Int J Speech Technol (2017) 20:89–98
DOI 10.1007/s10772-016-9391-z

Single channel noise reduction system in low SNR

Nasir Saleem[1]

Received: 17 August 2016 / Accepted: 13 November 2016 / Published online: 19 November 2016

Abstract We propose a two stage noise reduction system for reducing background noise using single-microphone recordings in very low signal-to-noise ratio (SNR) based on Wiener filtering and ideal binary masking. The proposed system contains two stages. In first stage, the Wiener filtering with improved a priori SNR is applied to noisy speech for background noise reduction. In second stage, the ideal binary mask is estimated at every time–frequency channel by using pre-processed first stage speech and comparing the time–frequency channels against a pre-selected threshold T to reduce the residual noise. The time–frequency channels satisfying the threshold are preserved whereas all other time–frequency channels are attenuated. The results revealed substantial improvements in speech intelligibility and quality over that accomplished with the traditional noise reduction algorithms and unprocessed speech.

Keywords Ideal binary masking · SNR · Speech intelligibility · Speech quality · Wiener filtering

1 Introduction

Noise reduction systems are extensively used telecommunication systems to enhance the quality of the speech communication in noisy environments. Although, an improved noise reduction can be realized by using microphone array system, but for economic reasons, most of these systems are based on single microphone. In principle, a single microphone noise reduction system uses adaptive filtering operations to attenuate time–frequency (T–F) units of the noisy speech that have low SNR and retain the T–F units with high SNR. By doing so, the essential regions of speech are preserved whereas the noise level is greatly reduced, leading to an enhanced speech with reduced noise level. Countless noise reduction systems are available in literature along this line (Boll 1979; Lim and Oppenheim 1978; Scalart and Filho 1996; Ephraim and Malah 1984, 1985). Wiener filter (Scalart and Filho 1996; Abd El-Fattah et al. 2014) is a linear filter employed to recover original speech signal from the noisy signal by minimizing the mean square error (MSE) between estimated/enhanced signal and the original one. In Wiener filtering, some attenuation rules are used to decide which T–F unit of noisy speech need to be attenuated and how much. Usually, these attenuation rules are optimized in such a way that the enhanced speech is as close as possible to the clean speech. Clearly, the quality of single microphone noise reduction systems is determined by the suppression rule. In general, a suppression rule with strong attenuation will lead to a less noisy speech, however, strong attenuation results in more distortion. On other hand, a moderate attenuation introduces less distortion but achieved limited amount of noise reduction. For this reason, a balance trade-off has to be made to achieve a speech signal with low distortion and high quality. To end this, ideal binary masking (IdBM) which is successfully applied in noise reduction systems. These masks are constructed to retain time–frequency (T–F) units when estimated speech is stronger than intrusive noise (SNR > 0 dB) and removes T–F units when intrusive noise is dominant (SNR ≤ 0 dB). The estimate of these masks can be achieved either using the single-microphone or the multi-microphone systems. A widespread literature

✉ Nasir Saleem
nasirsaleem@gu.edu.pk

[1] Department of Electrical Engineering, Gomal University, Dera Ismail Khan 29050, KPK, Pakistan

review on time–frequency masking can be found in the (Wang 2008). Methodologies employing binary masks have revealed generous quality improvements even at extremely low SNRs with less distortion. These optimistic results have reinvigorated the researchers to develop/estimate binary masks and suggested it as the goal of computational auditory scene analysis (CASA) (Wang 2005). With these evidences of quality/intelligibility improvement, research is done in the recent past in trying to estimate these masks (Boldt et al. 2008; Saleem et al. 2015a, 2015; Loizou 2009).

In this study a two-stage noise reduction system for the noise reduction is proposed which is based on Wiener filtering, employing an improved a priori SNR [to reduce one-frame delay offered by the decision-direct approach (Ephraim and Malah 1984)] and the ideal binary mask (Wang 2005). The ideal binary mask can be defined by relating a priori SNR estimate against the threshold (usually 0 dB). However, instead of a priori SNR, ideal binary mask estimation needs access to local instantaneous SNR which is defined as ratio of power spectrum of speech to the power spectrum of noise at every T–F unit. The performance of the proposed systems is evaluated with two different intruder's noise (babble, white noise) in terms of the speech distortion and residual noise. The rest of the paper is arranged as; in Sect. 2, a review of the proposed noise reduction system is presented, the Sect. 3 presents experimental setup, the Sect. 4 shows the results and analysis. Finally, the concluding remarks are given in the Sect. 5.

