is the posterior SNR, which is defined as the ratio of the power spectrum of the noisy speech signal to that of the estimated noise. In the low SNR case, the gain is set to zero when the power spectrum of noisy speech is less than that of the estimated noise.

The gain function in equation (12) acts as a posterior SNR dependent attenuator in the time-frequency domain. The input noisy speech is attenuated more heavily with decreasing posterior SNR and vice versa with increasing SNR. This filtering function can be readily implemented since posterior SNR can be easily measured on the noisy speech. Although it can effectively filter out the background noise from the noisy speech, the processing itself introduces musical residual noise. This processing distortion is due to the random variation of noise spectrum and the nonlinear mapping of the negative or small-valued spectral estimate. To make this residual noise "perceptually white," Virag [11] introduced several flexible subtraction parameters. These parameters are chosen to adapt to a criterion associated with the human auditory perception. The generalized spectral subtraction method in [11], considers only the simultaneous masking property of human auditory system, and the adaptation parameters are empirically determined. In this work, the generalized time-frequency subtraction method is developed in perceptual domain and the close form expressions for its optimal adaptation parameters.

4. PERCEPTUAL WEIGHTING FUNCTION

Let $\hat{S}(m) = G(m).Y(m)$ be the spectrum of the estimated enhanced speech and $G(m)$ the perceptual weighting function. The error spectrum is defined as

$$E(m) = \hat{S}(m) - S(m) = G(m).Y(m) - S(m) \quad = \left[G(m) - 1\right]\!.S(m) + G(m).\hat{N}(m) \qquad (13)$$

The first term describes the speech distortion caused by the spectral gain and the second term is the residual noise. Musical residual noise results from the pure tones present in the residual noise. Let

$$E_S^2 = E\left|\left\|E_S(m)\right\|^2\right| = E\left|\left|S(m)\right|^2\right|.\left(G(m) - 1\right)^2 \qquad (14)$$

Be the spectral energy of the speech distortion. Similarly,

$$E_N^2 = E\left|\left\|E_N(m)\right\|^2\right| = E\left|\left|\hat{N}(m)\right|^2\right|.G^2(m) \qquad (15)$$

denote the spectral energy of the residual noise. Where $\hat{N}(m)$ represent the spectra of noise signals. Assuming the noise signal is additive and uncorrelated with a speech signal. The perceptual weighting function $G(m)$ can be optimized by minimizing the short-term spectral energy associated with the speech distortion subject to a constant on the short-term spectral energy related to residual noise below the noise masking threshold (NMT).

$$\min_{g(m,\omega)} \left\{E_S^2(m)\right\} \qquad (16)$$

subject to the constant $E_N^2(m) \le T(m)$

where $T(m)$ is the noise masking threshold (NMT) corresponding to the frequency bin ω. The perceptual gain factor is designed such that it minimizes the speech distortion over all linear filters which result in the permissible residual noise level $T(m)$. The perceptual gain factor is estimated in lieu of (16) and the Kuhn-Tucker conditions for constrained minimization [12]. Then cost function J is formulated according to speech distortion and residual noise spectral energy as

$$J = E_S^2(m) + \mu\left[E_N^2(m) - T(m)\right] \quad (17)$$

where μ is the Lagrangian multiplier. It will be zero if the level of residual noise is under a given threshold. Substituting (14) and (15) in (17), the cost function becomes

$$J = E\left[|S(m)|^2\right](G(m)-1)^2 + \mu\left[E\left[\left|\hat{N}(m)\right|^2\right]G^2(m) - T(m)\right] \quad (18)$$

In order to minimize the cost function, first equation (18) is partially differentiated with respect to the Lagrangian multiplier and set the result to zero. The perceptual spectral weighting function is obtained as

$$G(m) = \sqrt{\frac{T(m)}{E\left[\left|\hat{N}(m)\right|^2\right]}};0 \le G(m) \le 1 \quad (19)$$

by equating the gain function of (12) to the perceptual weighting function of (19), the closed-form expressions for the subtraction parameters, α and β can be derived as follows:

$$1 - \frac{\alpha}{SNR_{post}} - \delta\frac{\left|\hat{N}(m)\right|}{|Y(m)|} = \frac{T(m)}{E\left[|N(m)|^2\right]} \quad (20)$$

