Presented at National Conference on Communication, NCC 2004, IISc Bangalore , Jan 30 – Feb 1, 2004

A COMPARISON STUDY OF SPECTRAL SUBTRACTION SPEECH ENHANCEMENT METHODS

K. Manohar and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
{manohar,prao}@ee.iitb.ac.in

ABSTRACT

Spectral subtraction is a widely employed method for the suppression of additive noise in speech signals. Several variants of the basic approach have been proposed over the years to address certain shortcomings, chiefly the quality of the remnant noise and its trade-off with speech distortion. In this paper, we present a unified view of the various forms of the spectral subtraction speech enhancement method. Typical realizations chosen from each of three broad categories are described. Results of objective and subjective testing of performance are reported and discussed. We show that mild over-subtraction of noise, followed by spectral weighting, can achieve a better speech quality-intelligibility trade-off than a purely over-subtraction-based approach.

1.INTRODUCTION

Among the many available single channel enhancement methods, "spectral subtraction" has been one of the most popular techniques for its simplicity and effectiveness. The basic or power spectral subtraction method [1] estimates the clean speech power spectrum by subtracting the estimated noise power spectrum from the noisy speech power spectrum. A limitation of the basic spectral subtraction method is that it often results in excessive remnant noise. The remnant noise is described as comprising of two types of components, residual noise and musical noise. Residual noise refers to the broadband noise that has the same perceptual characteristics as the original noise. Musical noise refers to the synthetic musical tones due to the presence of random short-duration spectral peaks in the remnant noise spectrum arising from a mismatch between the noise spectrum estimate and the instantaneous noise spectrum. Several researchers have made modifications to the basic method to minimize the residual noise and musical noise artifacts.

In this paper, we present a unified view of the various modified spectral subtraction methods that have been suggested in the literature. The methods, we find, can be classified into 3 broad categories. We next evaluate the performance of three methods each of which represents one category and also comment on the effectiveness of different objective measures.

2.MODIFIED SPECTRAL SUBTRACTION METHODS

The basic spectral subtraction algorithm obtains a frame-based estimate of the speech spectrum by the subtraction of an estimate of the noise spectrum from the noisy speech spectrum.

$$\left|\hat{S}(k)\right|^2 = \left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2 \quad \text{if } \left|\hat{S}(k)\right|^2 > 0$$
$$= 0 \quad \text{otherwise} \qquad (1)$$

where $\left|\hat{S}(k)\right|, \left|Y(k)\right|$, and $\left|\hat{D}(k)\right|$ refer to speech magnitude spectrum estimate, the noisy speech magnitude spectrum and noise magnitude spectrum estimate respectively for an input speech frame and "k" is the frequency index. Nearly all modified methods proposed in the literature can be represented using the following parametric formulation of the basic method [2],

$$\left|\hat{S}(k)\right|^{\alpha} = a_k \left|Y(k)\right|^{\alpha} - b_k \left|\hat{D}(k)\right|^{\alpha} \qquad (2)$$

where a_k and b_k are the parameters .

The modified methods can be classified into three categories. While the first category employs linear spectral subtraction, the other two categories use nonlinear or frequency-dependent subtraction method. The categories are:

1. Methods which use same parameters for all frequency bins. For example Boll [3] uses

$a_k = b_k = 1$, and the over-subtraction method proposed by Berouti et al. [4] is obtained when we set $a_k = 1$ and $b_k = \nu$ as the oversubtraction factor.

2. Methods which incorporate spectrum weighting function to the basic spectral subtraction speech estimate by setting $a_k = b_k = f(k)$, as in [2]. As a variant of this approach, some apply spectrum weighting function after a mild over-subtraction [5].
3. Methods based on over-subtraction but with a frequency dependent factor, i.e. with $a_k = 1$ and $b_k = \nu(k)$, a function of frequency bin; [6] and [7] are examples of this.

Recent studies have focused on frequency dependent methods in view of the fact that noise does not affect the entire frequency range uniformly always. In the following sections, we discuss and compare Berouti's over-subtraction method [4], the constrained parametric method [2] and the multi-band spectral subtraction [6] methods, each representing one of the three categories.

