Available online at www.elixirpublishers.com (Elixir International Journal)

Electrical Engineering

Elixir Elec. Engg. 104 (2017) 45785-45789

ISSN: 2229-712X

Speech Enhancement Combined with Post processing Technique in Non-Stationary Noise Environments

Subbarao Genikala

Asst.prof, Department of ECE, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh-517325,India.

ARTICLE INFO

Article history:

Received: 20 January 2017;

Received in revised form:

26 February 2017;

Accepted: 8 March 2017;

Keywords

Speech enhancement,

Estimators,

Harmonic regeneration.

ABSTRACT

A two stage novel speech enhancement algorithm is presented in this paper. First stage is designed such that it can substantially improve the signal to residual spectrum ratio by combining statistical estimators of the spectral magnitude of the speech and noise. By expressing the signal to residual spectrum ratio as a function of the estimators gain function. We derive a hybrid strategy that can improve the signal to residual spectrum ratio when the apriori and the posteriori SNR are detected to the lower than 0dB. The enhanced signals still suffer from undesirable speech distortion due to harmonic distortion. Some harmonics are considered as noise only components and are suppressed. To overcome harmonic distortion introduced in enhanced speech, in the proposed method the suppressed harmonics are regenerated. Objective and subjective tests were carried out to demonstrate improvement in the perceptual quality of speeches by the proposed technique.

1. Introduction

The problem of enhancing speech degraded by additive background noise, when only a single channel is available, remains challenging due to the insufficient information available to separate the underlying speech from the uncorrelated noise. In the past many algorithms have been proposed to solve this problem, such as the power spectral subtraction [1], the MMSE short time spectral amplitude estimator [2] and wiener filter based algorithms. These methods can be viewed in terms of applying spectral gain to each frequency bin in a short time frame of the noisy speech signal. The same gain function is applied regardless of the underlying SNR of each frequency bin. This can be problematic since some estimators can over attenuate the signal when operating at extremely low SNR levels. On the other hand, other estimators tend to apply little attenuation at the expense of introducing significant amounts of residual noise.

In most spoken languages, voiced sounds represent a large amount (around 80%) of the pronounced sounds. In the classic short-time suppression techniques some harmonics are considered as noise only components and are consequently suppressed by the noise reduction process. This is one major limitation of those methods. To overcome this limitation, a method, called regeneration of suppressed harmonics that takes into account the harmonic characteristic of speech, is proposed. In this approach, the output signal of classic noise reduction technique is further processed to create an artificial signal where in the missing harmonics are automatically regenerated. This artificial signal is used to refine the apriori SNR used to compute a spectral gain.

2. Speech Signal Estimator

A noisy speech signal $y(m,n)$ can be modelled as the sum of clean speech $s(m,n)$ and additive noise $d(m,n)$ in the frame m of the time domain, i.e.,

$$y(m,n) = s(m,n) + d(m,n) \quad (1)$$

and in frequency domain

$$Y(m,\omega) = S(m,\omega) + D(m,\omega) \quad (2)$$

where $Y(m,\omega)$, $S(m,\omega)$ and $D(m,\omega)$ are the spectral magnitudes of the noisy signal, clean speech signal and noise respectively. The spectral estimate of speech signal $\hat{S}(m,\omega)$ is obtained by multiplying a spectral gain factor $G(m,\omega)$ with the noisy speech spectrum $Y(m,\omega)$ as given in the *equ*.3.

$$\hat{S}(m,\omega) = G(m,\omega).Y(m,\omega) \quad (3)$$

where $G(m,\omega)$ assumed to be a function of apriori and posteriori SNRs. The apriori and a posteriori SNRs are defined as $\xi(m,\omega) = \frac{\lambda_s}{\lambda_d}$ and $Y(m,\omega) \cong \frac{|Y(m,\omega)|^2}{\lambda_d}$, where λ_s and λ_d are the variances of speech and noise respectively.

