EFFICIENT SNR-BASED SUBBAND POST-PROCESSING FOR RESIDUAL NOISE REDUCTION IN SPEECH ENHANCEMENT ALGORITHMS

Frédéric Mustiere, Martin Bouchard, Miodrag Bolic

School of Information Technology and Engineering,
University of Ottawa, Ottawa, ON, K1N 6N5, Canada

ABSTRACT

While current speech enhancement algorithms can significantly reduce background noise, the output speech is commonly unacceptably damaged – a strong penalty for sensitive applications. Alternatively, reducing the aggressiveness leads to more background residual noise – another rejection criterion in practice. In this work, a cost-effective technique for residual noise reduction is presented as a postprocessor for less aggressive enhancement algorithms. The main motivation is to keep their beneficial characteristics, and use the noisy and pre-enhanced signals to remove the remaining noise. The proposed method decomposes pre-enhanced signals into subbands, then performs framewise scaling of the downsampled subband time series based on the estimated Signal-to-Residual-Noise Ratio. Since many popular enhancement algorithms already operate in subbands, the application of the postprocessor is appealing from a computational standpoint. Results show the method consistently reduces background noise, with no further apparent speech damage, as reported by several objective measures and informal listening experiments.

1. INTRODUCTION

One of the central issues in speech enhancement consists of the tradeoff between noise reduction and intelligibility [1], and it is in fact rare for a method to consistently improve intelligibility. Rather than trying to improve it, practitioners usually set the more reasonable goal of at least not affecting it in the noise removal process. In sensitive applications where intelligibility and naturalness are important, non-aggressive setups for speech enhancement algorithms are thus privileged, at the cost of the presence of a larger amount of background residual noise in the enhanced speech. In this work, a post-processing technique is proposed with the following objectives in mind:

1. Remove surplus background residual noise while retaining the positive features of (pre)enhanced speech (i.e. intelligibility, low distortion, naturalness, etc)
2. As simple and efficient implementation as possible (i.e, aim for low computational complexity).

Both objectives are treated here with equal importance: indeed, if the second objective is not respected, one might as well rework and upgrade the pre-enhancement scheme. On the other hand, if the first objective can be attained with very small additions, then the appeal is more significant for real-world applications already employing certain well-established algorithms. Indeed, in many real-word applications, real time requirements are to begin with hardly met and hence we are interested in improving performance with adding very little computational requirements. Such a concern would for example be applicable to the post-processing method shown in [2], in which the non-negligible additional workload consists of a harmonic analysis combined with pitch tracking on the pre-enhanced signal, followed by a (pre-trained) codebook mapping for the restoration of the parts of the signal that were damaged during the initial noise reduction algorithm. In addition, note that the primary goal of restoring damaged speech components is fundamentally different from our first objective of removing excessive residual noise.

In this work, the objective of the post-processor is not enhancement per se, but rather noticeable background noise removal. Other methods with similar objectives have appeared in the literature; for example the post-filtering method of [3], based on the detection of formant locations and spectral valleys, is found to perform well for narrowband speech in AWGN. In contrast, the proposed post-processor shown in this paper is designed to be incorporated naturally as a module to already existing subband enhancement architectures, and is meant to operate in the same complex noise conditions. The paper is organized as follows: In Section II, the procedure is formally introduced, accompanied by qualitative explanations. In Section III, several tests are performed with well-established algorithms in various conditions (babble, factory, military vehicle noise, and car interior noises). The performance between direct output quality and post-processed quality is compared using several objective measures and the conclusions of informal listening tests are reported. Then, conclusions are given in Section IV.

2. THE PROPOSED POST-PROCESSING SCHEME

In simple terms, the idea consists of scaling, on a frame-by-frame basis, the pre-enhanced signals depending on the respective estimated levels of speech and residual noise. However, even in ideal conditions, it is not desirable to apply such volume-scaling in a fullband setup, as it would perceivably modulate the amplitude of the signal in a potentially disturbing manner. Thus, the method is chosen to be applied in the subband domain, as in the generic structure shown in Figure 1. A similar form of subband-signal scaling structure has been successfully applied as the core of a "standalone" subband speech enhancement algorithm (as opposed to a mere "post-processor") in [4], where subband gains are directly applied to the incoming noisy speech, and are determined from a VAD-based estimation of the a posteriori Signal-to-Noise Ratio. In our context however, the goal is to determine scaling factors to be applied to pre-enhanced subband speech signals, for which an estimate of the SNR has already been determined or is directly accessible.

