Spectral Subtraction Speech Enhancement with RASTA Filtering

Ram Singh

Department of Electrical Engineering,
Indian Institute of Technology, Bombay
rams.rcma@gmail.com

Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay
prao@ee.iitb.ac.in

Abstract

Spectral subtraction based speech enhancement methods are known to be effective for the suppression of additive stationary, broadband noise. Tonal noises such as car horn sounds are found to cause serious degradation of the output speech quality. A method is proposed in this work that incorporates RASTA processing within the framework of spectral subtraction in order to achieve better suppression of tonal noises. It is shown that the proposed method significantly outperforms both, spectral subtraction and RASTA speech enhancement methods, in the presence of simultaneous broadband and tonal noises.

1. Introduction

Suppression of acoustic background noise in single-channel noisy speech has been typically carried out by spectral subtraction (SS). The basic SS technique [1] involves subtracting an estimate of the noise spectrum from the noisy speech spectrum. Crucial to the performance of the technique is the voice activity detector (VAD) required to determine regions of speech pause for the noise estimate update. In stationary background noise, the SS method is simple and quite effective. Specific shortcomings include the presence of residual noise and an artifact known as musical noise. Satisfactory solutions to the problem of musical noise have been obtained by advanced approaches such as parametric spectral subtraction (PSS) [2] which adapt the subtraction parameters based on a computed a-priori SNR. A different approach to noisy speech enhancement is filtering by relative spectral processing (RASTA) [3][4] based on the assumption that the corrupting noise varies only slowly with respect to speech. The temporal trajectory of each short-time spectrum component is filtered to separate speech from noise. This approach has the advantage of not requiring a VAD. However it is not as effective as SS-based methods for broadband random noise backgrounds.

While the PSS method successfully suppresses stationary broadband noise, it is observed that certain strongly tonal noises disrupt the working of the VAD and remain unattenuated in the enhanced speech. A common example is car horn noise, ubiquitous in urban environments where cell phones are heavily used.

In this paper, the performances of the PSS and the RASTA methods are compared. A new method is proposed that combines the advantages of the two methods to achieve more effective suppression in the simultaneous presence of broadband and tonal noises. Objective and subjective quality test results of the enhanced speech are presented.

2. Parametric spectral subtraction method

Spectral subtraction based speech enhancement methods described in the literature can be represented by

$$\left|\hat{S}(k)\right|^2 = a_k\left|Y(k)\right|^2 - b_k\left|\hat{D}(k)\right|^2 \qquad (1)$$

where a_k and b_k are the algorithm parameters; $\hat{S}(k)$ is estimate of clean speech spectrum, $Y(k)$ is noisy speech spectrum and $\hat{D}(k)$ is estimate of noise spectrum. A parametric formulation of the basic spectral subtraction method based on statistical optimization [2] has been shown to alleviate the musical noise problem associated with spectral subtraction. In parametric spectral subtraction (PSS) method parameters a_k and b_k of Eq. (1) is set to 1 and the estimate of clean speech spectrum after performing the parametric optimization is given as

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

where "average a-priori SNR" $\xi(k)$ is

$$\xi(k) = (1-\eta)\underbrace{\frac{\left|S(k)\right|^2_{curr}}{\left|\hat{D}(k)\right|^2_{curr}}}_{\approx currentSNR} + \eta\underbrace{\frac{\left|\hat{S}(k)\right|^2_{prev}}{\left|\hat{D}(k)\right|^2_{prev}}}_{\approx previousSNR} \qquad (3)$$

where $\left|S(k)\right|_{curr} = \max\left(\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2, 0\right)$ and η is smoothing factor. A spectral floor with lower bound $\overline{\mu Y}$ is used to limit the attenuation of speech signal. The use of a-priori SNR provides temporal smoothing which is mainly responsible for suppression of musical noise contributed by random spectral peaks due to subtraction in the enhanced speech.

3. Relative spectral processing

Relative spectral processing [RASTA] based speech enhancement [3][4] involves linear filtering of the trajectory of the short-term power spectrum of noisy speech signal, as shown in Fig. 1. The spectral values of input speech signal are compressed by a nonlinear compression rule (a=2/3) before performing the filtering operation and expanded after filtering (b=3/2).

