LETTER

Speech Enhancement Based on Data-Driven Residual Gain Estimation

Yu Gwang JIN[†a)], ***Nonmember***, **Nam Soo KIM**[†b)], ***and*** **Joon-Hyuk CHANG**[††c)], ***Members***

SUMMARY In this letter, we propose a novel speech enhancement algorithm based on data-driven residual gain estimation. The entire system consists of two stages. At the first stage, a conventional speech enhancement algorithm enhances the input signal while estimating several signal-to-noise ratio (SNR)-related parameters. The residual gain, which is estimated by a data-driven method, is applied to further enhance the signal at the second stage. A number of experimental results show that the proposed speech enhancement algorithm outperforms the conventional speech enhancement technique based on soft decision and the data-driven approach using SNR grid look-up table.

key words: *speech enhancement, noise reduction, data-driven approach, residual gain estimation*

1. Introduction

The quality of a speech signal may significantly deteriorate when additive noise is present in the background. The speech enhancement technique, which estimates the clean speech when only the noisy signals are available, is the most popular approach to reduce the effect of noises. In the past few decades, many approaches to speech enhancement have been proposed. Among them, the minimum mean-square error (MMSE) estimators of the spectral amplitudes [1] or log spectral amplitudes [2] are widely used for noise reduction. The performance of those speech enhancement algorithms has been further improved with the incorporation of the soft decision scheme [3], [4].

In recent years, several speech enhancement algorithms based on the data-driven techniques have been proposed [5], [6]. In these approaches, a proper gain is obtained from a look-up table which is trained off-line in a variety of noisy conditions. The index for the gain is usually found by using signal features provided by a conventional speech enhancement algorithm. More recently, several approaches, instead of deriving the gain directly, estimate the noise power [7] and the *a priori* signal-to-noise ratio (SNR) [8] in a data-driven manner.

In this letter, we propose a novel speech enhancement algorithm based on data-driven residual gain estimation which consists of two stages: a conventional speech enhancement module and the gain adjustment module. At the first stage, the noisy input signal is processed by a conventional enhancement algorithm from which both the enhanced signal and several SNR-related parameters are obtained. At the second stage, the gain obtained from the first stage is further adjusted by incorporating the residual gain, which means the difference between the optimal gain and the one extracted from the first stage's enhancement algorithm. A data-driven strategy is employed to estimate the residual gain, for which we use the SNR-related parameters provided by the first stage enhancement algorithm. From a number of experiments, we can see that the proposed residual gain estimation approach shows better performance compared with the soft decision based enhancement algorithm [3] which is used as the first stage module in this work and the data-driven approach proposed in [5].

Manuscript received May 2, 2011.
Manuscript revised August 12, 2011.

[†]The authors are with the School of Electrical Engineering and the INMC, Seoul National University, Seoul 151–744, Korea.
[††]The author is with the School of Electronic Engineering, Hanyang University, Seoul 133–791, Korea.
a) E-mail: ygjin@hi.snu.ac.kr
b) E-mail: nkim@snu.ac.kr
c) E-mail: jchang@hanyang.ac.kr

DOI: 10.1587/transinf.E94.D.2537

2. Conventional Approaches to Speech Enhancement

Let $Y_k(l)$, $X_k(l)$ and $D_k(l)$ denote the discrete Fourier transform (DFT) coefficients of the noisy speech, clean speech and noise signal, respectively, for the k-th frequency bin at frame l. Then the spectral component of the noisy speech $Y_k(l)$ is given by

$$Y_k(l) = X_k(l) + D_k(l). \tag{1}$$

In the spectral subtraction techniques, the spectral component of the noisy speech can be enhanced by

$$\hat{X}_k(l) \triangleq \hat{G}(k,l) \cdot Y_k(l) \tag{2}$$

where $\hat{X}_k(l)$ and $\hat{G}(k,l)$ represent the spectral component of the enhanced speech and the corresponding spectral gain, respectively.

A number of methods to calculate a suitable gain $\hat{G}(k,l)$ have been developed. Among them, the MMSE estimators of the spectral amplitudes [1] or log spectral amplitudes [2] are well-known. In these algorithms, the gain $\hat{G}(k,l)$ is derived as $\hat{G}(k,l) = F[\hat{\xi}(k,l), \hat{\gamma}(k,l)]$ where $F[\cdot]$ is a complicated non-linear function and $\hat{\xi}(k,l)$ and $\hat{\gamma}(k,l)$ are estimates for the *a priori* SNR and the *a posteriori* SNR, respectively. The performances of these algorithms have been further improved with the incorporation of a soft decision scheme [3], [4] such that $\hat{G}(k,l) = F[\hat{\xi}(k,l), \hat{\gamma}(k,l), \hat{p}(k,l)]$ where $\hat{p}(k,l)$ indicates the speech absence probability (SAP) computed for the k-th frequency bin at frame l. Additionally, the noise power spectrum estimate can be updated not only during the

periods of speech absence but also when there exists an active speech component by taking the SAP into account [3].