$$\Rightarrow \alpha = \left[1 - \frac{T(m)}{E\left[\left|\hat{N}(m)\right|^2\right]} - \delta\frac{\left|\hat{N}(m)\right|}{|Y(m)|}\right] \cdot SNR_{post} \quad (21)$$

$$\frac{\beta}{SNR_{post}} = \frac{T(m)}{E\left[\left|\hat{N}(m)\right|^2\right]} \quad (22)$$

$$\Rightarrow \beta = SNR_{post} \cdot \frac{T(m)}{E\left[\left|\hat{N}(m)\right|^2\right]} \quad (23)$$

equations (21) and (23) ensures that the subtraction parameters α and β are adapted to the masking threshold of human auditory system to achieve a good trade-off between the residual noise, speech distortion and background noise. In high SNR condition, the parameter α is increased to reduce the residual noise at the expense of introducing more speech distortion. On the contrary, in low SNR condition, the parameter β is increased to trade the residual noise reduction for an increased background noise in the enhanced speech. To make the residual noise inaudible, the subtraction parameters are set such that the residual noise stays just below the masking threshold $T(m)$. If the masking threshold is low, the subtraction parameters will be increased to reduce the effect of residual noise. If the masking threshold is high, the residual noise will naturally be masked and become inaudible. Therefore, the subtraction parameters can be kept at their minimal values to minimize the speech distortion.

5. ESTIMATION OF NOISE MASKING THRESHOLD

Noise masking threshold (NMT) can be estimated on the spectra of enhanced speech $\hat{S}(m,\omega)$ as indicated in equation (9). The subband energy $\varepsilon(k)$ is computed by

$$\varepsilon(k) = \sum_{\omega=\omega_{k,l}}^{\omega_{k,h}} \left|\hat{S}(m)\right|^2 \qquad (24)$$

where $\omega_{k,h}$ and $\omega_{k,l}$ representing the upper and lower frequencies at the k^{th} critical band. The upper and lower frequencies of a critical band can be found in [11][13].

In order to take into account the masking properties between different critical bands, an excitation pattern $B(k)$ can be thought of as an energy distribution along the basilar membrane. $B(k)$ is determined by convolving the subband energy $\varepsilon(k)$, with the spreading function $SF(k)$, which can be found [13].

$$B(k) = SF(k) * \varepsilon(k) \qquad (25)$$

A relative threshold offset specifies whether the speech frame is tone-like or noise-like. This threshold should be imposed to adjust the log subband energy. Therefore, a threshold $\hat{B}(k)$, is evaluated as the sum of the log energy of the excitation pattern and the offset $O(k)$.

$$\hat{B}(k) = 10 \log_{10} B(k) + O(k) \qquad (26)$$

Convolving the subband energy with the spreading function increases the energy in each subband. Thus, the $\hat{B}(k)$ should be normalized to obtained the simultaneous masking threshold $Th(k)$, it is given by

$$Th(k) = \hat{B}(k) - G(k) \qquad (27)$$

where $G(k)$ denotes the gain factor between the spread energy $B(k)$, and the subband energy $\varepsilon(k)$, at the k^{th} subband. $G(k)$ is expressed as

$$G(k) = 10 \cdot \log_{10}\left(\frac{B(k)}{\varepsilon(k)}\right). \qquad (28)$$

The simultaneous masking threshold is compared with the absolute-hearing threshold which is frequency dependent and can be closely approximated by the expression

$$Aht(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 0.001 f^4 [\mathrm{dB}] \qquad (29)$$

With f in Kilohertz. Finally, the NMT $T(k)$ is obtained by

$$T(k) = \max\{Aht(f), Th(k)\} \qquad (30)$$

where F is chosen to be the central frequency of the k^{th} critical band.

To check the performance of the proposed scheme three objective measures namely Degree of noise reduction, Time-domain average segmental SNR measure and Modified Bark Spectral Distortion measure (MBSD) are used. The performance of proposed scheme is compared with that of power spectral subtraction [2 4] and critical band spectral subtraction [13 14].

5.1 Degree of noise reduction

There will be a trade-off between the noise suppression and speech quality. The performance of the proposed system is evaluated by the noise reduction, $NR(i)$, defined as

$$NR(i) = \frac{P_{in}(i)}{P_{out}(i)} \tag{31}$$

Where $P_{in}(i)$ the background noise level is in the corrupted speech signal and $P_{out}(i)$ is the noise level in the enhanced signal.