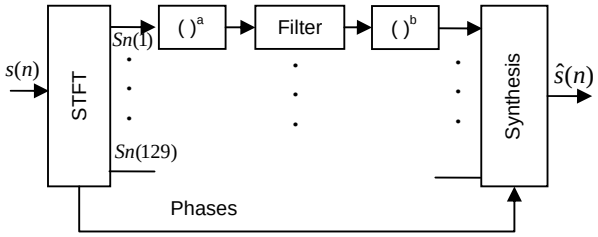

Figure 1: *Block diagram of RASTA method.*

To obtain better noise suppression for communication systems the fixed RASTA filters [3] were replaced by a bank of non-causal FIR Wiener-like filters [4]. For 256 point FFT, 129 unique filters are required. The output of each filter is given as

$$\hat{S}_i(k) = \sum_{j=-M}^{M} w_i(j)\, Y_i(k-j) \qquad (4)$$

Here, $\hat{S}_i(k)$ is estimate of clean speech in frequency bin "i" and frame-index "k", $Y_i(k)$ is noisy speech spectrum, $w_i(j)$ are the weights of the filter and M is order of the filter. In this method the weights $w_i(j)$ are obtained such that $\hat{S}_i(k)$ is least square estimate of clean speech $S_i(k)$ for each frequency bin i. The order $M = 10$ corresponds to 21 tap non-causal filters. The filters were designed based on optimization on 2 minutes of speech of a male speaker recorded at 8 kHz sampling over public analog cellular line from a relatively quiet library. The published response of the filter corresponding to bins in the frequency range 300 Hz to 2300 Hz is a band-pass filter, emphasizing modulation frequency around 6-8 Hz. Filters corresponding to the 150-250 Hz and 2700- 4000 Hz regions are low gain, low-pass filters with cut off frequency of 6 Hz. For very low frequency bins (0-100 Hz) the filters have flat frequency response with 0 dB gain. To perform the RASTA filtering an analysis window of 256 samples length with an overlap of 80 samples was used.

In our implementation the band specific non-causal FIR like Weiner filters are approximated by fourth order Butterworth filters and termed as frequency dependent RASTA (FD-RASTA) filter. For very low frequency band (0-100 Hz) no filtering is performed. The filters for the band (300-2300 Hz) are approximated by band-pass filter with lower cut-off frequency of 1 Hz and higher cut-off frequency of 15 Hz. The filters for the band (100-300 Hz) and (2300-4000 Hz) are approximated by low-pass filters with cut-off frequency of 10 Hz. To perform the FD-RASTA filtering an analysis window of 256 samples length with an overlap of 128 samples is used.

4. Performance Evaluation of PSS and RASTA Methods

The PSS and RASTA methods are experimentally compared in terms of objective speech quality measures on a data set of speech sentences corrupted with additive noise. The PSS method uses VAD [5] to detect the speech pauses to update the background noise. The objective tests used are weighted spectral slope (WSS) distance measure [6] and PESQ-MOS (Perceptual evaluation of speech quality mean opinion score, ITU-T recommendation P.862 [7]). The test speech sample used for objective tests is generated by concatenation of ten sentences from the TIMIT database [8] spoken by 5 male speakers and 5 female speakers. The different background noises [9][10] were used for generating the noisy speech sample. Though the measures have been observed over a wide range of SNRs and noises, only a few results are reported in Tables 1 and 2.

Table 1*: Objective quality score under white noise.*

Methods	0 dB SNR		5 dB SNR	
	WSS	PESQ - MOS	WSS	PESQ - MOS
Noisy speech	48.29	1.33	39.08	1.63
FD-RASTA	91.21	1.61	82.36	1.87
PSS	53.18	2.13	44.58	2.44

Table 2*: Objective quality score under horn noise.*

Methods	0 dB SNR		5 dB SNR	
	WSS	PESQ - MOS	WSS	PESQ-MOS
Noisy speech	137.29	1.77	113.22	2.03
FD-RASTA	76.56	1.98	70.98	2.16
PSS	114.92	1.91	90.25	2.20

