NOISE-DEPENDENT POSTFILTERING

Volodya Grancharov, Jonas Samuelsson and W. Bastiaan Kleijn

KTH (Royal Institute of Technology)
Department of Signals, Sensors and Systems
10044 Stockholm, Sweden
{volodya.grancharov,jonas.samuelsson,bastiaan.kleijn}@s3.kth.se

ABSTRACT

This paper introduces a modification of the commonly used postfilter that improves performance when acoustic background noise is present. The modification consists of replacing the nonadaptive postfilter parameters that govern the degree of spectral emphasis (commonly denoted as γ_1 and γ_2) with parameters that adapt to the noise statistics. We describe an effective mapping from the noise statistics to the emphasis parameters and provide a low complexity noise estimation algorithm that is sufficient for this application. The resulting noise-adaptive postfilter successfully attenuates the background noise and naturally converges to the conventional postfilter at high SNR conditions. Thus, the speech enhancement problem is solved with minimal modification of legacy codecs, since the existing structure of the speech codec is used. Test results indicate that the presented algorithm significantly outperforms the standard postfilter with non-adaptive parameters.

1. INTRODUCTION

The performance of a speech communication system can degrade in the presence of acoustic background noise and quantization noise. Due to their different nature, these two problems have been addressed independently. The perceived quantization noise is typically reduced by means of a postfilter [1], [2]. Background noise is attenuated by noise suppression systems such as Spectral Subtraction schemes and Wiener filtering [3], [4]. In this paper, we achieve both goals simultaneously by means of a simple extension of the conventional postfilter.

Postfilters are used in most speech-coding standards. They reduce the effect of quantization noise in a low bit-rate speech codec by emphasizing the formant frequencies and deemphasizing the spectral valleys. The transfer function of the most commonly used postfilter [5] is given by

$$H(z) = GH_s(z). \tag{1}$$

In equation 1, G is a gain factor and $H_s(z)$ is a filter of the form

$$H_s(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}(1 - \mu\, z^{-1}), \tag{2}$$

where $A(z)$ is the adaptive short term prediction-error filter, γ_1 and γ_2 are fixed *emphasis parameters* that control the degree of spectral emphasis (the frequency response) and μ controls the tilt compensation filter. The factor G aims to compensate for the gain difference between synthesized speech $s(n)$ and postfiltered speech $s_f(n)$. It is continuously adapted to the local-energy ratio of the noisy and the unnormalized postfiltered signal, but not to the background noise statistics.

The fixed emphasis parameters γ_1 and γ_2 are usually optimized based on listening tests for clean speech. To keep the values of emphasis parameters and the spectral-tilt parameter μ constant is consistent with the notion that the quantization noise of a codec can be approximated as a white noise of known energy. However, there is no apparent motivation for the current practice of using the same parameter values for noisy input speech. In this paper, we will show that adaptation of γ_1 and γ_2 to the noise level can improve performance significantly.

The motivation for adapting the emphasis parameters γ_1 and γ_2 is strengthened by a qualitative comparison of the behavior of a postfilter and a Wiener filter. We note that a Wiener filter renders an optimal (in the mean-square sense) clean-signal estimate from a noisy signal that is the addition of independent, stationary noise and clean-signal sources, given the power spectrum of the noise. A Wiener filter attenuates spectral regions of the noisy signal increasingly with decreasing SNR. In most practical settings this means that the spectral valleys of the noisy speech signal are attenuated and the formants structures are largely unaffected. The result is an emphasis of formants over spectral valleys that becomes stronger with decreasing overall SNR. The main qualitative difference between the Wiener filter and the conventional postfilter is that the emphasis of formants over spectral valleys does not depend on the noise level in the postfilter. This further motivates us to study the usage of the postfilter for suppression of the background noise through the adaptation of its parameters.

It has been noted in earlier work [5], [6] that the postfiltering concept can also be used for the suppression of background noise. However, no studies on how to design a postfilters for the application of background noise suppression were undertaken.