2.1.Berouti over-subtraction method

The algorithm of the basic method given by (1) is modified as

$$\left|\hat{S}(k)\right|^2 = \left|Y(k)\right|^2 - \alpha\left|\hat{D}(k)\right|^2 \quad \text{if } \left|\hat{S}(k)\right|^2 > \beta\left|\hat{D}(k)\right|^2$$
$$= \beta\left|\hat{D}(k)\right|^2 \quad \text{otherwise} \qquad (3)$$

where α is the subtraction factor and β is the spectral floor parameter. Over-subtraction of the noise spectrum and the introduction of a spectral floor serve to minimize residual noise and musical noise. The α varies linearly with SNR in dB on per frame basis as

$$\alpha = \alpha_0 - s \times (\mathrm{SNR}) \qquad (4)$$

where α_0 is the value of α at SNR = 0 dB, s is slope of line and SNR is the estimated frame signal to noise ratio in dB.

2.2.Constrained parametric spectral subtraction method

J.S. Chang et al. [2] use the parametric formulation of the original method given by (2) with the constraint '$a_k = b_k$', to derive a short-time spectral amplitude estimator of the speech signal that is optimized by minimizing the mean squared error between the original (ideal) spectrum and the estimated spectrum.

The optimized speech spectrum magnitude estimate is given by

$$\left|\hat{S}(k)\right| = \left\{\frac{\xi^2(k)}{\xi^2(k)+0.5}\left(\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2\right)\right\}^{1/2} \qquad (5)$$

where the average a priori SNR '$\xi(k)$' is

$$\xi(k) \approx (1-\eta)\underbrace{\frac{\left|S(k)\right|^2_{est}}{\left|\hat{D}(k)\right|^2}}_{\approx \text{ current SNR}} + \eta\underbrace{\frac{\left|\hat{S}(k)\right|^2_{prev}}{\left|\hat{D}(k)\right|^2_{prev}}}_{\text{previous SNR}} \qquad (6)$$

where $\left|S(k)\right|^2_{est} = Max\left(\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2, 0\right)$

and 'η' is the smoothing constant. A smoothed lower bound $\overline{\mu Y}$ is used to limit the signal attenuation. The final constrained parametric estimator is implemented as follows.

$$\left|\overline{S}(k)\right| = \begin{cases} \left|\hat{S}(k)\right| & \text{, if } \left|\hat{S}(k)\right| \geq \mu\left|Y(k)\right| \\ \overline{\mu Y}, & \text{otherwise} \end{cases} \qquad (7)$$

This method represents the second category of methods wherein a spectrum weighting is applied to the basic spectral subtraction estimate. The spectrum weighting is a function of the average a priori SNR as given in (6). It serves to suppress random spectral noise peaks which have improbable high instantaneous SNR values compared to the average SNR '$\xi(k)$' by higher attenuation of these components, thus reducing musical noise.

2.3.Multi-band spectral subtraction

This recently proposed method [6] advocates varying the over-subtraction factor in Berouti's method with frequency. Speech spectrum is divided into N non-overlapping bands and spectral subtraction is performed separately in each band as

$$\left|\hat{S}_i(k)\right|^2 = \left|Y_i(k)\right|^2 - \alpha_i\delta_i\left|\hat{D}_i(k)\right|^2 \quad b_i \leq k \leq e_i \qquad (8)$$

where b_i and e_i are the beginning and the ending frequency bins of the ith frequency band, α_i is the oversubtraction factor of the ith band and δ_i is a tweaking factor that is prefixed for each frequency band. The negative values in the speech estimate are floored to β fraction of the noisy spectrum. Segmental SNR_i of the ith frequency band is calculated as:

$$SNR_i(dB) = 10\log_{10}\left(\frac{\sum_{k=b_i}^{e_i}|Y_i(k)|^2}{\sum_{k=b_i}^{e_i}|\widehat{D}_i(k)|^2}\right) \qquad (9)$$

Based on the above SNR, the α_i value is calculated as

$$\alpha_i = \begin{cases} 5 & SNR_i < -5 \\ 4 - \frac{3}{20}(SNR_i) & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \qquad (10)$$

3. IMPLEMENTATION

For all the algorithms, speech sampled at 8 kHz is Hamming windowed using a 32-ms window with a 16-ms overlap between the frames and 256 point FFT is applied. The noise estimate is updated during the silence frames by using an averaging rule. In the present exercise, an ideal VAD is assumed (i.e. frames are identified as speech or silence manually). The algorithm specific details are given next.