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for funding this work.

For simplicity, assume that each subband domain signal (i.e., each of the decimated signals at the outputs of the filters of the filterbank) are here real-valued and locally viewed as time-domain signals.

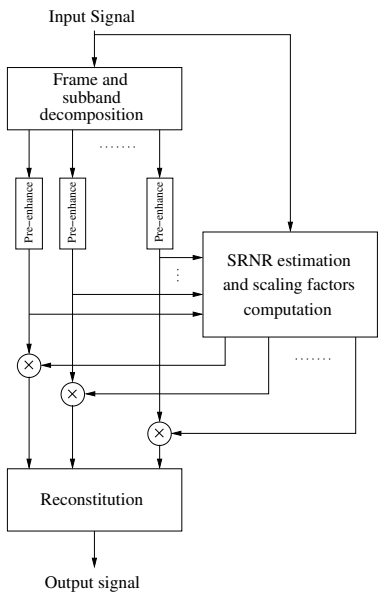

Figure 1: The proposed post-processing scheme. The Signal to Residual Noise Ratio (SRNR) is estimated from both the noisy fullband and the enhanced subband signals to produce scaling factors which are then applied before reconstruction.

To determine the scaling function in this context, we begin by assuming that the speech and noise statistics are fixed over small frames. Denote by $y_m(:,i)$ the pre-enhanced decimated speech vector at subband m and at the ith frame, assumed to contain the sum of the clean subband vector $x_m(:,i)$ and some residual noise $r_m(:,i)$. Next, suppose that over the ith frame, $x_m(:,i)$ and $r_m(:,i)$ are approximately i.i.d. with respective distributions $\mathcal{N}(0;\sigma_x(i)^2)$ and $\mathcal{N}(0;\sigma_r(i)^2)$ (the sequences can indeed be negative-valued, as opposed to spectral amplitudes in usual frequency-domain processing for example). With these assumptions, it is easy to show that, for all k indexing the subband frame:

$$p\left(x_m(k,i)|y_m(:,i)\right) =$$

$$\mathcal{N}\left(x_m(k,i)|y_m(k,i)\frac{\sigma_x(i)^2}{\sigma_x(i)^2+\sigma_r(i)^2};\frac{\sigma_x(i)^2\sigma_r(i)^2}{\sigma_x(i)^2+\sigma_r(i)^2}\right) \quad (1)$$

From the above, we can thus write the conditional expected value of $x_m(:,i)$ in terms of the Signal-to-Residual-Noise-Ratio, denoted here by $SRNR_m(i)$, to obtain the post-processed enhanced series $\hat{x}_m(:,i) = \mathcal{E}\left(x_m(k,i)|y_m(:,i)\right)$ as:

$$\hat{x}_m(:,i) = \left(1+SRNR_m(i)^{-1}\right)^{-1} y_m(:,i) \quad (2)$$

The gain function is shown in Figure 2. As the reader will have noted, in its form the gain function given in Eqn. 2 is essentially superimposable to an SNR-based frequency-domain filtering formulation of a spectral subtractive gain. Besides the distinct decimated filterbank context and assumptions regarding the nature of the intervening signals,

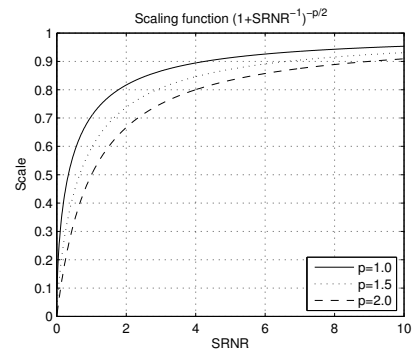

Figure 2: The proposed subband scaling function. When the subband Signal-to-Residual-Noise Ratio *SRNR* is low, the subband frame is strongly scaled down. As it can also be seen, the variable $p \geq 1$ is an aggressiveness factor.

there are notable differences of practical nature: what is proposed here is to reduce the gain to a single number per band and per frame – i.e., to locally reduce it to a scaling factor. In other words, we take advantage of the time/frequency localization of each small frame of data at the output of the decimated filters to formulate some simplifying assumptions resulting in the application of a fixed gain within one subband over a few consecutive samples. To respect this criterion, a "medium" amount of subbands and a relatively small subband frame size is required. As an important practical advantage, our proposed method is both embeddable in existing enhancement algorithms already using filterbanks, and the resulting scaling is very efficient.