Recently, several data-driven approaches to optimize speech enhancement methods have been proposed [5], [6]. In these approaches, the gain is obtained from an SNR look-up table trained in a variety of noisy conditions. Within the framework of table look-up, the gain is found such that

$$\hat{G}(k,l) = Table[\lfloor\hat{\xi}(k,l)\rfloor, \lfloor\hat{\gamma}(k,l)\rfloor] \tag{3}$$

where $Table[\alpha,\beta]$ denotes the gain corresponding to the (α,β) component of the look-up table and $\lfloor\cdot\rfloor$ indicates the operation of matching to the nearest table grid index. In these approaches, the conventional speech enhancement system works only as an estimator for the *a priori* and the *a posteriori* SNR's which become the input feature for the table look-up.

Data-driven methods have been found advantageous in optimizing the speech enhancement system to a particular environment with enough training database collected in the target environment. Since, however, the performance totally depends on the trained table, it may deteriorate if there exist some mismatches between the training and test environments.

3. Data-Driven Residual Gain Estimation

In this section, we propose a two-stage speech enhancement algorithm consisting of the conventional speech enhancement module and the gain adjustment module. At the first stage, the noisy input signal is enhanced and several SNR-related feature parameters are estimated. Unlike previous data-driven methods, the first stage module works for not only estimating parameters but also enhancing the input signal. In the subsequent processing, the speech signal is further enhanced by applying the residual gain. Estimation of the residual gain is performed based on a data-driven approach in which the codebooks are trained on a collected training database.

In a conventional speech enhancement algorithm, the gain $\hat{G}(k,l)$ is applied to the spectral component of an input signal as given by (2). Let $G_{opt}(k,l)$ denote the optimal gain for the k-th frequency bin at frame l, which is given by $G_{opt}(k,l) \triangleq X_k(l)/Y_k(l)$. Then,

$$\begin{aligned} X_k(l) &= G_{opt}(k,l)\cdot Y_k(l) \\ &= (G_{res}(k,l)\cdot\hat{G}(k,l))\cdot Y_k(l) \\ &= G_{res}(k,l)\cdot\hat{X}_k(l) \end{aligned} \tag{4}$$

where $G_{res}(k,l)$ is a residual gain which represents the difference between the optimal gain and the gain derived by the conventional speech enhancement algorithm.

As seen from (4), the optimal output signal can be obtained by applying the residual gain $G_{res}(k,l)$ to the enhanced signal $\hat{X}_k(l)$, so a successful estimation of the residual gain will lead to an improved speech enhancement performance. In this work, we estimate the log residual gain

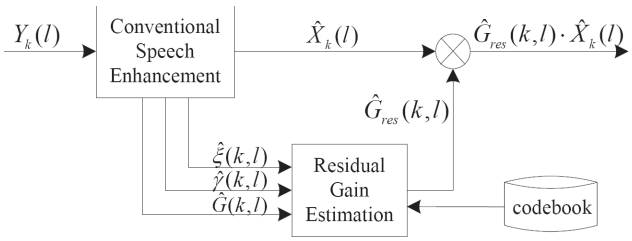

Fig. 1 Block diagram of the proposed two-stage speech enhancement system.

defined by

$$\begin{aligned} H(k,l) &\triangleq \log G_{res}(k,l) \\ &= \log G_{opt}(k,l) - \log\hat{G}(k,l) \end{aligned} \tag{5}$$

to minimize $E\{\|X_k(l) - \exp(\hat{H}(k,l))\cdot\hat{X}_k(l)\|^2\}$ where $\hat{H}(k,l)$ is an estimate for $H(k,l)$. In order to estimate the log residual gain, we employ a data-driven approach using the SNR-related parameters which are obtained from the first stage speech enhancement module. The block diagram of the proposed speech enhancement algorithm is shown in Fig. 1. As for the first stage module, we apply the speech enhancement algorithm proposed in [3] to compute $\hat{\xi}(k,l)$, $\hat{\gamma}(k,l)$ and $\hat{G}(k,l)$.