5.2 Time-domain Average segmental SNR measures

The time-domain Average segmental SNR (Avg.SNR$_{\text{seg}}$) measure was computed as per [15-16] given by

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} s^2(k)}{\sum_{n=Nm}^{Nm+N-1}\left(s(k)-\hat{s}(k)\right)^2} \tag{32}$$

where $s(k)$ is the input (clean) signal, $\hat{s}(k)$ is the processed (enhanced) signal, N is the frame length (chosen to be 20 ms), and M is the number of frames in the signal.

5.3 Modified Bark Spectral Distortion measure (MBSD)

This is a relatively new measure that comes under the family of perceptual distortion measures [17].This method is proposed as a modification over the conventional BSD method. The processing can be divided into three major steps. 1) Loudness calculation, 2) Noise masking threshold calculation, and 3) MBSD distortion computation. The speech signal is converted to loudness domain. The MBSD uses the noise masking thresholds for calculation of determination of the audible distortion. The noise-masking threshold is calculated by critical band analysis, spreading function application and absolute threshold computation. The loudness of the noise-masking threshold is compared to the loudness difference of the original and enhanced speech signal to determine whether the difference is perceptible. If the difference is below the masking threshold, then it is imperceptible and is not included in the calculation of MBSD. In order to define the distortion for the MBSD, an indicator of perceptible distortion $M(i)$ is given, where i is the i^{th} critical band. $M(i)$ is 1 when distortion is perceptible and 0 otherwise. The MBSD distortion is defined as the average difference of the estimated loudness and can be given as:

$$MBSD = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{k} M(i)\left|L_s^{(j)}(i) - L_{\hat{s}}^{(j)}(i)\right|^n\right] \tag{33}$$

where N is the number of samples, K is the number of critical bands, $L_s^j(i)$ and $L_{\hat{s}}^j(i)$ are the bark spectrum of the *j-th* frame of the original speech and enhanced speech respectively.

6. RESULTS AND DISCUSSION

In the simulation, noisy speech signals are obtained by adding a clean speech signal with white, airport, babble (speech-like), car, street, train, station and restaurant noise signals which are extracted from NOIZEUS database [18]. Four SNR levels, namely 0 dB, 5 dB, 10 dB and 15 dB, are used to evaluate the performance of a speech enhancement system. The noise power spectral density is estimated based on averaging the noisy speech power spectrum using time and frequency dependent smoothing factors [19]. These factors are adjusted based on signal-presence probability in individual frequency bins.

Table 2 shows the noise reduction values in dB, from this table it is observed that for 0 dB SNR values the background noise is greatly reduced. Table 3 presents the performance comparisons in terms of the average segmental SNR values. Proposed scheme gives better results (lowest SNR values) comparatively with other two methods.

Figure 1 indicates the output MBSD values .There is a marked improvement in the quality of speech obtained by the proposed method over other two approaches as evident from the Figure 1, wherein the proposed method outperformed the other approaches for 0 dB, 5 dB and 10 dB scenarios.

Babble noise is a speech like noise and to remove this kind of noise from a corrupted noisy speech signal is difficult. Figure 2 is the waveform plots of a speech signal uttered by a male speaker corrupted by babble noise with 0 dB SNR. Comparing the waveforms of enhanced speech shown in

Table 2. Noise reduction values (dB)

Type of noise and SNR(dB)	Power SS	Critical band SS	Proposed scheme
Airport-0	18.74	13.11	25.37
Airport-5	19.98	16.35	24.68
Airport-10	20.29	21.79	23.21
Airport-15	22.78	22.91	23.04
Babble-0	18.12	13.21	24.73
Babble-5	20.19	16.50	24.70
Babble-10	21.56	19.74	23.91
Babble-15	21.56	23.29	22.99
Car-0	18.45	13.38	25.63
Car-5	21.07	16.17	25.11
Car-10	22.68	19.47	24.20
Car-15	23.70	22.96	23.07
Street-0	19.08	13.97	25.06
Street-5	19.38	16.53	25.43
Street-10	22.64	19.64	24.14
Street-15	23.20	22.76	23.54
Train-0	17.45	13.59	25.87
Train-5	20.06	16.61	24.89
Train-10	21,12	19.87	24.20
Train-15	23.72	23.68	22.86