2 The overview of the proposed noise reduction system

This section provides as overview of the proposed noise reduction system. In classical noise reduction model, the noisy speech is given by equation;

$$\mathrm{y}(t) = \mathrm{s}(t) + \mathrm{e}(t) \tag{1}$$

where s(t) and e(t) specify clean speech and the noise respectively. Let $\mathrm{Y}(m,\omega_m)$, $\mathrm{S}(m,\omega_m)$ and $\mathrm{E}(m,\omega_m)$ categorized ω_m spectral component of short-time frame m of noisy speech y(t), clean speech s(t) and noise e(t) respectively. Both speech and noise are non-stationary in nature, however, in short-intervals (10–30 ms), both are supposed to be stationary, hence, the quasi-stationary nature is supposed in frame analysis. To reduce noise level, a spectral gain $\mathrm{G}(m,\omega_m)$ is multiplied to every short-time spectrum of the $\mathrm{Y}(m,\omega_m)$. Figure 1 demonstrates the block diagram of the proposed system. Practically, the spectral gain is involved in calculation of two prime SNR estimations, a posteriori and a priori SNR and is given as:

$$\gamma(m,\omega_m) = \frac{|\mathrm{Y}(m,\omega_m)|^2}{\mathrm{E}\{|\mathrm{E}(m,\omega_m)|^2\}} = \frac{|\mathrm{Y}(m,\omega_m)|^2}{\sigma_e^2(m,\omega_m)} \tag{2}$$

$$\xi(m,\omega_m) = \frac{|\mathrm{S}(m,\omega_m)|^2}{\mathrm{E}\{|\mathrm{E}(m,\omega_m)|^2\}} = \frac{\sigma_\mathrm{S}^2(m,\omega_m)}{\sigma_\mathrm{e}^2(m,\omega_m)} \tag{3}$$

where E{.} is expectation operator, $\gamma(m,\omega_m)$ and $\xi(m,\omega_m)$ is a posteriori and a priori SNR respectively. In real-world applications of a noise reduction systems, the power spectrum density of the clean speech $|\mathrm{S}(m,\omega_m)|^2$ and the noise $|\mathrm{E}(m,\omega_m)|^2$ are unidentified as merely the noisy speech is reachable. Therefore; both the instantaneous and a priori SNR are needed to be estimated. The power spectral density of noise can be estimated through speech gaps exploiting the standard recursive relation, given as:

$$\hat{\sigma}_\mathrm{e}^2(m,\omega_m) = \zeta\hat{\sigma}_\mathrm{e}^2(m-1,\omega_m) + (1-\zeta)\tilde{\sigma}_\mathrm{Y}^2(m-1,\omega_m) \tag{4}$$

where, ζ is the smoothing factor and $\tilde{\sigma}_\mathrm{Y}^2(m-1,\omega_m)$ is the estimate from existing frame. The two signal-to-noise ratios can be computed as:

$$\mathrm{SNR}_\mathrm{INSTANT}(m,\omega_m) = \frac{|\mathrm{Y}(m,\omega_m)|^2}{\sigma_\mathrm{e}^2(m,\omega_m)} - 1 \tag{5}$$

$$\xi_\mathrm{PRIO}^\mathrm{DD}(m,\omega_m) = \beta\frac{|\mathrm{G}(m-1,\omega_m) * \mathrm{Y}(m,\omega_m)|^2}{\hat{\sigma}_\mathrm{e}^2(m,\omega_m-1)} + (1 - \beta)\mathrm{F}\{\ \mathrm{SNR}_\mathrm{INSTANT}(m,\omega_m)\} \tag{6}$$

where $\xi_\mathrm{PRIO}^\mathrm{DD}(m,\omega_m)$ represents the a priori SNR calculation using decision-direct (DD) approach and F{.} shows the full-wave rectification. The decision-direct is computationally effective technique and performs remarkable in noise reduction applications, however, in this technique, the a priori SNR tails the shape of instantaneous SNR which leads to one-frame delay. In order to reduce single-frame delay, the improved version of the a priori SNR is used by introducing momentum terms to improve the tracking speech of proposed system. The improved version of a priori SNR can be written as:

$$\begin{gathered}\xi_\mathrm{PRIO}^\mathrm{DD-MT}(m,\omega_m) = \beta\frac{|\mathrm{G}(m-1,\omega_m) * \mathrm{Y}(m,\omega_m)|^2}{\hat{\sigma}_\mathrm{e}^2(m,\omega_m-1)} \\ + \lambda(m,\omega_m) + (1-\beta)\mathrm{F}\{\ \mathrm{SNR}_\mathrm{INSTANT}(m,\omega_m)\} \\ \lambda(m,\omega_m){=}\psi((\xi_\mathrm{PRIO}^\mathrm{DD}(m,\omega_m-1) - \xi_\mathrm{PRIO}^\mathrm{DD}(m,\omega_m-2))\end{gathered} \tag{7}$$

$\xi_\mathrm{PRIO}^\mathrm{DD-MT}(m,\omega_m)$ shows a priori SNR calculation using modified decision-direct method by inserting momentum terms, the $\lambda(m,\omega_m)$ is the momentum terms, $\psi(m,\omega_m)$ is called the momentum parameter ($\psi = 0.998$) and $\beta(m,\omega_m)$ is the smoothing parameter (usually $\beta = 0.98$) The estimated power spectrum of the clean speech $\mathrm{S}_\mathrm{EST}(m,\omega_m)$ is

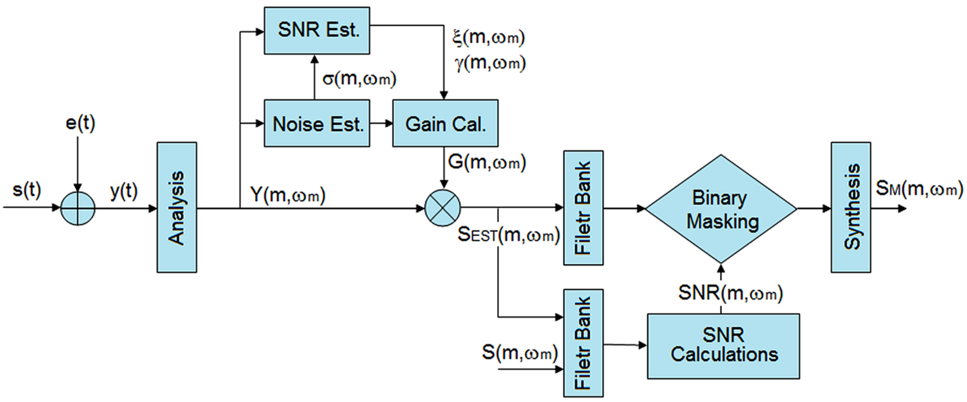


Fig. 1 Block diagram of proposed system

computed from the noisy speech $\mathrm{Y}(m,\omega_m)$ by multiplying with Wiener filter gain function:

$$|\mathrm{S}_{\mathrm{EST}}(m,\omega_m)| = |\mathrm{Y}(m,\omega_m)| * G^{\mathrm{DD-MT}}_{\mathrm{SQWF}}(m,\omega_m) \tag{8}$$

The square root Wiener gain function $\mathrm{G}^{\mathrm{DD}}_{\mathrm{SQWF}}(m,\omega_m)$ is given by equation:

$$\mathrm{G}^{\mathrm{DD}}_{\mathrm{SQWF}}(m,\omega_m) = \sqrt{\frac{\xi^{\mathrm{DD}}_{\mathrm{PRIO}}(m,\omega_m)}{\xi^{\mathrm{DD}}_{\mathrm{PRIO}}(m,\omega_m)+1}} \tag{9}$$

With improved a priori SNR, the gain function $\mathrm{G}^{\mathrm{DD-MT}}_{\mathrm{SQWF}}(m,\omega_m)$ in Eq. (8) becomes:

$$\mathrm{G}^{\mathrm{DD-MT}}_{\mathrm{SQWF}}(m,\omega_m) = \sqrt{\frac{\xi^{\mathrm{DD-MT}}_{\mathrm{PRIP}}(m,\omega_m)}{\xi^{\mathrm{DD-MT}}_{\mathrm{PRIP}}(m,\omega_m)+1}} \tag{10}$$