It is generally known that the speech spectra possess high level of spectral and temporal correlations. To take advantage of the temporal and spectral correlations inherent in the speech signals, we perform a grouping of the SNR-related parameters and gains both in the frequency and time domains. For a frequency-time grid point (k,l), we incorporate the frequency components with frequency bin index from $k-M$ to $k+M$ and the temporal components with frame index from $l-N$ to $l+N$ except for the grid point itself. Let $\Xi_{k,l}$, $\Gamma_{k,l}$, and $G_{k,l}$ be the grouped components defined at the frequency-time point (k,l) of the *a priori* SNR, the *a posteriori* SNR and the gain, respectively. Then they can be described by $((2M+1)(2N+1)-1)$-dimensional vectors as given by

$$\begin{aligned} \Xi_{k,l} \triangleq [&\check{\xi}_{k,l}(-M,-N), \check{\xi}_{k,l}(-M,-N+1), \cdots, \check{\xi}_{k,l}(-M,N), \\ &\check{\xi}_{k,l}(-M+1,-N), \cdots, \check{\xi}_{k,l}(0,-1), \check{\xi}_{k,l}(0,1), \cdots, \\ &\cdots, \check{\xi}_{k,l}(M,N)], \\ \Gamma_{k,l} \triangleq [&\check{\gamma}_{k,l}(-M,-N), \check{\gamma}_{k,l}(-M,-N+1), \cdots, \check{\gamma}_{k,l}(-M,N), \\ &\check{\gamma}_{k,l}(-M+1,-N), \cdots, \check{\gamma}_{k,l}(0,-1), \check{\gamma}_{k,l}(0,1), \cdots, \\ &\cdots, \check{\gamma}_{k,l}(M,N)], \\ G_{k,l} \triangleq [&\check{g}_{k,l}(-M,-N), \check{g}_{k,l}(-M,-N+1), \cdots, \check{g}_{k,l}(-M,N), \\ &\check{g}_{k,l}(-M+1,-N), \cdots, \check{g}_{k,l}(0,-1), \check{g}_{k,l}(0,1), \cdots, \\ &\cdots, \check{g}_{k,l}(M,N)], \end{aligned} \tag{6}$$

where

$$\begin{aligned} \check{\xi}_{k,l}(i,j) &\triangleq \log\hat{\xi}(k+i,l+j) - \log\hat{\xi}(k,l), \\ \check{\gamma}_{k,l}(i,j) &\triangleq \log\hat{\gamma}(k+i,l+j) - \log\hat{\gamma}(k,l), \\ \check{g}_{k,l}(i,j) &\triangleq \log\hat{G}(k+i,l+j) - \log\hat{G}(k,l), \end{aligned}$$

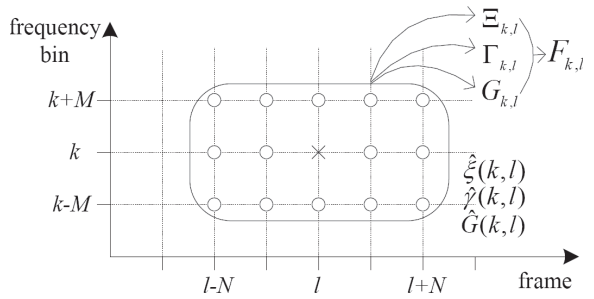

Fig. 2 Combination of feature parameters in the time-frequency domain.

$$for\ \ i = -M, \cdots, M, \quad j = -N, \cdots, N. \tag{7}$$

This grouping is illustrated in Fig. 2. In order to tabulate $\Xi_{k,l}$, $\Gamma_{k,l}$, and $G_{k,l}$ jointly, we apply the vector quantization (VQ) technique. The VQ codebook is obtained for each frequency component separately. The input to this VQ is the supervector $F_{k,l} = [\Xi_{k,l}, \Gamma_{k,l}, G_{k,l}]$ of which dimension equals $12MN + 6M + 6N$. Since the dimension of the supervector is usually high, direct application of it to the VQ codebook construction may be inefficient. In order to alleviate this problem, we apply the principal component analysis (PCA) technique as follows:

$$F'_{k,l} = P_k \cdot \{F_{k,l} - m_k\}, \tag{8}$$

where P_k is the matrix of which rows are the PCA basis vectors and m_k denotes the mean vector. Since P_k consists of reduced number of basis vectors, $F_{k,l}$ is projected to a low-dimensional representation $F'_{k,l}$ through (8). Since different frequency components have different characteristics of the spectral and temporal correlations, in this work we apply the PCA technique separately to the individual frequencies for an effective processing. Now a codebook is constructed for each frequency bin by applying a conventional VQ training algorithm to $\{F'_{k,l}\}$. In the training procedure, since it is assumed that we know the exact values of $X_k(l)$, $Y_k(l)$ and their ratio $G_{opt}(k, l)$ which are accompanied with a supervector $F'_{k,l}$, we can associate each codeword of the VQ codebook with an estimate for the residual gain. A log residual gain associated to a specific codeword is computed by averaging the log residual gains obtained from the supervectors that are assigned to that codeword.

