Speech Enhancement based on Residual Noise Shaping

Jong Won Shin, Seung Yeol Lee, Hwan Sik Yun and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul, Korea
{jwshin, sylee, hsyun}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we propose a novel approach to speech enhancement, which incorporates a new criterion based on residual noise shaping. In the proposed approach, our goal is to make the residual noise perceptually comfortable although the power of the residual noise is relatively high. In contrast to the conventional techniques, the proposed approach regulates not only the power of the signal distortion and residual noise but also the spectral shape of the residual noise. A predetermined 'comfort noise' is provided as a target for the spectral shaping. Three different versions of enhancement algorithms adopting the proposed criterion are presented. Subjective listening test results show that the proposed algorithm outperforms the conventional spectral enhancement techniques which are based on soft decision and the noise suppression module implemented in IS-893 Selectable Mode Vocoder.
Index Terms: speech enhancement, residual noise, comfort noise.

1. Introduction

The quality of a spoken language processed by speech coders and the performance of automatic speech recognition systems degrade seriously under the presence of additive background noise. Speech enhancement technique, which estimates the clean speech robustly when only the noisy signal is available, has become an indispensable part of practical speech processing systems.

One of the predominant approaches to speech enhancement is the spectral subtraction algorithm, which subtracts estimated spectral components of the noise from the input spectrum. Recently, the spectral subtraction technique has been generalized to include all the approaches that apply some spectral gain to the noisy speech spectra. A variety of criteria are used for the spectral subtraction task such as the minimum mean square error (MMSE) (which leads to the Wiener filter) [1], MMSE spectral amplitude [2], and MMSE log spectral amplitude criteria [3]. The performance of these speech enhancement algorithms has been enhanced with the incorporation of soft decision scheme [4].

More recently, there has appeared a new class of speech enhancement algorithms called the signal subspace technique [5]. Signal subspace techniques have shown a fairly good performance by making a compromise between the signal distortion and residual noise. Though effective in enhancing speech quality, these algorithms require a heavy computational burden. Some efforts have been made to reduce the computational amount [6]. Furthermore, several variations of signal subspace approach have been developed in order to keep the residual noise inaudible by constraining its level to be placed below the masking threshold of hearing [7]-[9].

It is known that there exist some noises that are heard more comfortable and pleasant in terms of human auditory perception. Many standard codecs generate perceptually comfortable noises to fill in the silence intervals for which only a small number of bits are allocated [10]-[13]. Some noises are felt more comfortable for human ears than others even if their power is higher. For that reason, it would be beneficial to control not only the average power but also the shape of the residual noise spectrum for a successful speech enhancement.

In this paper, we propose a novel criterion for speech enhancement which shapes the residual noise such that it can be heard more comfortable. The proposed approach regulates not only the power of speech distortion and residual noise, but also the spectral shape of the residual noise. Three different versions of the algorithm are presented based on the proposed criterion. A simplified implementation of the algorithm is found to have very little computational complexity and shows better performance compared with the algorithm proposed in [4] and that employed in IS-893 Selectable Mode Vocoder (SMV) [11].

2. Residual noise shaping in Speech Enhancement

2.1. General linear estimator case

Let $\mathbf{x}$ and $\mathbf{n}$ be K-dimensional time domain sample vectors of the clean speech and noise, respectively. If the noise is assumed to be additive, the observed noisy speech vector $\mathbf{y}$ can be described as

$$\mathbf{y} = \mathbf{x} + \mathbf{n}. \tag{1}$$

Let $\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}$ be a linear estimator of $\mathbf{x}$ with $\mathbf{H}$ being a $K \times K$ matrix. Then, the error vector $\mathbf{e}$ is given by

$$\begin{aligned} \mathbf{e} &= \hat{\mathbf{x}} - \mathbf{x} \\ &= (\mathbf{H} - \mathbf{I})\mathbf{x} + \mathbf{H}\mathbf{n} \\ &\triangleq \epsilon_{\mathbf{x}} + \epsilon_{\mathbf{n}} \end{aligned} \tag{2}$$

where $\epsilon_{\mathbf{x}} \triangleq (\mathbf{H} - \mathbf{I})\mathbf{x}$ represents the signal distortion and $\epsilon_{\mathbf{n}} \triangleq \mathbf{H}\mathbf{n}$ is the residual noise [5]. Let $\overline{a}$ denote an ensemble average of the random variable a. Then, a typical criterion used in signal subspace speech enhancement approach is given as follows:

$$\begin{aligned} &\mathbf{H}_{opt} = \arg\min_{\mathbf{H}} \overline{\epsilon_{\mathbf{x}}^2} \\ &\text{subject to :} \quad \frac{1}{K}\overline{\epsilon_{\mathbf{n}}^2} \le \alpha\sigma^2 \end{aligned} \tag{3}$$

where $0 \le \alpha \le 1$ and σ^2 is the variance of noise signal if the noise is assumed to be white [5], or some positive constant if the noise is

assumed to be colored [6], [14]. According to this criterion, the enhancement algorithm tries to minimize the signal distortion while maintaining the residual noise level below a certain threshold. It is noted that (3) concentrates on the power of the residual noise regardless of its spectral shape. In order to alleviate this limitation, a number of approaches that utilize the masking property of human auditory system have been developed [7]-[9]. These approaches determine the threshold based on the masking property resulting in a sort of residual noise shaping. The major purpose of the noise shaping in these techniques is to make the residual noise inaudible and they still focusses on the relative power of the residual noise with respect to the speech power spectra.

In contrast to the masking-based techniques, our approach tries not only to keep the residual noise power small but also to make it more comfortable for human hearing. The difference between the proposed criterion and that adopting masking properties lies on whether the purpose of residual noise shaping is to make the residual sound perceptually comfortable or to make it inaudible.

Given the K-dimensional desired comfortable noise vector $\mathbf{s}$, the proposed criterion modifies (3) such that

$$\begin{aligned} \mathbf{H}_{opt} &= \arg\min_{\mathbf{H}} \overline{\epsilon_{\mathbf{x}}^2} \\ \text{subject to :} \quad & \overline{\epsilon_{\mathbf{n}}^2} \le \alpha \\ \text{and} \quad & \mathrm{E}[||\mathbf{Hn} - g\mathbf{s}||^2] \le \beta \end{aligned} \tag{4}$$

where $\mathrm{E}[\cdot]$ means the expectation of the enclosed random variable and the additional second constraint attempts to make the spectral power distribution of the residual noise match that of the desired target, $\mathbf{s}$, more closely. In (4), the gain g is chosen to minimize the distance between $\mathbf{Hn}$ and $g\mathbf{s}$, i.e.,

$$g = \arg\min_{k} ||\mathbf{Hn} - k\mathbf{s}||^2 = \frac{(\mathbf{s}^T\mathbf{Hn})}{\mathbf{s}^T\mathbf{s}} \tag{5}$$

with T denoting matrix transposition. Note that the second constraint in (4) is described in terms of the Euclidean distance between the time domain residual noise and the scaled desired signal vectors, which is equivalent to the spectral distance according to the Parseval's theorem. The desired comfortable noise vector $\mathbf{s}$ is treated as a fixed deterministic vector which can vary from frame to frame.

How to select the desired comfortable noise is an important issue. Most of the standard coders design the comfortable noise as a white signal processed by the linear prediction (LP) synthesis filter [10]-[12], or a synthesized signal using randomly generated pitch lags, adaptive codebook gains and fixed codebook excitations [13]. It is generally known that some noises such as the vehicular noise are more comfortable for human ears than other types of noises with the same signal-to-noise ratio (SNR). We can use the output of an LP synthesis filter with an appropriate excitation as a desired signal as in standard codecs, or can use a set of perceptually pleasant noise vectors instead. In principle, any signal which is comfortable for human ears can be the target comfortable noise, $\mathbf{s}$.

The problem stated in (4) is a convex optimization task, and we can construct an objective function using the method of Lagrange multipliers as follows:

$$J = \overline{\epsilon_{\mathbf{x}}^2} + \mu_1(\overline{\epsilon_{\mathbf{n}}^2} - \alpha) + \mu_2(\mathrm{E}[||\mathbf{Hn} - g\mathbf{s}||^2] - \beta) \tag{6}$$

where $\mu_1, \mu_2 \ge 0$ are the Lagrange multipliers. From (2) and (5), the objective function J in (6) becomes

$$\begin{aligned} J = \mathrm{tr}\,(\mathbf{H}-\mathbf{I})R_{\mathbf{x}}(\mathbf{H}-\mathbf{I})^T + \mu_1(\mathrm{tr}\,\mathbf{H}R_{\mathbf{n}}\mathbf{H}^T - \alpha) \\ +\mu_2(\mathrm{tr}\,\mathbf{H}R_{\mathbf{n}}\mathbf{H}^T - \frac{\mathbf{s}^T\mathbf{H}R_{\mathbf{n}}\mathbf{H}^T\mathbf{s}}{\mathbf{s}^T\mathbf{s}} - \beta) \end{aligned} \tag{7}$$