Station-0	19.06	13.10	25.41
Station-5	20.91	16.12	25.13
Station-10	22.47	19.48	24.15
Station-15	23.86	22.91	23.10
Restaurant-0	16.95	13.92	25.44
Restaurant-5	17.93	16.69	24.81
Restaurant-10	19.98	19.95	23.63
Restaurant-15	22.53	23.32	22.92

Figures 2(c), 2(d) and 2(e), the proposed method significantly improve the performance of the speech enhancement system in removing background noise. The proposed method significantly reduces the amount of residual noise in speech-pause regions and the enhanced speech signal has not been severely deteriorated during speech-dominant regions. Therefore, the speech quality can be maintained at an acceptable level.

Table 3. Time-domain average segmental SNR values

Type of noise and SNR(dB)	Power SS	Critical band SS	Proposed scheme
Airport-0	-16.20	-18.94	-5.55
Airport-5	-13.12	-13.83	-2.61
Airport-10	-9.17	-8.95	-0.39
Airport-15	-7.74	-2.82	0.18
Babble-0	-16.66	-18.64	-5.87
Babble-5	-13.35	-13.81	-1.96
Babble-10	-9.65	-9.01	-0.55
Babble-15	-6.99	-3.92	0.27
Car-0	-15.01	-19.18	-4.55
Car-5	-11.03	-14.21	-1.48
Car-10	-8.63	-8.70	0.23
Car-15	-6.55	-4.60	0.48
Street-0	-16.26	-18.34	-4.21
Street-5	-10.79	-13.99	-1.82
Street-10	-7.91	-7.75	-0.74
Street-15	-7.33	-3.97	0.45
Train-0	-16.32	-18.43	-3.69
Train-5	-12.11	-13.31	-0.98
Train-10	-9.64	-14.76	-1.95
Train-15	-7.08	-3.66	0.46

Station-0	-16.26	-18.34	-4.21
Station-5	-10.79	-13.99	-1.82
Station-10	-7.91	-7.75	-0.74
Station-15	-7.33	-3.97	0.45
Restaurant-0	-17.02	-17.73	-5.23
Restaurant-5	-13.33	-12.96	-1.86
Restaurant-10	-9.76	-8.40	-0.57
Restaurant-15	-7.67	-3.43	0.42

Figure 3 shows the spectrograms of a speech signal uttered by a male speaker corrupted by babble noise with 0 dB SNR. Observing the spectrograms of enhanced speech during speech-pause regions, the proposed method shown in Figure 3(e) is better able to remove background/residual noise than the other two methods shown in Figures 3(c) and 3(d). Consequently, by deriving the subtraction parameters in terms of noise masking threshold, improves the performance of the speech enhancement system. An informal subjective listening test also reveals that the enhanced speech produced by the proposed method sounds like less annoying that that produced by the power spectra subtraction and critical band spectral subtraction methods.