A. Speech signal estimation using noise statistics

In this paper, clean speech signal is estimated using the statistical estimators of the speech and noise. The statistical estimators of the speech and noise have a certain symmetry, in that one can derive the estimator for the noise signal based on the estimator of the speech signal by making the appropriate substitution of the constituent parameters $\xi(m,\omega)$ and $\gamma(m,\omega)$.

According to the decision directed algorithm [2] the apriori SNR estimate

$$\hat{\xi}(m,\omega) = \alpha Y(m-1,\omega) G^2(m-1,\omega) + (1-\alpha) P[Y(m,\omega)-1] \quad (4)$$

Tele:

E-mail address: subbarao.genikala@gmail.com

where α $(0 \le \alpha < 1)$, is the smoothing factor and the operator $P[\cdot]$ is defined by

$$P[l] = \begin{cases} l & (l \ge 0), \\ 0 & (otherwise) \end{cases} \tag{5}$$

And also the estimator of the speech magnitude spectrum is defined by

$$G_s(m,\omega) = \Gamma(1.5)\frac{\sqrt{v(m,\omega)}}{\hat{\gamma}(m,\omega)}\exp\left(-\frac{v(m,\omega)}{2}\right) \cdot \left[(1+v(m,\omega))I_0\left(\frac{v(m,\omega)}{2}\right) + v(m,\omega)I_1\left(\frac{v(m,\omega)}{2}\right)\right] \tag{6}$$

$\Gamma(\cdot)$ denotes the gamma function, $I_0(\cdot)$ and $I_1(\cdot)$ denotes the modified Bessel functions of zero and first-order, respectively. Moreover, $v(m,\omega)$ is defined by

$$v(m,\omega) = \frac{\hat{\xi}(m,\omega)}{1+\hat{\xi}(m,\omega)}\hat{\gamma}(m,\omega) \tag{7}$$

Where the parameters $\xi_d(m,\omega)$ and $\gamma_d(m,\omega)$ are defined as:

$$\xi_d(m,\omega) = \lambda_d / \lambda_s = 1/\xi_s(m,\omega),$$
$$\gamma_d(m,\omega) = \Upsilon_k^2(m,\omega)/\lambda_s = \gamma_s(m,\omega)/\xi_s(m,\omega) \tag{8}$$

The new gain function, denoted as H_m, applies the same rule as its counterpart $G_s(m,\omega)$, and the parameters are now defined with respect to the noise. The estimator of the noise magnitude spectrum can be computed using

$$H_{\log MMSE(\xi(m,\omega),\gamma(m,\omega))} = \frac{\xi_d(m,\omega)}{\xi_d(m,\omega)+1}\exp\left\{\frac{1}{2}\int_{\upsilon_d}^{\infty}\frac{e^{-t}}{t}dt\right\}$$
$$= \frac{1}{\xi_s(m,\omega)+1}\exp\left\{\frac{1}{2}\int_{\upsilon_d}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{9}$$

Where

$$\upsilon_d = \gamma_s(m,\omega) \Big/ \{\xi_s(m,\omega)(\xi_s(m,\omega)+1)\}.$$

Now we can easily compute the speech spectral component as follows:

$$\hat{S}(m,\omega) = \Upsilon(m,\omega) - \hat{D}(m,\omega)$$
$$= \left[\Upsilon(m,\omega) - \hat{H}_m(\xi,\gamma)\Upsilon(m,\omega)\right]$$
$$= \Upsilon(m,\omega)\left[1-\hat{H}_m(\xi,\gamma)\right]$$
$$= \Upsilon(m,\omega)\cdot G_d(m,\omega) \tag{10}$$

where $G_d(m,\omega)$ is the speech estimator derived from the noise spectrum estimate. Note that in the case of the wiener estimator, due to its symmetry, the $G_s(m,\omega)$ and $G_d(m,\omega)$ gain functions are the same. This is not the case for other estimators such as the MMSE and logMMSE estimators.