There are several advantages in the proposed two-stage speech enhancement technique. First, the proposed residual gain estimation approach considers the error characteristics of the conventional enhancement module and tries to recover them. So a better or similar performance can be anticipated compared with the conventional method even in the worst environmental condition. It makes the whole system more robust. Second, the proposed scheme of estimating the residual gain instead of the optimal gain itself can be considered more robust because the dynamic range of the residual gain is usually smaller than that of the gain. The robustness may be improved further if the first stage enhancement algorithm produces more exact spectral gains. Third, the proposed approach considers both the spectral and temporal correlations of the speech signal for table look-up. Through the incorporation of these correlations, some mistakes in gain estimation occurring at the first stage enhancement module can be partially recovered. Finally, each element of the supervector does not represent the original SNR-related parameter or gain directly but the relative difference, as given by (7). Since the distribution of this relative difference is usually more compact and better balanced, it is highly likely to result in a good VQ codebook.

4. Experimental Results

To verify the performance of the proposed speech enhancement algorithm, we carried out a number of objective quality measurements under various noisy conditions. We compared the speech quality obtained from the proposed residual gain estimation algorithm (denoted as RGE) with those of the conventional speech enhancement algorithm based on soft decision [3] (denoted as SESD) which was adopted in this work as the first stage module, and the data-driven approach proposed in [5] using the SNR grid look-up table (denoted as SGLT). For the test material, NOIZEUS corpus [9] was used which consisted of 30 IEEE sentences, and these speech data were corrupted by different types of additive noises at various SNR's. In this work, noisy files corrupted by airport, street and train noises were used with 0, 5, 10 and 15 dB SNR. Each file was sampled at 8 kHz. In the procedure for VQ codebook training, the white noise was added to a separate speech data of length 456 seconds by varying the SNR from −10 dB to 30 dB so that the total length of data became 4104 sec. For RGE, parameter grouping as shown in (6) was performed with M=1 and N=3, hence the dimension of each supervector became 60 which was further reduced to 10 with the application of PCA. A VQ codebook with 128 codewords was trained for each frequency bin based on the 10-dimensional feature vectors. The SNR grid look-up table used in the SGLT scheme was trained for the values of the *a priori* and *a posteriori* SNR's, $\hat{\xi}(k, l)$ and $\hat{\gamma}(k, l)$, in the range [−20 dB, 40 dB] with a 1 dB step size [1].

In order to evaluate the performance of the proposed speech enhancement technique, the segmental SNR (SSNR) and the perceptual evaluation of speech quality (PESQ) [10] measurements were carried out. For a comparison among the tested algorithms, SESD, RGE and SGLT, we computed the SSNR improvement (SSNR+) score, the difference of the SSNR of the signal which was processed by each enhancement technique from that of the unprocessed noisy signal. The results are shown in Table 1 where we can see that in most of the tested cases, the proposed RGE algorithm produced better SSNR+ and PESQ scores than both the SESD used in the first stage and the SGLT algorithms.

To discuss the difference between each method in terms of the amount of noise reduction and speech distortion, we evaluated the segmental noise attenuation (segmental NA) and the segmental speech-to-speech-distortion ratio (segmental SSDR) [6]. The larger the values of the segmental NA and segmental SSDR become, the less residual noise

Table 1 Results of segmental SNR improvement (SSNR+) and perceptual evaluation of speech quality (PESQ).

		SSNR+ [dB]			PESQ [point]		
noise	SNR	SESD	RGE	SGLT	SESD	RGE	SGLT
airport	0 dB	5.70	7.14	4.88	1.78	1.84	1.84
	5 dB	4.43	5.82	4.14	2.14	2.23	2.22
	10 dB	3.62	5.00	3.56	2.53	2.57	2.54
	15 dB	3.05	4.21	3.17	2.90	2.93	2.86
street	0 dB	5.77	7.17	5.18	1.71	1.80	1.79
	5 dB	4.12	5.39	4.15	2.07	2.17	2.16
	10 dB	3.15	4.27	3.35	2.49	2.54	2.51
	15 dB	2.45	3.63	2.83	2.81	2.84	2.80
train	0 dB	6.06	7.53	5.62	1.69	1.79	1.80
	5 dB	4.91	6.27	4.76	2.01	2.14	2.13
	10 dB	4.27	5.54	4.23	2.42	2.48	2.45
	15 dB	3.42	4.55	3.56	2.80	2.85	2.80
Avg.		4.25	5.54	4.12	2.28	2.35	2.33