in which $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$ are the autocorrelation matrices of $\mathbf{x}$ and $\mathbf{n}$, respectively. Differentiating this objective function with respect to $\mathbf{H}$ and setting it to zero leads us to

$$\mathbf{H}\{R_{\mathbf{x}} + (\mu_1+\mu_2)R_{\mathbf{n}}\} - R_{\mathbf{x}} - \frac{\mu_2}{\mathbf{s}^T\mathbf{s}}\mathbf{s}\mathbf{s}^T\mathbf{H}R_{\mathbf{n}} = 0 \tag{8}$$

where the last term makes the algorithm different from the conventional signal subspace approach. It is not easy to derive a closed form solution of this equation, but it can be solved since it poses a set of $K \times K$ linear equations in $K \times K$ unknown variables.

2.2. Spectral subtraction case

For a simpler and computationally less expensive implementation, we can consider the spectral subtraction approach where $\mathbf{H}$ is constrained to have the form $\mathbf{H} = \mathbf{F}^H\mathbf{G}\mathbf{F}$ in which $\mathbf{F}$ denotes the discrete Fourier transform (DFT) matrix and $\mathbf{G} = \mathrm{diag}[g_1, g_2, \cdots, g_K]$. With this structured form of estimation, the objective function (7) becomes

$$\begin{aligned} J = \quad & \mathrm{tr}\,(\mathbf{F}^H\mathbf{G}\mathbf{F} - \mathbf{F}^H\mathbf{F})R_{\mathbf{x}}(\mathbf{F}^H\mathbf{G}\mathbf{F} - \mathbf{F}^H\mathbf{F}) \\ & +(\mu_1+\mu_2)\mathrm{tr}\,\mathbf{F}^H\mathbf{G}\mathbf{F}R_{\mathbf{n}}\mathbf{F}^H\mathbf{G}\mathbf{F} \\ & -\mu_2\frac{\mathbf{s}^T\mathbf{F}^H\mathbf{G}\mathbf{F}R_{\mathbf{n}}\mathbf{F}^H\mathbf{G}\mathbf{F}\mathbf{s}}{\mathbf{s}^T\mathbf{s}} - \mu_1\alpha - \mu_2\beta. \end{aligned} \tag{9}$$

Let $R_{\mathbf{X}}$ and $R_{\mathbf{N}}$ be defined as follows:

$$\begin{aligned} R_{\mathbf{X}} &= \mathrm{E}[\mathbf{F}\mathbf{x}\mathbf{x}^H\mathbf{F}^H] = \mathbf{F}R_{\mathbf{x}}\mathbf{F}^H \\ R_{\mathbf{N}} &= \mathrm{E}[\mathbf{F}\mathbf{n}\mathbf{n}^H\mathbf{F}^H] = \mathbf{F}R_{\mathbf{n}}\mathbf{F}^H \end{aligned} \tag{10}$$

and let $\mathbf{S}$ denote the DFT vector of $\mathbf{s}$, i.e., $\mathbf{S} = \mathbf{F}\mathbf{s}$. Then, using the properties $\mathbf{F}^H\mathbf{F} = \mathbf{I}$ and $\mathrm{tr}\,(\mathbf{ABC}) = \mathrm{tr}\,(\mathbf{CAB})$, the objective function J can be reduced into

$$\begin{aligned} J = \mathrm{tr}(\mathbf{G}-\mathbf{I})^2R_{\mathbf{X}} + (\mu_1+\mu_2)\mathrm{tr}\mathbf{G}^2R_{\mathbf{N}} \\ -\mu_2\frac{\mathbf{S}^H\mathbf{G}R_{\mathbf{N}}\mathbf{G}\mathbf{S}}{\mathbf{S}^H\mathbf{S}} - \mu_1\alpha - \mu_2\beta, \end{aligned} \tag{11}$$

or

$$\begin{aligned} J = \sum_j (g_j-1)^2R_{\mathbf{X}jj} + (\mu_1+\mu_2)\sum_j g_j^2R_{\mathbf{N}jj} \\ -\mu_2\frac{\sum_{jk}\mathbf{S}_j^*g_jR_{\mathbf{N}jk}g_k\mathbf{S}_k}{\sum_j|\mathbf{S}_j|^2} - \mu_1\alpha - \mu_2\beta \end{aligned} \tag{12}$$

where $\mathbf{A}_{ij}$ represents the (i, j) th element of a matrix $\mathbf{A}$ and $\mathbf{a}_i$ is the i th element of a column vector $\mathbf{a}$. Like in the general linear estimator case, we can obtain a set of equations that $\mathbf{G}$ must satisfy by differentiating J with respect to g_i's for $1 \le i \le K$ and setting them to zero. After some algebra, we have

$$\begin{aligned} \{R_{\mathbf{X}ii} + (\mu_1+\mu_2)R_{\mathbf{N}ii}\}g_i - R_{\mathbf{X}ii} \\ -\mu_2\sum_j R_{\mathbf{N}ij}\frac{\mathrm{Re}\,[\mathbf{S}_j^*\mathbf{S}_i]}{\sum_k|\mathbf{S}_k|^2}g_j = 0, \quad 1 \le i \le K \end{aligned} \tag{13}$$

where the last term makes the solution different from the conventional Wiener filter.