B. Speech signal estimation by measuring segmental SNR measure in frequency domain

The frequency-weighted segmental SNR measure [3][4] has been found in [5][6] to correlate highly with speech quality and intelligibility. This ratio denoted as SNR_{seg} is defined as:

$$SNR_{seg}(\xi(m,\omega)) = \frac{\mathrm{E}\left\{|S(m,\omega)|^2\right\}}{\mathrm{E}\left\{\left|S(m,\omega)-\hat{S}(m,\omega)\right|^2\right\}}$$
$$= \frac{\mathrm{E}\left\{|S(m,\omega)|^2\right\}}{\mathrm{E}\left\{|S(m,\omega)-G_k\Upsilon(m,\omega)|^2\right\}}$$
$$= \frac{\xi(m,\omega)}{(1-G_k(m,\omega))^2 + G_k{}^2(m,\omega)} \tag{11}$$

where $G_k(m,\omega)$ is the gain function defining a statistical estimator. By substituting the value of $G_s(m,\omega)$ and $G_d(m,\omega)$ in the *equ.*11, variation of segSNR in terms of $\xi(m,\omega)$ and $\gamma(m,\omega)$ verified experimentally. From the experimental results it is clear that the $G_d(m,\omega)$ estimator is more appropriate for low SNR regions, while the $G_s(m,\omega)$ estimator is more appropriate for higher SNR regions. This becomes more evident when $\gamma_s(m,\omega) < 0dB$. A similar SNR_{seg} vs. $\xi_s(m,\omega)$ pattern was observed with other estimators. We can conclude that in order to maximize the signal-to-residual spectrum ratio (and subsequently speech quality), the following rule needs to be adopted:

$$\hat{S}(m,\omega) = G_c \cdot \Upsilon(m,\omega) = \begin{cases} G_d \cdot \Upsilon(m,\omega)\ \xi_s(m,\omega), \gamma_s(m,\omega) \le 0dB, \\ G_s \cdot \Upsilon(m,\omega)\ otherwise \end{cases} \tag{12}$$

where the subscript c denotes the "combined" or hybrid estimator.

3. Speech Harmonic Regeneration

Plapous [7] introduced a simple and efficient way to restore speech harmonics in a noisy speech. In such approach, a non-linear function NL is applied to a speech signal. Then the restored harmonics $s_{harm}(m,n)$ is obtained by

$$s_{harm}(m,n) = NL(\hat{s}(m,n)) \tag{13}$$

Note that the restored harmonics $s_{harm}(m,n)$ are generated at the same positions as the clean speech ones. The signal $s_{harm}(m,n)$ cannot be used directly as clean speech estimation, because the harmonic amplitudes of this artificial signals are biased compared to clean speech. However, it contains very useful information that can be exploited to improve the estimation of the a-priori SNR.

$$\hat{\xi}_{harm}(m,\omega) = \frac{\rho(m,\omega)\hat{S}(m,\omega) + (1-\rho(m,\omega))\hat{S}_{harm}(m,\omega)}{\hat{S}_{noise}(m,\omega)} \tag{14}$$

The parameter $\rho(m,\omega)$ is used to control the mixing level of $\hat{S}(m,\omega)$ and $\hat{S}_{harm}(m,\omega)$ depending on the chosen non-linear function (0< $\rho(m,\omega)$<1). It is necessary to combine $\hat{S}(m,\omega)$ and $\hat{S}_{harm}(m,\omega)$, because the harmonic function can restore harmonics at the desired frequencies, but also with biased amplitudes.

The improved a-priori SNR, $\hat{\xi}_{harm}(m,\omega)$ can be used to obtain a new suppression gain, $G_{harm}(m,\omega)$, which will be able to preserve most of the harmonics of the speech signal.

$$G_{harm}(m,\omega)=v\left(\hat{\xi}_{harm}(m,\omega),\hat{\xi}(m,\omega)\right) \qquad (15)$$

Finally, the resulting speech spectrum is estimated as follows

$$\hat{S}(m,\omega)=G_{harm}(m,\omega)Y(m,\omega) \qquad (16)$$

Although the suppression gain $G_{harm}(m,\omega)$ has the ability to preserve the harmonics suppressed with most of the common speech enhancement algorithms, it could not avoid the distortions if the enhanced signal $\hat{S}(m,\omega)$is distorted already.