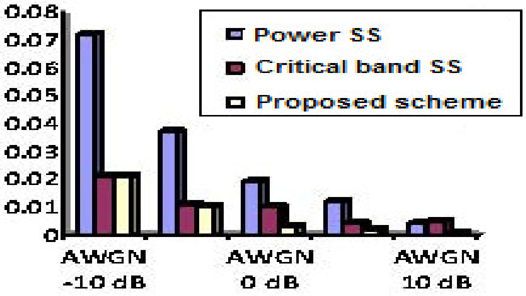

Figure 1. Output MBSD values

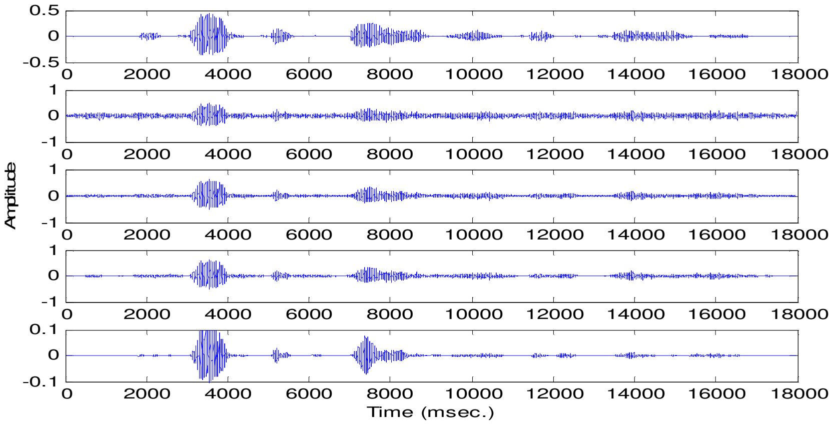

Figure 2. Time plots of (a) clean speech (b) noisy speech corrupted by babble noise with 0 dB SNR, (c) enhanced speech using power spectral subtraction, (d) enhanced speech using critical band spectral subtraction and (e) enhanced speech using proposed method

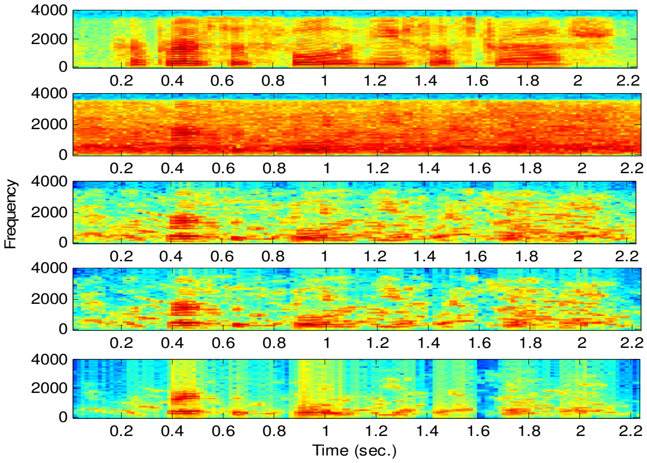

Figure 3. Spectrograms of (a) clean speech, (b) noisy speech corrupted by babble noise with 0 dB SNR, (c) enhanced speech using power spectral subtraction, (d) enhanced speech using critical band spectral subtraction method and (e) enhance speech using proposed method.

An examination of the results indicate that by using the proposed scheme there is improvement in speech quality and reduction in residual background noise as compared with that obtained from some spectral subtraction approaches.

7. CONCLUSION

In this paper, a new speech enhancement system has been proposed. It can be optimized into several powerful processing techniques that exploit the physiology of human auditory system to recover high-quality speech from noise contaminated speech. The proposed system consists of two functional stages, one is *mel*-scale spectral subtraction and other one is perceptual weighting function. The noisy speech is first decomposed into mel-scale critical bands and then performed noise reduction for each critical band using spectral subtraction. This spectral subtraction method that takes into account the non-uniform effect of colored noise on the spectrum of speech and cross correlation between back ground noise and speech signal. The subtraction parameters are adaptively adjusted using the masking threshold. Performance of this method is improved by combining with perceptual weighting function. Experimental results shows better performance it can also effectively employ for suppressing speech-like disturbances.

REFERENCES

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

[2]Y.Ephraim and D.Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", IEEE Trans.Speech Audio Processing, Vol.ASSP-32, pp. 1109-1121, 1984.

[3]Yang Gui and H.K.Kwan, "Adaptive subband Wiener filtering for speech enhancement using critical-band gammatone filterbank", Circuits and Systems, 48th Midwest Symposium on Circuits and Systems, vol.1, pp.732-735, Aug. 2005.

[4] S.Boll, "Suppression of Acoustic Noise inSpeech Using Spectral Subtraction", IEEE Transactions on Speech and Audio Processing, vol.27, no.2, pp. 113-120, 1979.

[5]M.Berouti, R.Schwartz, J.Makhoul, "Enahcement of Speech Corrupted by AcousticNoise",Proc.IEEE,Int.Conf.Acoust.,Speech and Signal Proc., pp. 208-211, Apr. 1979.

[6] J.johnston, “transform Coding of Audio Signals Using Perceptual Noise criteria”, IEEE Journal on Selected Areas of Communication, vol. 6,pp. 314-323, February 1988.

[7] S. Kamath and P. Loizou, “A Multi-band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise”, Proceedings of ICASSP-2002, Orlando, FL, May 2002.

[8] E.Zwicker and h.fastl, Psychoacoustics: Facts and Models, Springer, New-York, 1990.