2.3. Modified Wiener filtering case with a further assumption

Further assumption can be adopted to make the computational complexity very low. If we assume that the frequency components of the noise signal are statistically uncorrelated, i.e., $R_{\mathbf{N}ij} = 0 \quad \text{for } i \neq j$, (13) further reduces to

$$\Big[R_{\mathbf{X}ii} + \big\{\mu_1 + \mu_2(1 - \frac{|\mathbf{S}_i|^2}{\sum_j |\mathbf{S}_j|^2})\big\}R_{\mathbf{N}ii}\Big]g_i = R_{\mathbf{X}ii}, \quad 1 \leq i \leq K. \quad (14)$$

If we let the SNR for the i-th frequency bin be denoted by $\xi_i = R_{\mathbf{X}ii}/R_{\mathbf{N}ii}$, the optimal g_i can be written in a closed form as follows:

$$g_i = \frac{\xi_i}{\xi_i + \mu_1 + \mu_2(1 - \frac{|\mathbf{S}_i|^2}{\sum_j |\mathbf{S}_j|^2})}, \quad 1 \leq i \leq K. \quad (15)$$

It is not difficult to see that the spectral gain increases as the power of the corresponding spectral component of the normalized desired comfortable noise gets larger and vice versa.

The Lagrange multipliers μ_1 and μ_2 should be computed from the thresholds α and β, which turns out to be rather complicated. For that reason, we choose to control μ_1 and μ_2 directly. In other words, instead of the constrained minimization problem stated in (4), we construct a simple minimization problem where the objective function is just a weighted summation of three subfunctions to be minimized such that

$$J = \overline{\epsilon_{\mathbf{x}}^2} + \mu_1\overline{\epsilon_{\mathbf{n}}^2} + \mu_2 \mathrm{E}[||\mathbf{F}^H\mathbf{G}\mathbf{F}\mathbf{n} - g\mathbf{s}||^2]. \quad (16)$$

Now, μ_1 and μ_2 are the weights that control the trade-off among the signal distortion, the power of the residual noise, and the spectral shape of the residual noise. Minimizing J once again leads us to the solution given in (15).

The assumption of statistical independence among noise spectral components is quite common even though it may deviate from the real measurement [2]-[4]. However, it is generally accepted that the correlation among different frequency components is weaker in the background noise than in voiced speech signals, which have a harmonic structure. Moreover, although the assumption is not quite accurate, the resulted gain function (15) has some intuitive meanings. It has a form similar to that of the Wiener filter except for the fact that there are two parameters: μ_1 adjusts the trade off between the signal distortion and the residual noise, and μ_2 controls the perceptual comfortableness of the residual noise. In this respect, the proposed algorithm can be considered as an extended version of the conventional Wiener filter, which explicitly incorporates the capability of residual noise shaping.

3. Experimental results

The simplified spectral gain function shown in (15) was applied to the speech enhancement task. Processing of data and the SNR estimation for each frequency bin were carried out in exactly the same way as those used in [4]. The distributions for DFT coefficients of noise and noisy speech signal were modeled in terms of the complex Gaussian distributions and the global speech absence probability (SAP) was computed based on these parametric models. Estimates for the relevant parameters such as variances of the clean speech and background noise components were updated based on a soft decision scheme. Interested readers are referred to [4]. In this experiment, what is differentiated from [4] is that we applied (15) for gain computation instead of the Ephraim-Malah's rule. Except for the gain computation step, all the other parts of speech enhancement were maintained the same to those employed in [4]. In this respect, the proposed approach can be considered as a post-processing unit that is concatenated to a speech enhancement system and modifies the noise-suppressed output so that it can be heard more comfortable in human perception.