3. Proposed Post-Processing Technique

As discussed previously, the enhanced speech $\hat{S}(m,\omega)$ suffers from distortions since some weak components of speech are considered as the background noise and are suppressed together with the noise by common noise reduction algorithms. In order to correct this problem, a post-processing technique is proposed.

To allow the harmonic regeneration technique to be more effective, implement based on a-priori SNR estimator proposed in [8] to the MMSE-LSA in order to get an enhanced speech with higher SNR. Then, the harmonic regeneration technique is applied to preserve the speech harmonics. A Max function is used as the non-linear function NL in *equ.*13, which follows,

$$\hat{s}_{harm}(m,n)=Max\left(\hat{s}_{MMSE-LSA}(m,n),0\right) \qquad (17)$$

Following the approach in [7], the parameter $\rho(m,\omega)$ in *equ.*14, is selected to be the wiener filter gain function. However, a-priori SNR estimator $\xi(m,\omega)$ is proposed for computing the wiener estimator as below:

$\rho(m,\omega)=G_{WF}$, where

$$G_{WF}=\frac{\hat{\xi}(m,\omega)}{1+\hat{\xi}(m,\omega)} \qquad (18)$$

In this case, G_{WF} inherits the advantages of low variance structural noise and emphasized speech spectral peaks. Then, the a-priori SNR is improved as below:

$$\hat{\xi}_{harm}(m,\omega)=\frac{G_{WF}^{T}\hat{S}(m,\omega)+\left(1-G_{WF}\right)\hat{S}_{harm}(m,\omega)}{\hat{S}_{noise}(m,\omega)} \qquad (19)$$

The proposed a-priori SNR estimator $\hat{\xi}_{harm}(m,\omega)$ is applied to *equ.*12, in order to obtain a new suppression gain function $G_{harm}(m,\omega)$. The enhanced speech can be obtained by applying the new suppression gain $G_{harm}(m,\omega)$ to the spectrum of noisy speech as follows:

$$\hat{S}_{harm}(m,\omega)=G_{harm}(m,\omega)Y(m,\omega) \qquad (20)$$

4. Results

To evaluate and compare the performance of the proposed two stage speech enhancement, simulations are carried out with the NOIZEUS, A noisy speech corpus for evaluation of speech enhancement algorithms, database [9].

The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real world noises at different SNRs. Speech signals were degraded with different types of noise at global SNR levels of 0 dB, 5 dB, 10 dB and 15 dB. In this evaluation only six noises are considered those are babble, car, train, airport and street noise. The following objective quality measures used for the evaluation of the proposed speech enhancement method

4.1 Time-domain SNR measure

The time-domain segmental SNR (SNR $_{seg}$) measure [10] was computed is given by

$$SNR_{seg}=\frac{10}{M}\sum_{m=0}^{\omega-1}\log\frac{\sum_{n=N_m}^{N_m+N-1}s^2(n)}{\sum_{n=N_m}^{N_m+N-1}\left(s(n)-\hat{s}(n)\right)^2} \qquad (21)$$

Where $s(n)$ the input is (clean) signal, $\hat{s}(n)$ is the processed (enhanced) signal, N is the frame length and M is the number of frames in the signal. Table 1 presents the performance Comparisons in terms of the average segmental (Avg. SegSNR). From table1 it is clear that the proposed approach significantly outperforms the Wiener filter estimator and the power spectral subtraction.

Table 1. Output Avg. SegSNR (dB).