3.1.Over-subtraction method

This method is implemented as in eq. (3). From prior informal listening, the parameter values were selected as follows: α_0 =5, 6 or 7 with β = 0.001.

3.2.Constrained parametric method

As given in eq. (5), (6) and (7), the values of 'η' and 'μ' have a control on the amount of residual noise and musical noise. If '$\xi(k)$' is oversmoothed using 'η' > 0.995, it gives almost complete removal of musical noise but results in smearing of the speech signal. Three different sets of typical values for 'η' and 'μ' have been tried which are (0.99, 0.05), (0.99, 0.01) and (0.98, 0.05).

3.3.Multi-band spectral subtraction method

The parameter values that have been suggested by the authors [6] are used in implementation. '4' bands have been used with β=0.002 and tweaking factors (δ_i) are fixed as '1.0 for first band (i.e. 0-1 kHz) ', '2.5 for second and third bands (i.e. 1-3 kHz)' and '1.5 for the fourth band (i.e. 3-4 kHz)'.

4. PERFORMANCE EVALUATION

Two main attributes of enhanced speech are quality and intelligibility. The two measures are independent of each other. Typically quality as judged by listening is related to the amount and nature of the remnant noise after spectral subtraction. The intelligibility is related to the extent of speech spectrum distortion. While at high SNRs, it is possible to improve significantly the speech quality while retaining the intelligibility, this is not true at low SNRs.

In this section, we provide the results of objective tests which quantify the overall degradation of the noisy and enhanced speech with respect to the reference clean speech. We also provide the results of subjective tests to separately evaluate the intelligibility of the speech.

4.1.Objective quality results

Eight sentences from the TIMIT database spoken by 4 male speakers and 4 female speakers have been used. Three different background noises were taken from the SPIB database [8], white noise, factory noise and the babble noise. The overall SNR is computed according to ITU P.56 standard [9].

The objective quality measures used are segmental SNR, weighted spectral slope measure (WSS), Itakura-Satio distortion measure (IS) [10] and PESQ-MOS (Perceptual Evaluation of Speech Quality Mean Opinion Score, ITU-T recommendation P.862 [11]). Though the measures have been observed over a wide range of SNR -3dB, 0dB, 3dB, 5dB and 10dB, only few are tabulated due to limitation of space in Tables 1, 2 and 3. From the objective measures, informal listening and spectrograms we draw the following conclusions. Sound samples can be found at [12].

Objective measures: WSS measure which is considered to have high correlation with subjective tests [10] gave the most consistent results among all measures. In most of the cases, the algorithm which was judged the best by informal listening and by observing the spectrograms had the lowest WSS score. Besides the WSS scores also showed high correlation with the subjective quality test, A-B comparison test results. A-B comparison test involves presenting listeners with a sequence of

two speech test files (A and B) and asking them to decide whether they preferred file A or file B. In our case the files A and B are the speech files obtained by enhancing noisy speech files (corrupted with white noise at 5 dB SNR) using different algorithms. On the other hand, IS measure proved to have a low correlation with subjective tests [10]. The segmental SNR values of all methods did not vary much among the methods for a particular noise, but clearly indicated reduction in background noise in all cases. Similarly, though the MOS scores indicated clear improvement in overall quality of the noisy speech after applying any algorithm, the variation of the scores among the algorithms was too low and inconsistent to decide the best performance.