To remove/reduce residual noise, the pre-processed signals are inserted to the second stage. Although, the pre-processed speech offers reasonable speech quality, however, residual noise remains deceptive and annoying under substantial noisy situations. The estimate of clean speech is used for computing instantaneous SNR in second stage. The $\mathrm{SNR}_{\mathrm{INSTANT}}$ is computed as:

$$\mathrm{SNR}_{\mathrm{INSTANT}}(\mathrm{m},\omega_m) = 10\log_{10}\left(\frac{|\mathrm{S}(\mathrm{m},\omega_m)|}{|\mathrm{S}_{\mathrm{EST}}(\mathrm{m},\omega_m)|}\right) \tag{11}$$

The short-term energies of filtered waveforms are calculated followed by the comparison stage. To reduce residual noise, ratio of estimated magnitude spectrum to clean speech ($|\mathrm{S}(m,\omega_m)|/|\mathrm{S}_{\mathrm{EST}}(m,\omega_m)|$) is compared against a predefined threshold T. The T–F units satisfying the constraint i.e. ($|\mathrm{S}(m,\omega_m)|/|\mathrm{S}_{\mathrm{EST}}(m,\omega_m)|$) > T are preserved whereas T–F units violating the constraints i.e. ($|\mathrm{S}(m,\omega_m)|/|\mathrm{S}_{\mathrm{EST}}(m,\omega_m)|$) < T are attenuated. The modified magnitude spectrum $\mathrm{S}_{\mathrm{M}}(m,\omega_m)$ is calculated as:

$$|\mathrm{S}_{\mathrm{M}}(\mathrm{m},\omega_m)| = \begin{cases} |\overline{\mathrm{S}}_{\mathrm{EST}}(\mathrm{m},\omega_m)| & |\mathrm{S}_{\mathrm{EST}}(\mathrm{m},\omega_m)|/|\mathrm{S}(\mathrm{m},\omega_m)| \geq \mathrm{T} \\ 0 & |\mathrm{S}_{\mathrm{EST}}(\mathrm{m},\omega_m)|/|\mathrm{S}(\mathrm{m},\omega_m)| < \mathrm{T} \end{cases} \tag{12}$$

Following the selection of the T–F units, an inverse STFT is applied to modified speech using the phase of the noisy speech spectrum followed by the overlap-and-add method to synthesize noise-suppressed/reduced speech.

3 Experiments: methodology and setup

This section offers experimental setup and methodology to assess the performance and suitability of the proposed noise reduction system. In experiments, the Noizeus (Hu and Loizou 2007) corpus was engaged which was composed of 30-phonetically balanced sentences belonging to three male and three female speakers. The sentences were sampled at 8 kHz frequency and filtered to simulate the frequency characteristics of telephone handsets. The corpus was originated with non-stationary noises at various SNRs. However; our experiments kept only clean sentences. The noisy stimuli were generated by adding clean sentences with babble and white noise using the ITU-T Recommendation P.56 (ITU-T P.56 1993). Three signal-to-noise ratio levels, including -5, 0, and 5 dB were used to assess the performance. The noise sources were taken from AURORA (Hirsch and Pearce 2000) database. The ITU-T Recommendation P.862 (PESQ) (Rix et al. 2001) was used to predict the mean opinion scores (MOS) and ITU-T Recommendation P.835 (ITU-T P.835 2003) was used to predict the amount of residual noise (BAK) and speech distortion (SIG). The spectrogram analysis was also performed to assess the proposed system. To measure the speech intelligibility, the normalized subband envelop correlation (NSEC) (Boldt and Ellis 2009) measure is used which is a good alternate to the speech intelligibility index (SII) and speech transmission index (STI).

4 Objective measures

A number of objective measures are derived in the literature to evaluate the performance of noise reduction systems

Table 1 Performance comparison in terms of PESQ–MOS scores, and improvement (ΔPESQ) between first and second stage

Noise type	SNR (in dB)	Un-processed	First stage	Second stage	ΔPESQ
Babble	−5	1.32	1.63	2.48	0.85
	0	1.51	1.87	2.78	0.88
	5	1.79	2.12	2.86	0.74
White	−5	1.32	1.51	2.27	0.76
	0	1.66	1.69	2.75	1.07
	5	1.91	2.02	2.93	0.91