From the reported objective quality results as well as informal listening and spectrogram analysis, it is observed that the speech enhanced by PSS method is superior to RASTA enhanced speech under white noise. Table 1 lists the objective quality score of noisy and enhanced speech under white noise. We observe that, as anticipated, the WSS distance for PSS enhanced speech is lower than FD-RASTA enhanced speech (implying an improvement in the speech quality). The anomaly in estimated WSS distance for noisy speech (i.e. WSS distance for noisy speech is better than that of the enhanced speech) is due to the simplicity of this objective measure and its consequent inability to predict different class of degradations which are subjectively perceived. The PESQ-MOS on the other hand is consistent with the subjectively perceived trend of an improvement in

speech quality. Under white noise the PSS method has better PESQ-MOS compared to FD-RASTA method. Under horn noise however, the FD-RASTA method has better WSS distance compared to PSS method and PESQ-MOS is comparable. The informal listening and spectrographic analysis reveals that the FD-RASTA method is more effective under horn noise.

5. Modified Speech Enhancement Method

The proposed modified speech enhancement method (MSE) combines the PSS and RASTA algorithmic techniques with a view to achieving simultaneous suppression of both random and tonal noises. The FD-RASTA filter is incorporated as a sub-module with SS algorithms and analyzed for its performance under white noise, pink noise, factory noise, horn noise, train noise and their combinations at different SNR levels. After systematic analysis of enhanced speech, it is observed that, by incorporating FD-RASTA as sub-module the horn noise is suppressed. However, white noise and its combination with horn noise causes severe distortion and musical noise in enhanced speech. To solve the distortion problem, the FD-RASTA filter is redesigned with higher pass-band and the spectral compression and expansion rule is set to a=2/2.15 and b=2.15/2. The further experiments carried out with redesigned FD-RASTA filter gives very low distortion, though musical noise is still present. White noise after FD-RASTA filtering gets converted into colored noise with sharp spectral peaks. Hence, the accuracy in noise estimation reduces; this causes the musical noise in enhanced speech. To solve this problem, the noise is estimated for filtered as well as unfiltered speech spectrum. The energy ratio of unfiltered and FD-RASTA filtered noise estimate is calculated and used along with unfiltered noise estimate for speech enhancement. The block diagram for MSE method is shown in Fig. 2. The expression for MSE method is given as

$$\begin{aligned} \left|\hat{S}_i(k)\right| &= G_i(k)\times\left|\overline{Y}_i(k)\right| \qquad if \quad \left|\hat{S}_i(k)\right| > \beta\left|D_i(k)\right|/\lambda_i \\ &= \beta\left|D_i(k)\right|/\lambda_i \qquad otherwise \end{aligned} \tag{5}$$

Here, $\left|\hat{S}_i(k)\right|$ and $\left|\overline{Y}_i(k)\right|$ are estimates of enhanced speech and filtered noisy speech spectra respectively under sub-bands used by FD-RASTA filters, similarly $\left|D_i(k)\right|$ is the estimate of unfiltered noise spectrum, $G_i(k)$ is gain function, β is spectral floor parameter set to 0.01 and λ_i is scaling factor for unfiltered noise estimate. The gain function $G_i(k)$ is calculated as

$$G_i(k) = \sqrt{\frac{\xi_i(k)}{\alpha + \xi_i(k)}} \tag{6}$$

Here, α is over subtraction factor calculated as $\alpha = \alpha_0 - s\times(SNR)$, the parameter α_0 is over-subtraction factor at 0 dB SNR and s is slope set to 1/20. $\xi_i(k)$ is a-priori SNR calculated as

$$\xi_i(k) = (1-\eta)\underbrace{\frac{\left|Y_i(k)\right|^2_{curr}}{\left|D_i(k)\right|^2_{curr}/\lambda_i}}_{\approx currentSNR} + \eta\underbrace{\frac{\left|\hat{S}_i(k)\right|^2_{prev}}{\left|D_i(k)\right|^2_{prev}/\lambda_i}}_{\approx previousSN\ R} \tag{7}$$

The RASTA filtered noisy speech spectrum is enhanced using unfiltered noise estimate (having smooth spectrum as compare to RASTA filtered noise estimate). To avoid over attenuation scaling factor λ_i is used for unfiltered noise estimate. The scaling factor λ_i is energy ratio of unfiltered ($\left|D_i(k)\right|$) and filtered ($\left|\overline{D}_i(k)\right|$) noise estimate given as

$$\lambda_i = \frac{\sum_{b_i}^{e_i}\left|D_i(k)\right|^2}{\sum_{b_i}^{e_i}\left|\overline{D}_i(k)\right|^2} \tag{8}$$