Method	0 dB SNR				5 dB SNR			
	SegSNR	MOS	WSS	IS	SegSNR	MOS	WSS	IS
Degraded	-5.42	1.44	74.2	4.61	-0.41	1.74	51.8	3.93
Ber-5	3.15	2.06	82.7	5.9	4.92	2.43	62.2	5.4
Ber-6	3.11	2.07	74.2	8.9	4.76	2.45	61.7	8.9
Ber-7	3.02	2.07	68.6	12.0	4.62	2.43	55.4	12.1
Param-1	2.71	2.03	60.0	7.7	4.23	2.39	50.1	9.2
Param-2	2.71	2.06	79.4	15.0	4.21	2.37	63.4	14.1
Param-3	3.04	2.11	66.6	4.3	4.65	2.45	51.7	5.1
Multi	3.11	2.10	73.5	6.4	4.92	2.44	55.8	7.2

Table 1. *Objective quality scores under white noise Ber-α_0: Berouti method with α_0 subtraction factor, Param-1,2,3: Parametric approach with 'η' and 'μ' as (0.99,0.05), (0.99,0.01), (0.98,0.05) respectively, Multi: Multi-band spectral subtraction*

Method	0 dB SNR				5 dB SNR			
	SegSNR	MOS	WSS	IS	SegSNR	MOS	WSS	IS
Degraded	-5.14	1.65	77.4	3.8	-0.1	2.01	63.9	2.96
Ber-5	2.42	1.99	86.5	7.9	4.44	2.43	69.5	5.6
Ber-6	2.42	1.98	82.5	10.6	4.36	2.42	68.7	8.1
Ber-7	2.38	2.00	80.4	12.9	4.26	2.41	68.5	10.6
Param-1	2.20	2.03	75.5	9.8	3.90	2.36	66.8	8.1
Param-2	2.42	2.06	82.5	10.6	4.36	2.42	68.7	8.1
Param-3	2.37	2.10	78.3	6.2	4.22	2.40	65.9	4.7
Multi	2.3	2.02	83.0	7.7	4.40	2.40	67.9	6.4

Table 2. *Objective quality scores for various algorithms under factory noise*

Method	0dB SNR				5dB SNR			
	SegSNR	MOS	WSS	IS	SegSNR	MOS	WSS	IS
Degraded	-4.75	1.77	76.7	3.2	0.26	2.1	60.1	2.62
Ber-5	1.34	1.94	112	6.0	3.92	2.28	74.7	4.6
Ber-6	1.38	1.92	110	8.5	3.87	2.27	70.7	6.6
Ber-7	1.38	1.90	109	10.8	3.82	2.26	68.2	8.6
Param-1	1.18	1.92	105	9.8	3.47	2.23	70.2	7.4
Param-2	1.17	1.77	120	15.0	4.21	2.37	83.2	14.1
Param-3	1.24	1.88	106	5.6	4.65	2.45	72.6	4.1
Multi	1.20	1.94	108	6.5	3.90	2.28	72.7	5.0

Table 3. *Objective quality scores for various algorithms under babble noise*

Comparison of algorithms: The parametric approach using set-1 values of 'η' and 'μ' (0.99 and 0.05) gave consistently better performance over the other algorithms. Both informal listening and spectrograms indicate good reduction in musical noise without increase in residual noise when compared to the other methods. The WSS score of it is the lowest among all algorithms in most cases clearly indicating its better performance; even the IS measure and the MOS scores are good enough. Besides most of the listeners preferred it to other algorithms in the A-B comparison tests. But slight smearing of the speech signal occurs due to averaging nature of the weighting function, because of which the intelligibility may be affected.

Optimum parameters: α_0 = 7 for the Berouti method and 'η' and 'μ' =0.99, 0.05 for the parametric method are the optimum values for good overall quality. They give the best trade-off between residual noise and musical noise.

Multi-band spectral subtraction: The multi-band spectral subtraction method which is supposed to work better than Berouti method in colored noise fails to do so. The reason for this can be attributed to two aspects, the usage of tweaking factors and the lack of accurate band SNRs. The authors [6] suggest the tweaking factors based on their assumption that most of the speech energy is concentrated below 1 kHz and hence lower subtraction is needed for avoiding speech distortion. But the suggested values are so widely separated that they dominate the variation of overall subtraction factor instead of the band SNRs. The band SNRs are also not accurate enough because of the inaccurate noise estimation in case of non-stationary noises.