(Rix et al. 2001; Hansen and Pellom 1998; Klatt 1982; Quackenbush et al. 1988; Kitawaki et al. 1988). The most extensively used objective measure includes PESQ–MOS and segmental SNR ($\mathrm{SNR_{SEG}}$) (Hansen and Pellom 1998). The PESQ–MOS measure which was not originally designed to assess the performance of noise reduction systems, however, it has been found to have good correlation with mean opinion score (MOS). It predicts the MOS scores which yields results from 1 to 5, where high score indicates better speech quality. Similarly, $\mathrm{SNR_{SEG}}$ is another widely used objective measure and it has the best correlation with background noise reduction. The $\mathrm{SNR_{SEG}}$ is defined as:

$$\mathrm{SNR_{SEG}}(\mathrm{m},\omega_m)=\frac{10}{\mathrm{M}}\sum_{\mathrm{m}=0}^{\mathrm{M}-1}\log_{10}\left(\frac{|\mathrm{S}(\mathrm{m},\omega_m)|^2}{||\mathrm{S}(\mathrm{m},\omega_m)|-|\mathrm{S_{EST}}(\mathrm{m},\omega_m)||^2}\right) \tag{13}$$

where $\mathrm{S}(\mathrm{m},\omega_\mathrm{m})$ and $\hat{\mathrm{S}}(\mathrm{m},\omega_\mathrm{m})$ shows the frames of clean and estimated speech respectively. To discard non-speech frames, every frame was threshold by a 0 dB lower bound and −35 dB upper bound. The performance of a noise reduction system has a trade-off among musical noise, speech distortion and noise reduction. Both PESQ–MOS and $\mathrm{SNR_{SEG}}$ cannot portray the whole picture of these trade-offs. Therefore, ITU-T Recommendation P.835 (composite measure) is used to measure the speech distortion and residual noise. The P.835 measure is formulated by relating the basic objective measures to establish composite measure (Loizou 2007), given as:

$$\begin{aligned}\mathrm{Csig} &= 3.093-1.029\mathrm{S_{LLR}}+0.603\mathrm{S_{PESQ}}-0.009\mathrm{S_{WSS}}\\ \mathrm{Cbak} &= 1.634+0.478\mathrm{S_{PESQ}}-0.007\mathrm{S_{WSS}}+0.063\mathrm{S_{SNR_{SEG}}}\end{aligned} \tag{14}$$

where $\mathrm{S_{PESQ}}$, $\mathrm{S_{LLR}}$, $\mathrm{S_{WSS}}$ and $\mathrm{S_{SNRSEG}}$ represents perceptual evaluation of speech quality (PESQ), log-likelihood ratio (LLR) and weighted-slope spectral (WSS) distance respectively.

4.1 Objective performance evaluation

The objective evaluation was performed for noisy (unprocessed) speech, Weiner filtering, spectral subtraction, ideal ratio making (IdRM) and the proposed system respectively. The measurements employed were PESQ–MOS, $\mathrm{SNR_{SEG}}$, and composite measure (speech distortion, SIG and residual noise, BAK). For all measuring parameters, the high scores indicate better speech quality.

4.2 PESQ evaluation

The Table 1 shows the performance comparison in terms of the perceptual evaluation of speech quality (PESQ–MOS) among noisy speech, first stage and second stage respectively. A remarkable improvement in PESQ–MOS was observed with proposed systems. The highest improvement in PESQ–MOS was observed with 0 dB white noise (Δ = 1.07) while the lowest improvement was observed with 5 dB babble noise. Significant improvements in PESQ–MOS were observed with proposed systems when compared to the noisy speech and the highest improvement is reported with 0 dB babble noise (Δ = 1.27) while the lowest improvement is obtained with −5 dB white noise (Δ = 0.95). The Table 2 shows observations in terms of speech quality (PESQ–MOS) of proposed system against noisy speech, speech processed by the spectral subtraction, Weiner Filtering, and Ideal ratio mask respectively. The highest PESQ–MOS scores are obtained with the ideal ratio mask (IRM) which is understandable. The boldface shows the best performance in reference to noisy speech, Spectral Subtraction and Weiner Filtering.

4.3 Segmental SNR evaluation

Table 3 shows the performance comparison in terms of the segmental SNR ($\mathrm{SNR_{SEG}}$) between the noisy speech, the first stage and the second stage respectively. An improvement in $\mathrm{SNR_{SEG}}$ was observed with proposed systems in all noise conditions. The highest and lowest improvements in $\mathrm{SNR_{SEG}}$ were noted.