As a target comfort noise, we tried two signals. One was extracted from the vehicular noise samples in NOISEX-92 database, and the other was generated by passing a white random noise through the 10th-order LP synthesis filter where the LP parameters were estimated from the input speech. The vehicular noise was selected since it is known perceptually comfortable even in low SNR conditions. On the other hand, the idea of exciting the LP synthesis filter with a white noise comes from many standard speech coders where only a rough information of the spectral envelope is transmitted during non-active speech periods [10]-[12]. The use of the spectral envelope generated by the LP synthesis filter can also somewhat account for the masking effect though the residual noise power is not strictly restricted below the masking threshold. From a number of experiments, we found that both of the two target comfort noises resulted in a comparable performance. All the experimental results shown in the remainder of this paper were obtained when we applied the LP synthesis filter to generate the desired comfort noise. We could observe a similar performance with the vehicular noise.

In order to evaluate the performance of the simplified version of the proposed enhancement algorithm, an informal subjective quality evaluation was performed. The test material consisted of eight 7.5-s-long speech files spoken by 4 male and 4 female speakers. Each file contained two sentences and was sampled at 8 kHz. These files were corrupted by three different kinds of additive noises at various SNR's. The added noises were extracted from the white, pink and F-16 cockpit noise waveforms collected in NOISEX-92 database. The frame size was 10 ms and 128-point fast Fourier transform was applied to extract the spectra. For each test file, nine listeners gave their opinion of the perceptual quality with a score among 5 (excellent), 4 (good), 3 (fair), 2 (poor) and 1 (bad). All the scores from the listeners were then averaged to yield the mean-opinion-score (MOS). In this experiment, the parameters that control the contribution of each distortion in (16) were set to $\mu_1 = 0.25$ and $\mu_2 = 0.8$.

Table 1 summarizes the results of the MOS tests. The performance of the proposed algorithm (denoted as SERNS, which is an abbreviation of Speech Enhancement based on Residual Noise Shaping) was compared with that of the spectral enhancement method based on the global soft decision presented in [4] (denoted as SEGSD) and was also compared with that of the algorithm adopted in IS-893 SMV. The proposed algorithm outperformed the SEGSD algorithm, which showed a slightly better performance than that used in SMV, by 0.18 in MOS score on average. Even though the performance gain was not quite large, the proposed algorithm demonstrated a consistently better performance than other algorithms at every SNR with any background noise type tested in the experiment. The experimental results show that even in the simplified implemental form, the proposed approach considering an appropriate residual noise shaping improves the performance of the conventional speech enhancement systems.

noise	white			pink			F-16 cockpit			Average
SNR(dB)	5	10	15	5	10	15	5	10	15	
SMV	2.89	3.61	4.04	1.53	2.33	3.28	1.76	2.69	3.50	2.85
SEGSD	2.97	3.64	4.04	1.60	2.42	3.42	1.79	2.69	3.53	2.90
SERNS	3.26	3.69	4.06	1.76	2.83	3.81	1.85	2.71	3.71	3.08

Table 1: MOS results for the proposed enhancement algorithm (SERNS), the one presented in [4] (SEGSD) and the one adopted in IS-893 Selectable Mode Vocoder (SMV) under various environmental conditions.

4. Conclusions

We have proposed a novel criterion of speech enhancement, which shapes the residual noise so that it can be heard more comfortable. Three different versions of the enhancement algorithms based on the proposed criterion are presented. The simplest one has a form similar to that of the Wiener filter and so it can be implemented with a very little computational complexity. The performance of the simplest version of proposed algorithm was shown to be superior to that of the conventional spectral subtraction algorithm based on global SAP and that of the algorithm used in IS-893 SMV. Further study areas may include the computationally efficient implementation of the more complicated versions of the proposed enhancement algorithm.

5. Acknowledgements

This work was partly supported by IT R&D Project funded by Korean Ministry of Information and Communications.

6. References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. New Jersey: Prentice Hall, 1993.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-33, no.2, pp. 443-445, Apr. 1985.

[4] J. -H. Chang and N. S. Kim, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251-266, Jul. 1995.

[6] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 2, pp. 87-95, Feb. 2001.

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

[8] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 700-708, Nov. 2003.

[9] Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 270-273, Feb. 2004.

[10] 3GPP2 Document C.S0014-0 v1.0, *Enhanced Variable Rate Codec (EVRC)*, Dec. 1999.

[11] 3GPP2 Document C.S0030-0 v3.0, *Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communication Systems*, Jan. 2004.

[12] 3GPP2 Document C.R0052-A v1.0, *Software Distribution for Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Service Options 62 and 63, Specification*, Aug. 2005.

[13] ITU-T, "A silence compression scheme for G.729 optimised for terminals conforming to ITU-T V.70," *ITU-T Rec. G.729 Annex B*, Nov. 1996.

[14] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 104-106, Apr. 2003.