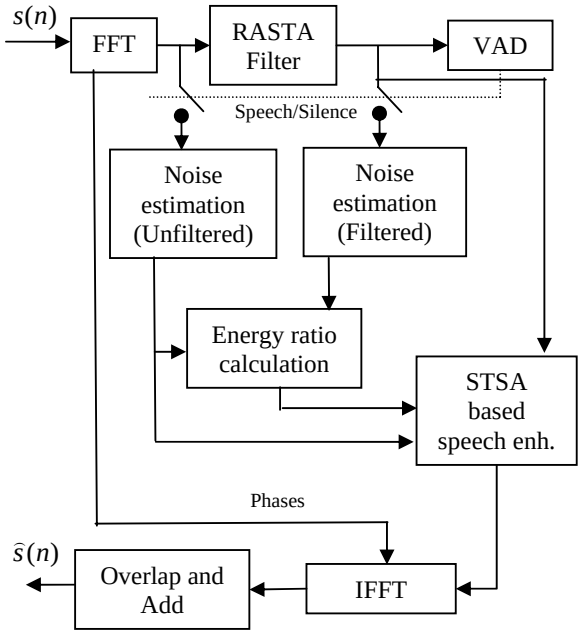

Figure 2: *Block diagram for modified speech enhancement (MSE) method.*

6. Performance Comparison of PSS and MSE Methods

The performances of PSS and MSE methods are compared based on the spectrographic analysis, quality and intelligibility of enhanced speech. Fig. 3 shows the

spectrograms of clean, noisy and enhanced speech for an utterance corrupted with additive white plus horn noises at overall SNR of 0 dB. We see that the PSS method suppresses the white noise to a large extent (with some loss of speech) but leaves the strong harmonics of the horn noise relatively unattenuated. The MSE method achieves better overall noise suppression. The quality of enhanced speech is evaluated using a suite of objective measures (WSS and PESQ-MOS) on a set of noisy sentences at various SNRs. The results are presented in Tables 3 and 4.

Table 3*: Objective quality score under horn noise.*

Methods	0 dB SNR		5 dB SNR	
	WSS	PESQ-MOS	WSS	PESQ-MOS
Noisy speech	137.29	1.77	113.22	2.03
PSS	114.92	1.91	90.25	2.20
MSE	87.60	1.97	72.44	2.28

Table 4*: Objective quality score under white+horn noise.*

Methods	0 dB SNR		5 dB SNR	
	WSS	PESQ-MOS	WSS	PESQ-MOS
Noisy speech	81.06	1.22	61.48	1.50
PSS	88.74	1.60	67.58	2.01
MSE	100.10	1.64	82.97	2.02

The modified rhyme test (MRT) [11] is carried out to measure the intelligibility of the enhanced speech. One speaker and six listeners were used. Percentage correct responses are scored for each listener and averaged across listeners. While the clean speech had an intelligibility of 96%, the intelligibility of speech enhanced under different noise condition at various SNR levels is listed in Table 5. The A-B comparison test is a subjective listening test carried out to compare the quality of speech enhanced by PSS and MSE method. One speaker and six listeners were used in this subjective test. The listeners indicate which of the two processed utterances they prefer. The test results under horn and wihte+horn noise conditions at various SNRs appear in Table 6.

Table 5*: Intelligibility for PSS and MSE methods.*

SNR (dB)	Configuration	Intelligibility Under horn noise (%)	Intelligibility Under white+horn noise (%)
0	PSS MSE	53.4 58.5	37.5 39.5
3	PSS MSE	59.4 63.9	45.3 49.2
7	PSS MSE	61.2 65.4	59.1 61.5

Table 6: *A-B comparison test result under white+horn noise.*

Preferred configuration	Horn			White+horn		
	0 dB	3 dB	7 dB	0 dB	3 dB	7 dB
PSS(%)	11	14	23	12	20	31
MSE(%)	76	71	66	75	62	38
Neutral(%)	13	15	21	13	18	31