4.2. Subjective test for intelligibility

A formal subjective intelligibility test, the modified rhyme test (MRT) [13] has been carried out to compare the parametric method with Berouti method. In the MRT, 50 sets of single-syllable words are used to test consonant intelligibility. Vowel intelligibility was included by adding one more set*: had, hid, hod,*

hud, head, heed. The listener hears one word from each set in the carrier phrase "Would you write..." and so on 51 times. The 6 possible words are presented on the test sheet. One speaker and four listeners were used. Percentage correct responses are scored for each listener for two trials and averaged across trials and listeners. While the clean speech had an intelligibility of 96%, the other results are listed in Table 4.

SNR (dB)	Configuration	Intelligibility (%)
0	Noisy	52.45%
	Ber-5	55.22%
	Ber-7	52.45%
	Param-1	50.00%
	Param-3	56.86%
5	Noisy	64.22%
	Ber-5	60.78%
	Ber-7	60.29%
	Param-1	61.27%
	Param-3	63.72%

Table 4. *Intelligibility for various algorithms under white noise*

From the results in Table 4, we observe that there is no significant change in speech intelligibility after enhancement. The parametric approach, Param-1, which was found to provide significantly better speech quality than any of the Berouti method configurations, is seen to be comparable with other methods in the context of intelligibility. Commenting on the effect of algorithm parameters on intelligibility, we observe that Ber-7 has slightly lower intelligibility than Ber-5 which can be attributed to higher speech distortion due to greater over-subtraction factor. Similarly, Param-1 has slightly lower intelligibility than Param-3 due to greater smearing of speech signal resulting from high value of 'η', the smoothing constant. But further comprehensive tests over wide range of SNR are needed to generalize these observations. . Besides at very low SNRs like 0 dB, it was observed that intelligibility is not only decided by the amount of speech distortion but also by the overall quality. This aspect also has to be investigated.

5.CONCLUSIONS

The constrained parametric approach, a method which essentially applies a (time-frequency) SNR-dependent spectrum weighting gives the best overall quality improvement among the three methods evaluated with comparable speech intelligibility. This demonstrates the potential of frequency dependent methods in reducing the remnant noise-speech distortion trade-off of linear spectral subtraction methods.

REFERENCES

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

[2] B.L. Sim, Y.C. Tong, J.S. Chang and C.T. Tan, "A Parametric formulation of the generalized spectral subtraction method,"*IEEE Trans. Speech and Audio Processing*, vol.6, no.4, pp.328-337, July 1998.

[3] Boll S.F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. ASSP-27, pp.113-120, Apr 1979.

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

[5] R.M. Crozier, B.M.G. Cheetham, C.Holt, and E.Munday, "Speech enhancement employing spectral subtraction and linear predictive analysis," *IEEE Trans. On Acoustic, Speech, and Signal Processing*, vol. ASSP-28, no.2, pp.137-145, April. 1980.

[6] Sunil D. Kamath and P.C.Loizou, "A multi-band spectral subtraction method for speech enhancement," *Proc. ICASSP,* pp. IV-4164, May 2002.

[7] Nathalie Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol.7, no.2, pp.126-137, March 1999.

[8] Signal Processing Information Base, "Noise data",http://spib.rice.edu/spib/select_noise.html, March 2003

[9] "Objective measurement of active speech level," ITU-T Recommendation P.56, Mar 1993.

[10] S.R. Quackenbush, T.P. Barnwell, M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, NJ, 1998.

[11] "Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation P.862, Feb.2001

[12] K.Manohar and Preeti Rao, "NCC-04 demo. Samples,"http://www.ee.iitb.ac.in/~prao/ncc2004/index.htm, Sept 2003

[13]Meyer Sound Laborataries, "Statistical measures of speech intelligibility," http://www.meyersound.com/support/papers/speech/section3.htm, Oct 2000