2. POSTFILTER WITH NOISE-DEPENDENT EMPHASIS

The proposed structure of a noise-dependent postfilter is illustrated in the block diagram of Figure 1. The reconstructed noisy signal is used as input to a routine that estimates the noise statistics. The noise statistics are mapped to an appropriate set of emphasis parameters. The emphasis parameters are used in a conventional postfilter structure.

In this section, we discuss practical methods that can be used to estimate the noise statistics and to obtain the mapping from the noise statistics to the emphasis parameters. We start with the discussion of our estimation of the noise statistics and follow that by a discussion of the mapping.

This work was funded by Nokia Corporation.

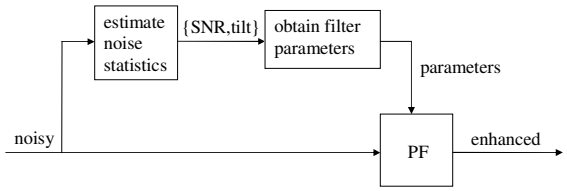

Fig. 1. The general scheme of the noise-dependent postfilter.

2.1. Estimation of the Noise Statistics

Existing noise estimation algorithms, such as [7], provide detailed estimates of the power spectral density of the acoustic background noise. The high computational effort required for such algorithms can be justified by the detailed description obtained. However, in contrast to many conventional speech enhancement systems such as Wiener filters, the enhancement by means of a postfilter is governed by only two parameters (γ_1 and γ_2) and this means that a detailed characterization is not necessary. Moreover, the computational effort of the existing noise estimation algorithms is not justified. Thus, we have chosen to characterize the noise statistics by the SNR and the spectral tilt only. The spectral tilt is quantified as the predictor coefficient of a first-order predictor filter. The results of section 3 confirm the soundness of this choice.

Our noise estimation algorithm is based on the same principle as that of [7]. However, in our case only the noise energy and the spectral tilt are estimated, significantly reducing the computational burden. The main steps of the algorithm are given in Table 1.

1. Initialize energy buffer $\{B_e(i)\}_{i=0,\cdots,N-1}$ and tilt buffer $\{B_t(i)\}_{i=0,\cdots,N-1}$:
   $B_e(i) = 0,\ i = 0,\cdots,N-1$
   $B_t(i) = 0,\ i = 0,\cdots,N-1$
2. For each successive frame j perform:
   (a) Update energy and tilt buffers:
   $B_e(i) = B_e(i-1),\ i = N, N-1,\cdots,1$
   $B_e(0) = e(j)$
   $B_t(i) = B_t(i-1),\ i = N, N-1,\cdots,1$
   $B_t(0) = t(j)$
   (b) Obtain noise energy estimate $\widehat{e}_n(i)$ and noise tilt estimate $\widehat{t}_n(i)$
   i. $\widehat{e}_n(i) = \min_{i=0,\cdots,N-1} B_e(i)$
   ii. $\widehat{t}_n(i) = B_t(\mathrm{argmin}_{i=0,\cdots,N-1} B_e(i))$

Table 1. Noise estimation algorithm. $e(j)$ and $t(j)$ are the energy and tilt of the current frame and $\widehat{e}_n(j)$ and $\widehat{t}_n(j)$ are the estimated noise energy and tilt.

Table 2 provides the estimation results of the noise spectra tilt for a frame size of 20 ms and a buffer size of 30 frames averaged over a database. The numbers were averaged over a database that consisted of ten clean speech sentences from the TIMIT [8] database that were contaminated with three types of stationary noise sources. The values in the column "True Tilt" were calculated over the noise frames and the values in the column "Estimated Tilt" were given by the noise estimation algorithm described above. Figure 2 illustrates the performance of the algorithm in terms of SNR estimation. In this example, the clean speech sentence is contaminated with white noise at 15 dB.

Noise Type	True Tilt	Estimated Tilt
Car 5 dB	0.99	0.96
Babble 10 dB	0.86	0.89
White 0 dB	0.04	0.08

Table 2. The average noise tilt.