Type of noise and SNR (dB)	Wiener filter	Spectral Subtraction	Proposed method
Airport-0	-4.37	-4.06	-3.31
Airport-5	-2.57	-2.23	-1.73
Airport-10	-0.06	-0.68	-0.04
Airport-15	1.88	0.77	1.83
Babble-0	-4.59	-4.40	-3.63
Babble-5	-1.39	-1.80	-1.49
Babble-10	0.03	-0.16	0.69
Babble-15	2.71	0.7	2.29
Car-0	-3.93	-4.48	-3.31
Car-5	-1.65	-1.90	-0.99
Car-10	0.68	-0.08	0.63
Car-15	2.31	0.75	2.18
Street-0	-2.88	-3.16	-2.47
Street-5	-2.13	-2.28	-1.87
Street-10	1.20	-0.31	0.89
Street-15	2.25	0.58	2.08
Train-0	-3.45	-3.45	-3.03
Train-5	-0.86	-1.75	-1.03
Train-10	-0.39	-0.81	0.11
Train-15	2.62	0.49	2.35
Station-0	-3.62	-3.72	-2.87
Station-5	-1.93	-1.83	-1.30
Station-10	0.95	-0.36	0.72
Station-15	2.72	0.76	2.28

4.2 Log-likelihood ratio (LLR) measure

The performance of the proposed method was evaluated The LLR measure for each 20-ms speech frame was computed as follows:

$$d_{LLR}=\log_{10}\left(\frac{a_y R_s a_y^T}{a_s R_s a_s^T}\right) \qquad (22)$$

Where a_s and R_s are the linear prediction coefficient vector and autocorrelation matrix of the original(clean) speech frame respectively, and a_y is the linear prediction coefficient vector of the enhanced speech frame. The LLR is a spectral distance measure which mainly models the mismatch between the formats of the original and enhanced signals. The mean LLR value was obtained by averaging the individual frame LLR values across the sentence. The LLR results are tabulated in Table 2, smaller spectral distance values (LLR) were obtained by the proposed method.

Table 2. LLR values (dB).

Type of noise and SNR (dB)	Wiener filter	Power spectral subtraction	Proposed method
Airport-0	1.23	1.10	0.73
Airport-5	0.77	0.83	0.54
Airport-10	0.61	0.71	0.42
Airport-15	0.44	0.62	0.26
Babble-0	1.22	1.16	0.72
Babble-5	0.98	0.91	0.53
Babble-10	0.70	0.68	0.35
Babble-15	0.42	0.60	0.23
Car-0	1.23	1.15	0.72
Car-5	0.89	0.83	0.50
Car-10	0.58	0.66	0.38
Car-15	0.46	0.59	0.22
Street-0	1.32	1.20	0.72
Street-5	1.15	1.07	0.51
Street-10	0.68	0.68	0.35
Street-15	0.54	0.66	0.21
Train-0	1.37	1.27	0.75
Train-5	0.98	0.95	0.52
Train-10	0.85	0.83	0.48
Train-15	0.67	0.68	0.35
Station-0	1.13	1.00	0.69
Station-5	0.87	0.83	0.57
Station-10	0.61	0.68	0.34
Station-15	0.53	0.63	0.27

The timing waveforms and spectrograms for enhanced speech signal are shown in Fig.1 and Fig.2 respectively. From timing waveforms and spectrograms, it is confirmed that a reduction of the residual noise and speech distortion is achieved with proposed method.

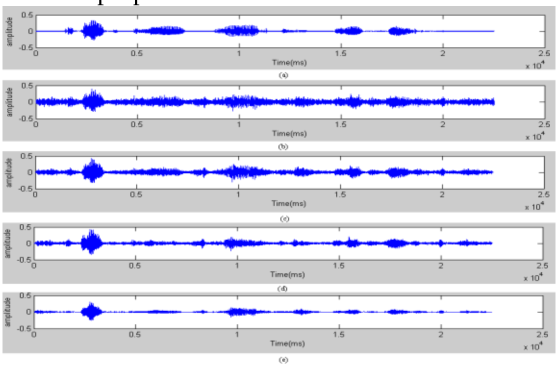

Fig 1. Timing waveforms of (a) clean speech, (b) noise corrupted speech signal with Babble noise at SNR 0dB and enhanced speech signals using (c) Wiener filtering (d) power spectral subtraction and (e) proposed method.