With the 0 dB white noise (Δ = 2.58) and −5 dB babble noise (Δ = 1.46) respectively. The improvement in $\mathrm{SNR_{SEG}}$ clearly shows that significant noise reduction was achieved with proposed systems (by applying the second stage). By observing the results in Table 3, the

Table 2 Performance comparison in terms of PESQ–MOS scores between different noise reduction algorithms

Noise type	SNR (in dB)	Un-processed	Spectral subtraction	Weiner filtering	Ideal ratio mask	Proposed system
Babble	−5	1.32	1.51	1.59	2.87	2.48
	0	1.51	1.81	1.81	3.11	2.78
	5	1.79	2.33	2.02	3.37	2.86
White	−5	1.32	1.58	1.61	2.97	2.27
	0	1.66	1.64	1.63	3.24	2.75
	5	1.91	2.21	2.01	3.53	2.93

Table 3 Performance comparison in terms of SNR_{SEG} scores, and improvement (ΔSNR_{SEG}) between first and second stage

Noise type	SNR (in dB)	Un-processed	First stage	Second stage	ΔSNR_{SEG}
Babble	−5	0.31	0.34	1.60	1.46
	0	1.06	1.11	2.74	2.42
	5	2.44	1.92	3.28	2.26
White	−5	0.27	0.29	2.14	2.02
	0	0.96	0.77	2.85	2.58
	5	2.44	1.95	3.34	2.39

Table 4 Performance comparison in terms of speech distortion (SIG), and improvement (ΔSIG) in distortion between first and second stage

Noise type	SNR (in dB)	Un-processed	First stage	Second stage	ΔSIG
Babble	−5	2.13	2.18	2.43	0.25
	0	2.45	2.48	2.97	0.49
	5	2.95	2.95	3.41	0.46
White	−5	1.45	1.49	2.39	0.90
	0	1.87	1.89	2.67	0.78
	5	2.29	2.31	3.29	0.98

Table 5 Performance comparison in terms of residual noise (BAK), and improvement (ΔBAK) in distortion between first and second stage

Noise type	SNR (in dB)	Un-processed	First stage	Second stage	ΔBAK
Babble	−5	1.41	1.88	2.45	0.57
	0	1.68	2.05	2.67	0.62
	5	2.04	2.29	2.82	0.53
White	−5	1.86	1.84	2.41	0.57
	0	2.01	1.97	2.68	0.71
	5	2.28	2.18	2.92	0.74

improvements in SNR_{SEG} after first stage were negligible when compared to noisy speech, however, considerable improvements in SNR_{SEG} were observed with proposed systems when compared to the noisy speech.

4.4 Composite measure evaluation

Both PESQ–MOS and SNR_{SEG} cannot portray the whole picture of the trade-off between residual noise and speech distortion. To measure the speech distortion introduced by the noise reduction system and the amount of residual noise, composite measure was used (discussed in Sect. 4). Table 4 shows the speech distortion introduced by first stage, second stage and the improvement in speech distortion (ΔSIG) respectively. A high amount of speech distortion was introduced by first stage of proposed system which is less evident in the second stage (high scores of SIG). The highest and lowest gains in SIG scores were observed at 5 dB white noise ($\Delta = 0.98$) and −5 dB babble noise (0.25) respectively. Table 5 shows the amount of residual noise (BAK) in enhanced speech after processed by first and second stage respectively. A considerable amount of background noise was reduced (i.e., less residual noise) by the first stage of proposed system (high BAK values) which was further reduced by the second stage (high BAK values for second stage). The highest and lowest gains in BAK scores were observed at 5 dB white noise ($\Delta = 0.74$) and 5 dB babble noise ($\Delta = 0.53$)