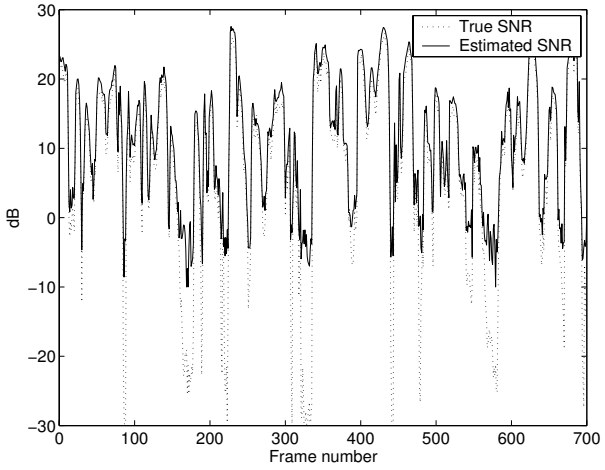

Fig. 2. True and estimated SNR.

2.2. Mapping from Noise Statistics to Emphasis Parameters

To adapt the postfilter to the noise statistics, a mapping from the SNR and the spectral tilt to the emphasis parameters γ_1 and γ_2 was implemented. Our objective was to create a mapping that selects the emphasis parameters that result in a minimum mean spectral distortion.

In our implementation we consider only postfiltering of the spectral envelope. Since we do not consider the spectral fine structure, the spectral distortion measure must be based on the spectral envelope only. We can use autoregressive (AR) modelling of the clean and noisy speech for this purpose. Let $A_y^{-1}(e^{j\omega})$ be the AR spectral envelope (corresponding to the transfer function of the tenth-order autoregressive filter) of the noisy speech. Furthermore, let $A_s^{-1}(e^{j\omega})$ be the tenth-order AR spectral envelope of the clean speech. The standard log spectral distortion [9] then becomes

$$SD^2 = \frac{1}{2\pi}\int \left(10\log_{10}\frac{|H(e^{j\omega})|^2|A_y^{-1}(e^{j\omega})|^2}{|A_s^{-1}(e^{j\omega})|^2}\right)^2 d\omega. \quad (3)$$

The mapping from the noise statistics (the SNR and the spectral tilt) to the postfilter parameters is based on a table that links a finite set of [SNR,tilt] pairs with a finite set of [γ_1,γ_2] pairs. To limit the *domain* of the mapping to a finite set of SNR and spectral tilt points, these input variables are subjected to uniform scalar quantizers with a finite range. The uniform scalar quantizers have

a step size of 1 dB for the SNR and 0.1 for the noise spectral tilt (prediction coefficient) and had a range of $0-30$ dB and $0.0-1.0$ respectively. The step sizes were selected sufficiently small that they did not affect the quality of the postfiltered speech. Similarly, the *range* of the mapping is limited to a finite set of uniformly spaced values for γ_1 and γ_2. The step size is 0.05 for both parameters and they ranged between 0 and 1. Again their values were selected to not affect the quality of the postfiltered speech. The mapping is then a discrete mapping that consist of quantizing the values of SNR and tilt and selecting the optimal set of γ_1 and γ_2 from a table.

The mapping was obtained by a searching procedure over a database of spectral descriptions. For each pair of SNR and spectral tilt input variables that is in the domain of the mapping, we search for the pair of emphasis parameters in the specified range that results in the lowest mean spectral distortion over a database. The procedure to obtain the mapping is shown in more detail in Table 3. Due to the fast convergency of γ_1 and γ_2, the training database used was based only on ten speech sentences from the TIMIT database. The noise sources were artificially created.

1. Create a database of clean speech power spectra, calculated over 20 ms segments of clean speech.
2. Perform for all desired SNR and tilt sets:
   (a) Add artificial noise power spectrum P_n of specified tilt to the clean power spectra P_s at the specified SNR.
   (b) For the entire database, select from all allowed parameter pairs γ_1 and γ_2 the pair that minimizes the mean SD.
   (c) Save the current input SNR level, spectral tilt and the corresponding parameters γ_1 and γ_2 in the lookup table.