Obviously, the above requires the knowledge of $SRNR_m(i)$, which is difficult to accurately estimate as it strongly depends on the method/algorithm used and on the noise conditions. Nevertheless, a practical solution consists of estimating it from $SNR_m(i)$, the Signal-to-Noise Ratio in the current subband frame (obtainable from the pre-enhancement stage) – the two are indeed strongly correlated. For this purpose, several methods can be envisioned: For example, using various training data obtained specifically using the chosen pre-enhancement algorithm, some mathematical relationship (e.g. linear regression) between the two sets of subband *SNR*s could be obtained. On the other hand, a complex scheme will threaten the crucial simplicity objective stated in our introduction. While attempting to efficiently approximate it, heuristically it was found that satisfactory results can be obtained by using the simple following rule:

$$SRNR_m(i) \simeq \max\{SNR_m(i),\ SNR(i)\} \quad (3)$$

In the above rule, the practical value used to represent the residual noise ratio in each subband is simply taken as the maximum between the fullband estimated *SNR* and the current subband estimated *SNR*. The rationale for incorporating the fullband SNR was initially based on the observation that in many situations the "local" subband SNR is found to be in discordance with the fullband *SNR* and thus some low-amplitude speech components that are still important for intelligibility are more at risk of being scaled down. Note also that from Eqn. 3 we necessarily have $SRNR_m(i) \geq SNR_m(i)$, which is consistent with the expected effect of the pre-enhancement scheme. In practice, to further

account for the effect of pre-enhancement, we found that the introduction of a parameter $p \geq 1$ is also beneficial and provides an accessible aggressiveness parameter, so as to obtain the final rule:

$$\hat{x}_m(:,i) = \left(1 + SRNR_m(i)^{-1}\right)^{-\frac{p}{2}} y_m(:,i) \qquad (4)$$

In our implementations, p is set to 1.15. The use of Eqn. 4 allows for a very low-cost post-processing (one of our primary goals), while the effectiveness of the above solution will be confirmed in practical tests in Section 3.

Regarding once again computational complexity, note that if the pre-enhancement scheme is already frame-based and employing subbands, the overall computational overhead is minimal.

3. EXPERIMENTAL RESULTS

3.1 Speech and noise material

The audio material used in this paper has a sampling frequency of 20 kHz. The clean speech material is obtained by concatenating multiple speakers from the TIMIT database [5], and inserting silences in order to obtain a 60% activity rate (as recommended for objective quality estimation in [6]). The total length of the clean speech material is approximately 30 seconds. The noise data was obtained from http://spib.rice.edu/spib/select_noise.html, containing examples from the NOISEX-92 database [7]: the babble, factory, and military vehicle noises were used. The obtained noisy signals are scaled with 3 different values to obtain various conditions, from low to high input SNR. Thus, 9 different conditions were tested for 3 different algorithms.

3.2 Objective quality measures

In order to illustrate the performance of our postprocessor, we resort to several objective speech quality measures. The tools used are the SNR and the average segmental SNR (ASNR) [8]; the Coherence Speech Intelligibility Index (CSII) [9]; the wideband extension for the PESQ score (WPESQ) from [10]; and the three composite measures shown in [1], meant to reflect the level of speech distortion (Csig), the level of background noise intrusiveness (Cbak), and the overall quality (Covl). For the CSII and WPESQ, the signals are resampled at 16 kHz beforehand, and for the Csig, Cbak, and Covl, at 8 kHz.

3.3 Choice of enhancement algorithms used as preprocessors and subband decomposition

First, in order to test our post-processor, we choose two well-established and well-recognized algorithms:

- The statistical-based (LMMSE) algorithm, presented in [11]. The performance of this method is very well rated amongs various algorithms (see [13]).
- The multi-band spectral subtractive algorithm (MSSUB) shown in [12], which is shown to largely outperform the traditional spectral subtraction algorithm.