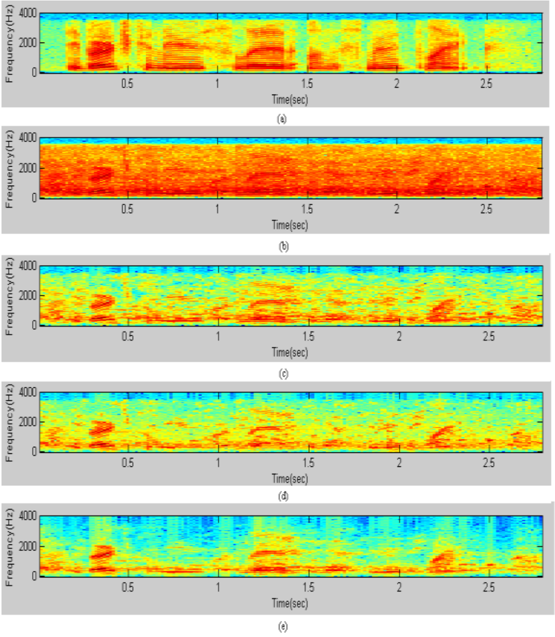

Fig 2. Spectrograms of (a) clean speech, (b) noise corrupted speech signal with Babble noise at SNR 0dB and enhanced speech signals using (c) Wiener filtering (d) power spectral subtraction and (e) proposed method.

Conclusion

In this research article suppressed harmonics are regenerated. Objective and subjective tests were carried out to demonstrate improvement in the perceptual quality of speeches by the proposed technique. The proposed approach significantly outperforms the Wiener filter estimator and the power spectral subtraction. Mean LLR value was obtained by averaging the individual frame LLR values across the sentence. The LLR results are smaller spectral distance values (LLR) were obtained by the proposed method. Timing waveforms and spectrograms, it is confirmed that a reduction of the residual noise and speech distortion is achieved with proposed method.

References

[1] Berouti, M., Schwartz, R., Makhoul, J., 1979.Enhacement of speech corrupted by acoustic noise. In: Proc. ICASSP-79, pp. 208-211.

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

[3] J. Tribolet, P. Noll, B. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," Proc.
IEEE Int. Conf. Acoust., Speech , Signal Process., pp. 586-590, 1978.

[4] S. Quackenbush, T. Barnwell, and M. Clements, Objective Measures of Speech Quality. Englewood Cliffs, NJ: Prentice-Hall,1988.

[5] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement." IEEE Trans. On Audio, Speech, and Language Process., vol. 16, no. 1, pp. 229-238, 2008.

[6] J. Ma, Y. Hu and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am., Vol. 125, no. 5, pp. 3387-3405, May 2009.

[7] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” IEEE Trans. Acoustic., Speech and Signal Process, vol. 14, no.6, Nov. 2006, pp. 2098-2108.

[8] Daniel P.K. Lun and Tai-Chiu Hsung, “Improved Wavelet Based A-Priori SNR Estimation for Speech Enhancement”, in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), Paris, France, May 2010, pp.2382-2385.

[9] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” IEEE Trans. Acoustic, Speech, and Signal Process., Vol.33, no. 2, Apr. 1985, pp. 443-445.

[10] Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504-512.

Author Profile

Subbarao Genikala received the M.Tech degree in Digital Electronics and Communication Systems from JNTU Kakinada, India in 2014 .He completed his B.Tech in the stream of Electronics and Communication Engineering from Acharya Nagarjuna University, India in 2012, received the Diploma from the Department of Electronics and Communication Engineering, Bapatla, India in 2009.Since 2014, he has been a Assistant Professor in the Department of Electronics and Communication Engineering, Madanapalle Institute of Technology & Science, Madanapale, Andhra Pradesh, India. He has published one International Journal. Where he is currently working towards the Ph.D degree. His current research interest is focused on Bio-medical Microwave Antennas.