Table 3. The algorithm to find the mapping. The algorithm operates on the power-spectral representation.

Figure 3 shows the mapping to the emphasis parameters as a function of SNR at two different tilt values. The smooth evolution of the filter parameters with changing noise energy ensures stable performance under errors in the estimated noise parameters. It can be seen that the level of suppression depends on the tilt of the spectrum. More attenuation is performed for the noise sources with a flat spectrum. This is natural since the structure of the postfilter leads to a similar noise suppression across the entire spectrum. For practical noise sources, the performance of the noise-dependent postfilter generally does not degrade for strongly colored spectra since their their energy is generally concentrated in a less audible regions.

In general, it is not beneficial to enhance the spectral structure of the background noise signal. If the SNR for the current frame is estimated to be below 5 dB, the frame is classified as nonspeech. In that case the postfilter is not applied and only energy attenuation of 50% is performed. The suppression of the noise level in between speech segments has significant impact on the overall performance of the system.

Given the optimal γ_1 and γ_2 we determined the mapping from the noise statistics, to obtain the gain factor G than minimizes the SNR. That can be beneficial if the noise suppression system is

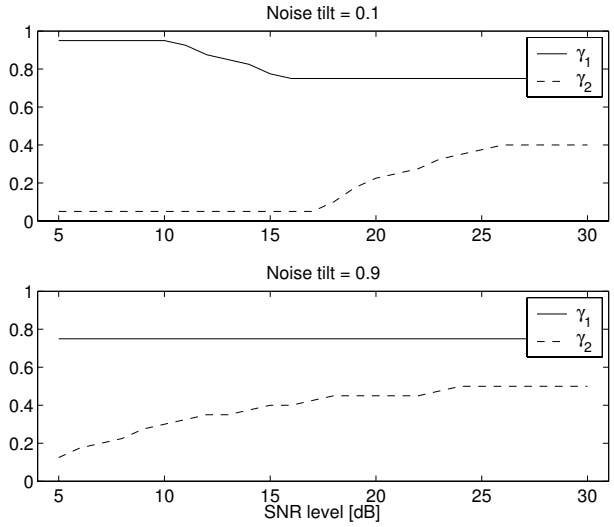

Fig. 3. Evolution of γ_1 and γ_2 with SNR.

used for parameter estimation, for example in a pre-processor to a speech recognition system. However, we found that for human perception is preferable to preserve some residual noise level and add additional constraints to smooth the level of attenuated noise from frame to frame. For the experiments, described in section 3 we retained the estimation of G, used in postfilters.

In preliminary experiments we also allowed the tilt control parameter μ to vary. However, we achieved the best perceptual performance when only the emphasis parameters were varied and when the μ fixed to the value 0.4. The reason is that when the noise spectrum tilt changes rapidly, varying μ may cause the unpleasant perceptual effect of nonconstant level of the residual noise.

3. PERFORMANCE

An A/B listening test has been carried out to evaluate the performance of the proposed system. Two male and two female speakers were arbitrarily chosen from the TIMIT database. The noisy signals were created with four real noise sources: car, rain, street and wind noise added to the speech at 15, 20, 15 and 10 dB input SNR. For the tests we used eight experienced listeners not familiar with the system. The noise-dependent postfilter used in the tests use the parameters γ_1 and γ_2 obtained by the algorithm described in Table 3. The noise parameters were estimated with the noise estimation algorithm described in Table 1.

We first tested the noise-dependent postfilter against the noise suppression system included in the IS-127 TIA/EIA standard for the Enhanced Variable Rate Codec (EVRC) [10]. Both systems were extracted from the codecs and applied directly to the noisy signal. The outputs of the systems were compared without further processing. From Table 4 can be seen that the systems perform essentially well, despite of the simplicity of the noise-dependent postfilter. We also observed that this is valid for all noise types and SNR values used in the test.