The MATLAB implementations for these two methods were directly used from the accompanying CD-Rom from [1]. The post-processed version of these algorithms are denoted by LMMSE-P and MSSUB-P. Next, our postprocessor is chosen to operate in 32 subbands, obtained via pseudo-QMF filterbanks decomposition [14]. For further illustrative purposes, the LMMSE algorithm above is adapted to function in these subbands, in a configuration identical to that shown in Figure 1. To distinguish this method with the full-band LMMSE, we denote it by LMMSE-S.

3.4 Results and analysis

For the sake of clarity and concision, the output scores were averaged over all types of noise for each of the three SNR conditions – yielding three tables. Some example waveforms in military noise conditions are given in Figure 3.

Average scores for Low SNR conditions							
Type of algorithm	Objective measures						
	SNR	ASNR	CSII	WPESQ	Csig	Cbak	Covl
Noisy	-0.32	-5.16	0.04	1.04	1.39	1.22	1.12
LMMSE	9.02	0.22	0.77	1.21	1.48	1.60	1.25
LMMSE-P	10.61	1.91	0.97	1.32	1.60	1.76	1.38
Difference	1.59	1.69	0.20	0.11	0.12	0.16	0.13
LMMSE-S	10.04	0.85	0.86	1.26	1.53	1.67	1.35
LMMSE-S-P	11.34	2.30	0.97	1.36	1.66	1.78	1.40
Difference	1.30	1.45	0.11	0.10	0.13	0.11	0.05
MSSUB	9.45	-0.05	0.86	1.28	1.71	1.73	1.45
MSSUB-P	10.01	1.21	0.97	1.41	1.73	1.85	1.52
Difference	0.56	1.26	0.11	0.13	0.02	0.12	0.07

Table 1: Average results for **low SNR** input conditions. Each result reported in this table are an average over 4 simulations for each method (corresponding to the 4 types of noise used).

Average scores for Medium SNR conditions							
Type of algorithm	Objective measures						
	SNR	ASNR	CSII	WPESQ	Csig	Cbak	Covl
Noisy	5.69	-1.64	0.89	1.12	1.88	1.64	1.50
LMMSE	12.83	2.83	0.98	1.43	1.87	2.02	1.62
LMMSE-P	13.26	3.91	0.99	1.52	1.85	2.10	1.70
Difference	0.43	1.08	0.01	0.09	-0.02	0.08	0.08
LMMSE-S	13.45	3.23	0.99	1.48	2.03	2.04	1.70
LMMSE-S-P	13.80	4.04	0.99	1.54	2.10	2.09	1.75
Difference	0.35	0.81	0	0.06	0.07	0.05	0.05
MSSUB	13.24	2.73	0.98	1.65	2.26	2.22	1.95
MSSUB-P	13.43	3.56	0.99	1.78	2.34	2.37	2.24
Difference	0.19	0.83	0.01	0.13	0.08	0.15	0.29

Table 2: Average results for **medium SNR** input conditions. Each result reported in this table are an average over 4 simulations for each method.

Average scores for High SNR conditions							
Type of algorithm	Objective measures						
	SNR	ASNR	CSII	WPESQ	Csig	Cbak	Covl
Noisy	11.71	2.26	0.99	1.40	2.41	2.17	1.99
LMMSE	16.20	5.32	0.99	1.66	2.25	2.40	1.97
LMMSE-P	16.73	6.34	0.99	1.71	2.31	2.47	2.06
Difference	0.53	1.02	0	0.05	0.06	0.07	0.09
LMMSE-S	15.98	5.16	0.99	1.68	2.36	2.35	1.99
LMMSE-S-P	16.64	5.98	0.99	1.73	2.38	2.37	2.11
Difference	0.66	0.82	0	0.05	0.02	0.02	0.12
MSSUB	15.25	4.86	0.99	2.01	2.71	2.63	2.38
MSSUB-P	15.80	5.69	0.99	2.12	2.72	2.68	2.45
Difference	0.55	0.83	0	0.11	0.01	0.05	0.07

Table 3: Average results for **high SNR** input conditions. Each result reported in this table are an average over 4 simulations for each method.