Table 2 Results of segmental noise attenuation (segmental NA) and segmental speech-to-speech-distortion ratio (segmental SSDR).

		segmental NA [dB]			segmental SSDR [dB]		
noise	SNR	SESD	RGE	SGLT	SESD	RGE	SGLT
airport	0 dB	15.31	19.69	10.28	3.16	3.55	6.91
	5 dB	14.26	18.36	9.64	6.01	6.90	10.07
	10 dB	12.86	16.94	9.27	10.04	11.45	13.70
	15 dB	11.47	15.68	8.80	15.28	16.50	18.14
street	0 dB	15.62	20.22	11.39	2.93	3.07	6.38
	5 dB	14.59	18.80	10.51	6.09	6.86	9.96
	10 dB	12.73	16.84	9.54	12.29	13.57	15.52
	15 dB	11.82	15.95	9.14	15.80	17.13	18.73
train	0 dB	15.71	20.25	11.70	2.64	2.83	6.31
	5 dB	14.64	18.77	10.79	5.94	7.01	10.09
	10 dB	12.95	17.07	9.97	11.09	12.83	14.69
	15 dB	11.38	15.52	8.97	17.06	18.60	19.99
Avg.		13.61	17.84	10.00	9.03	10.03	12.54

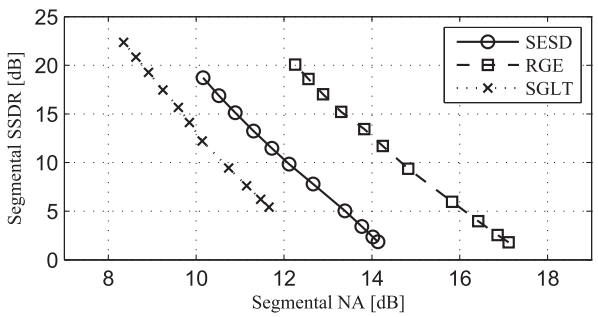

Fig. 3 Segmental SSDR vs. segmental NA tested under white noise, by applying the SESD, RGE and SGLT algorithm.

and speech distortion remain, and the better the algorithm performs. The segmental SSDR was averaged over the frames where speech was present, while the segmental NA was averaged over all frames. The results are summarized in Table 2, and the result for the white noise is plotted in Fig. 3 where the markers of each curve indicate the performances for the input SNR in the range [0 dB, 20 dB] with a 2 dB step size. Compared with the SGLT algorithm, RGE approach showed a higher level of segmental NA though slight degradation in terms of segmental SSDR. Furthermore, the residual noise of the SGLT algorithm was found relatively nonstationary because the gain estimated from each frame was less correlated with that of the previous frame.

5. Conclusions

In this letter, we have proposed a novel speech enhancement algorithm based on data-driven residual gain estimation. At the first stage, the input signal is enhanced by a conventional speech enhancement module while several SNR-related parameters are estimated simultaneously. The residual gain, which is estimated by a data-driven method, is applied to further adjust the gain obtained from the first stage. Experimental results show that the proposed algorithm performs better than both the conventional speech enhancement technique based on soft decision [3] and the data-driven approach using SNR grid look-up table [5].

Acknowledgments

This research was supported in part by Basic Science Research Program through the NRF funded by the Ministry of Education, Science and Technology (No. 2011-0020407) and by the MKE, Korea, under the ITRC support program supervised by the NIPA (NIPA-2010-C1090-1021-0007).

References

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.33, no.2, pp.443–445, April 1985.

[3] N.S. Kim and J. -H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Process. Lett., vol.7, no.5, pp.108–110, May 2000.

[4] J. -H. Chang and N.S. Kim, "Speech enhancement: new approaches to soft decision," IEICE Trans. Inf.& Syst., vol.E84-D, no.9, pp.1231–1240, Sept. 2001.

[5] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," Speech Commun., vol.49, no.7-8, pp.530–541, July-Aug. 2007.

[6] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," IEEE Trans. Audio Speech Lang. Process., vol.16, no.4, pp.825–834, May 2008.

[7] J. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," IEEE Trans. Audio Speech Lang. Process., vol.16, no.6, pp.1112–1123, Aug. 2008.

[8] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," IEEE Trans. Audio Speech Lang. Process., vol.19, no.1, pp.186–195, Jan. 2011.

[9] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp.153–156, 2006.

[10] ITU, "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 Rec. P. 862, 2000.