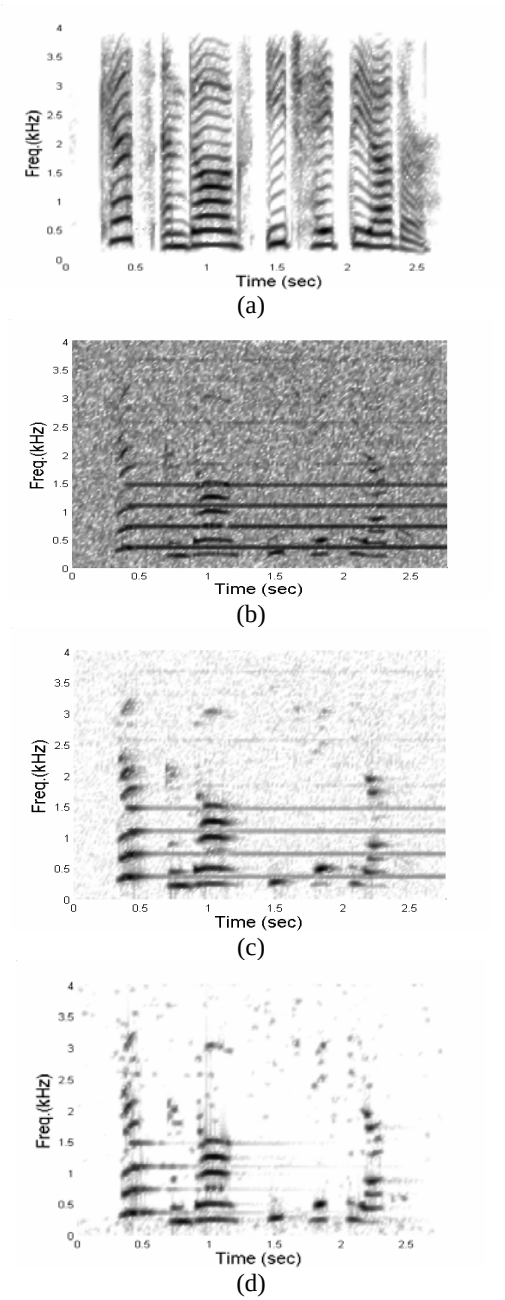

Figure 3: *Spectrograms of speech sample "Cats and dogs each hate the other" (a) clean, (b) noisy (c) PSS enhanced (d) MSE enhanced speech under white+horn noise at 0 dB SNR.*

7. Conclusions

The MSE method, proposed in this paper, incorporates RASTA processing within the framework of spectral subtraction in order to achieve better suppression of tonal noises such as car horn. Based on the objective quality tests and subjective listening tests for intelligibility and quality, it is observed that the speech enhanced by MSE method achieves improved quality and intelligibility under horn noise and white plus horn noise conditions. Under other noise conditions, the MSE method is observed to be comparable to the PSS method. Thus we conclude that the proposed method significantly outperforms both PSS and RASTA speech enhancement in the presence of simultaneous broadband and tonal noises.

8. References

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

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

[3] Hermansky, H. and Morgan, N., "RASTA processing of speech", *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 578-589, October, 1994.

[4] Hermanskey, H., Wan, E.A. and Avendano, C., "Noise suppression in cellular communications", *2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTT94), September 26-27, Kyoto, Japan, 1994.*

[5] Marzinzik, M. and Kollmeier, B., "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics", *IEEE Trans. Speech and Audio Processing*, vol.10, no.2, pp.109-118, February, 2002.

[6] Schuyler, R.Q., Thomas, P.B. and Mark, A.C., Objective measures of speech quality, Prentice Hall Publication, 1988

[7] 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, February, 2001.

[8] N. Institute of Standards, The TIMIT cdrom, 1989.

[9] Signal Processing Information Base, Noise data, http://spib.rice.edu/spib/select/noise.html, January, 2001.

[10] Horn sounds library http://www.autospeak.com/ library.htm , 12th December, 2005

[11] Meyer Sound Laboratories, Statistical measures of speech intelligibility, http://www.meyersound.com/ support/papers/speech/section3.htm, 21st December, 2005.