A comparison was performed between the ETSI Enhanced Full Rate (EFR) codec [11] with its standard postfilter and the EFR codec with the noise-dependent postfilter. The averaged re-

System type	Preference
Noise-dependent postfilter	56%
Noise suppression from the EVRC	44%

Table 4. System preference averaged over all speakers and noise types.

sults are presented in Table 5. As expected, the conventional postfilter performs poorly in the presence of the acoustic background noise. The difference becomes more significant with increasing the noise level as presented in Table 6. The results indicate that existing codec structures are sufficiently powerful to attenuate the background noise with a performance similar to that of separate noise-suppression algorithms.

System type	Preference
EFR + noise-dependent postfilter	75%
EFR + conventional postfilter	25%

Table 5. System preference averaged over all speakers and noise types.

Noise Type	System type	Preference
Rain 20 dB	EFR + noise-dependent postfilter	69%
	EFR + conventional postfilter	31%
Wind 10 dB	EFR + noise-dependent postfilter	88%
	EFR + conventional postfilter	12%

Table 6. Preference for different noise types.

The evaluations were also performed in terms of objective measures. The output of the EFR codec with and without noise-dependent postfilter was processed with PESQ [12]. The PESQ technology was not certified for this application, since the noise-dependent postfilter approximates the behavior of a noise suppression system. However, from listening to the test samples we concluded that the PESQ values closely follow the perceived quality.

Speech from five male and five female speakers from the TIMIT database was used for the evaluations. Ten kinds of background noise signals were added to the clean speech sentences at 5 to 25 dB SNR. The MOS scores produced by the PESQ measure are presented in Table 7. From the test results it is clear that the EFR codec with the proposed noise-dependent postfilter performs better than EFR with the conventional postfilter. The results may improve further if the mapping is optimized for the output of the codec tested.

Noise Type	EFR, no postfilter	EFR + postfilter	EFR + noise-dependent postfilter
White 15 dB	2.549	2.530	2.764
Street 20 dB	3.105	3.194	3.327
Wind 15 dB	3.192	3.127	3.328

Table 7. MOS values produced by PESQ.

4. CONCLUSIONS

Our results show that only small changes are needed in existing standard codecs to enhance significantly the quality of coded speech in acoustic background noise. The improvement can be attained by adapting the parameters of the postfilter that determine the degree of spectral emphasis and the signal gain. The proposed enhancement method is robust to a mismatch in the estimated noise characteristics and requires insignificant additional computational complexity. The new method does not affect the compatibility of the codec with existing standards. It is likely that our method can be improved further by replacing the currently used spectral distortion measures with psychoacoustically motivated measures.

5. REFERENCES

[1] V. Ramamoorthy, N. Jayant, R. Cox, and M. Sondhi, "Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE J. on Select. Areas Commun.*, vol. 6, pp. 364–382, Feb 1988.

[2] P. Kabal, F. Wang, D. O'Shaughnessy, and R. Ramachandran, "Adaptive postfiltering for enhancement of noisy speech in the frequency domain," in *Circuits and Systems, 1991., IEEE Int. Symp.*, pp. 312–315, Jun 1991.

[3] J. Lim, ed., *Speech Enhancement*. Englewood Cliffs, NJ: Prentice Hall, 1983.

[4] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans., Acoust., Speech, Signal Process.*, vol. 28, pp. 137 –145, Apr 1980.

[5] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 59–71, 1995.

[6] R. Conway, T. Sreenivas, and R. Niederjohn, "Adaptive postfiltering applied to speech in noise," in *Circuits and Systems, 1989., Proc. of the 32nd Midwest Symp.*, pp. 101–104, Aug 1989.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul 2001.

[8] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1*, 1990.

[9] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier Science Publishers, 1995.

[10] "TIA/EIA/IS-127, Enhanced Variable Rate Codec, speech service option 3 for wideband spread spectrum digital systems," Jan 1997.

[11] European Telecommun. Standards Institute (ETSI), "Enhanced Full Rate (EFR) speech transcoding (GSM 06.60)," 1996.

[12] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," *Acoust., Speech, Signal Process., IEEE Int. Conf. (ICASSP)*, vol. 2, pp. 749–752, 2001.