First of all, observing Tables 1, 2, and 3, it is clear that the post-processor consistently increases the objective scores obtained by the enhancement algorithms. This is especially the case for the average segmental SNR, the WPESQ, but

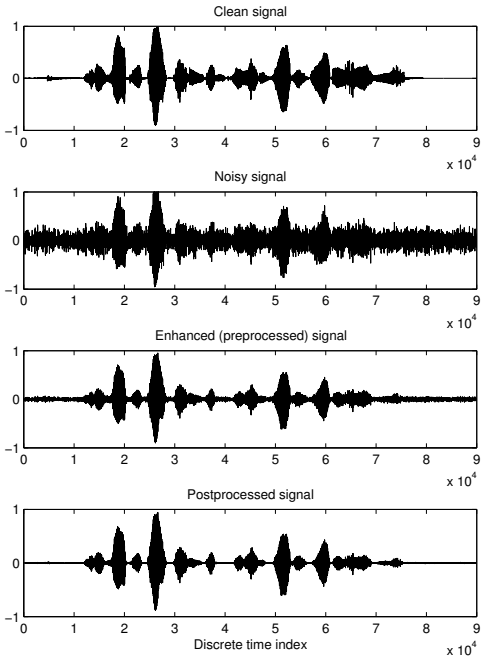

Figure 3: Visual example of the effect of the post-processor. The background noise is clearly reduced.

also the SNR and the Cbak measures. Moreover, it appears that the most benefits are seen at medium and low SNR, which correspond to situations where improvements are most needed. Interestingly, the ASNR (the measure that is most increased) and the Cbak measures have been shown to be mostly correlated with the level of background noise intrusiveness [1,15], and thus these results are consistent with our objective of reducing the residual noise. From informal listening tests, we also find that the proposed post-processor is able to remove a significant amount of background noise. This is particularly noticeable when no speech is present, but it can also be heard during speech utterances, especially when the original noise contains high frequencies. This is well observed in Figure 3. Note that the WPESQ, Csig, Cbak, and Covl measures do not take into account silences, confirming that the postprocessor also provides benefits *during* speech. Some audio demonstrations can be found at http://www.site.uottawa.ca/∼bouchard/papers/Eusipco_RNR.zip

4. CONCLUSION

This paper introduced a very simple and low-complexity add-on to speech enhancement algorithms, and it was shown that it can reduce the excess of residual noise in the enhanced speech without further damaging the remaining speech. The method is particularly advantageous when the enhancement algorithm used operates in subbands, in which case the additional complexity is minimal.

REFERENCES

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

[2] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, no. 4, May 2007.

[3] F-M. Wang, P. Kabal, R. Ramachandran, D. O'Shaugnessy, "Frequency domain adaptive post-filtering for enhancement of noisy speech," in *Speech Communication*, Vol. 12, no. 1, pp. 41–56, February 1993.

[4] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," Ch. 3 in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, eds., Kluwer Academic Publishers, 2004.

[5] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett and N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS Speech Disc," NTIS order number PB91-100354, 1990.

[6] International Telecommunication Union, *Recommendation P. Supplement 23: ITU-T coded-speech database*, Geneva, 1998.

[7] A. Varga, H.J.M. Steeneken, M. Tomlinson, D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Technical report, Defence Evaluation and Research Agency, Speech Research Unit, Malvern, United Kingdom, 1992.

[8] J.H.L. Hansen and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *ICSLP*, Vol. 7, pp. 2819–2822, Sydney, Australia, 1998.

[9] J. Kates and K. Arehart, "A model of speech intelligibility and quality in hearing aids," in *Proceedings of the IEEE Workshop on Applications of Signal processing to Audio and Acoustics*, New Paltz, New York, 2005.

[10] International Telecommunication Union, *Recommendation P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, Geneva, 2005.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 33, Issue 2, pp. 443–445, April 1985.

[12] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *ICASSP*, Orlando, 2002.

[13] Y. Hu and P.C. Loizou, "Subjective comparison of speech enhancement algorithms," in *ICSLP*, Pittsburg, USA, 2006.

[14] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," in *IEEE Transactions on Signal Processing*, Vol. 42, Issue 1, January 1994.

[15] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *ICSLP*, Pittsburg, PA, USA, 2006.