[9]Ch.V.Rama Rao, M.B.Rama Murthy and K.Anith Sheela, “ A New Technique for Street Noise Reduction in Signal Processing Applications”, Proc. IEEE Int. Conf. TENCON-08, Hyderabad, Nov. 2008.

[10] G.Farahani, S.M.Ahadi, M.M.Homayounpoor and A.Kashi, “Consideration of Correlation between Noise and Clean Speech Signals in Autocorrelation-Based Robust Speech Recognition”, 9th Int. Symposium on Signal Processing and its Applications, ISSPA 2007, pp. 1-4, 12-15 Feb., 2007.

[11] N.Virag, “Speech Enhancement Based on Masking Properties of the Auditory System”, Proc. ICASSP, pp. 796-799, 1995.

[12] D.G.Luenberger. Linear and Nonlinear Programming. Reading, MA: Addison-Wesley, 1984.

[13]K.Anitha Sheela, Ch.V.Rama Rao, K.SatyPrasad and A.V.N.Tilak,“A noisereduction preprocessor for mobile voice communication using perceptually weightd spectralsubtraction method”,Proc.,3rd Int.Conf., ObCom-2006, Mobile, Ubiquitous &PervasiveComputing, ISBN-10:0230-63011-1, ISBN-13:9780230-63011-6, VIT University,Vellore, vol.1, pp. 92-100, Dec. 16-19, 2006.

[14]L. Singh and S. Sridharan, “Speech enhancement using critical band spectral subtraction,” in Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP '98), pp. 2827–2830, Sydney, Australia, December 1998.

[15] Yi Hu and Philipos C.Loizou, “Evaluation of objective quality measures for speech enhancement”, IEEE Trans. On Audio, Speech and Language Processing, vol.16, no.1, pp. 229-238, Jan.2008.

[16]Jianfen Ma, Yi Hu and Philipos C.Loizou, “Objective Measures for Predicting Speech Intelligibility in Noisy Conditions Based on New Band-important functions”, Journal of Acoustical Society America, Vol. 125, No. 5, pp. 3387-3405, May 2009.

[17]Yang, W., Benbouchta, M. and Yantorno, R., “Performance of the modified bark spectral distortion as an objective speech quality measure,” ICASSP, Vol. 1, pp. 541-544, Seattle, 1998.

[18] http://www.utdallas.edu/~loizou/speech/noizeus/

[19]Ch.V.Rama Rao, Gohthami.A, Harsha.D, Rajkumar.L, M.B.Rama Murthy, K.Srinivasa Rao and K.Anitha Sheela, “Noise estimation for speech enhancement in non-stationary environments-a new method”, accepted for oral presentation at the Int. Conf. on Signal and Image Processing, held in Singapore, during August 25-27, 2010.

AUTHORS

CH.V.Rama Rao was born on 15th June 1979. He obtained his B.E., in ECE from Nagarjuna University in 2000, M.Tech. in Digital System and Computer Electronics from JNTU, Hyderabad in 2007 and he is pursuing Ph.D at a JNTU, Hyderabad. He has a teaching experience of over 10 years and research experience of over 2 years. He published 14 papers in international and national journals and conferences. His areas of interest are Speech Processing, Signal Processing and Communications.

M B Ramamurthy is in the academic filed for the past 36 years. Currently he is Professor ECE and Dean academics in Jayaprakash Narayan College of Engineering, Mahabubnagar India 509001. He has 66 publicaions to his credit in International Journals and Conferences. His areas of interest are Speech Processing, Signal Processing and Communications.He is Senior member IEEE, Life Fellow Institute of Electronics and Telecomunication Engineers India , Life member Institution of Engineers India

K.Srinivasa Rao was born on 16th January 1962. He obtained his B.E., in ECE from Andhra University in 1985, M.Tech. in Digital Systems Engineering from Osmania University in 1993 and Ph.D., in ECE from Andhra University in 2000. He has a teaching experience of over 25 years and research experience of over 8 years. He published 31 papers in various international and national journals and conferences. He is member of IEEE, Fellow IETE and life member of ISTE, SEMCE(I) and MBMESI. His areas of interest are Digital Electronics, Microwave Communications, Communication Systems, Antennas & Propagation, VLSI, Embedded Systems and Image Processing.