Fig. 2 NSEC based speech intelligibility prediction in various noisy backgrounds

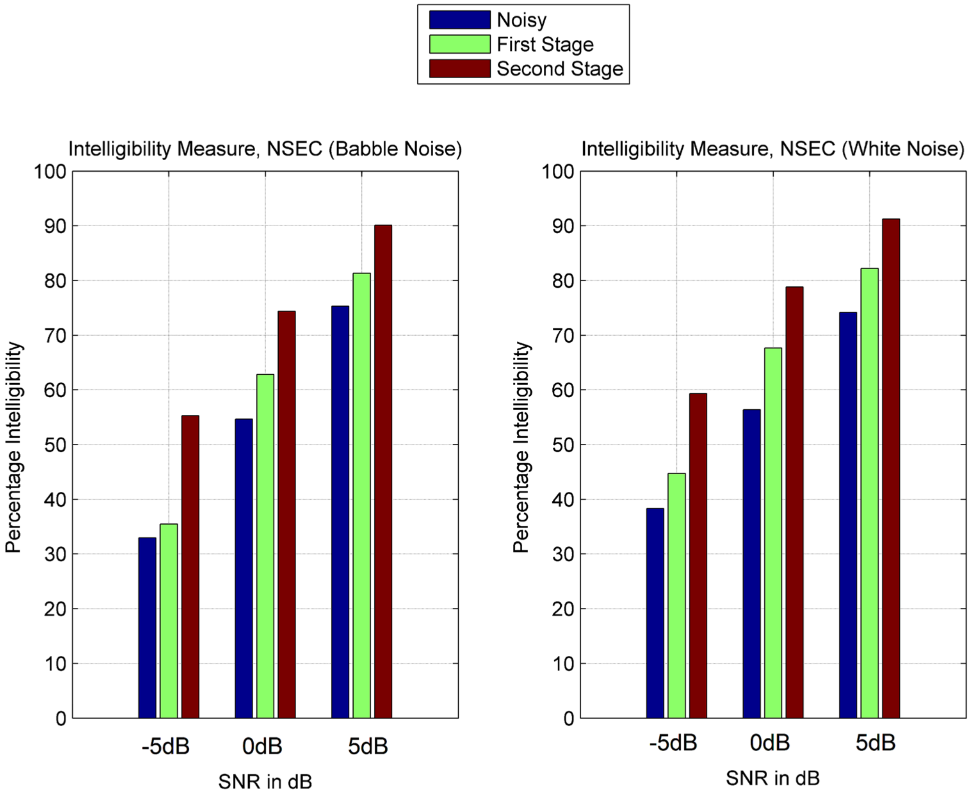
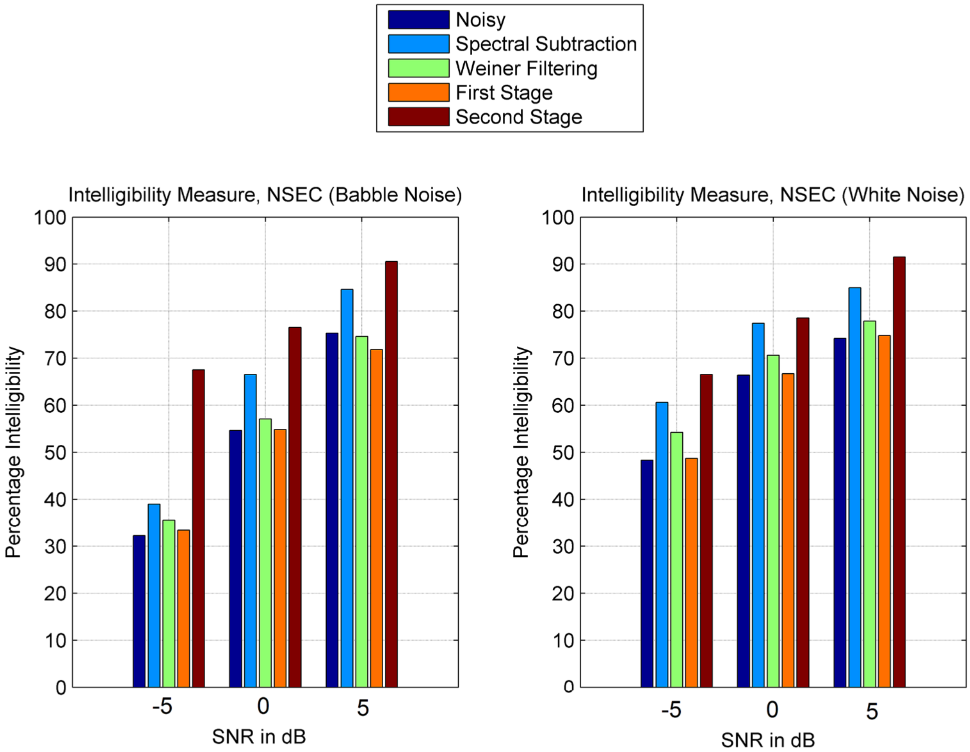


Fig. 3 NSEC based speech Intelligibility prediction for different start-of-the-art noise reduction systems



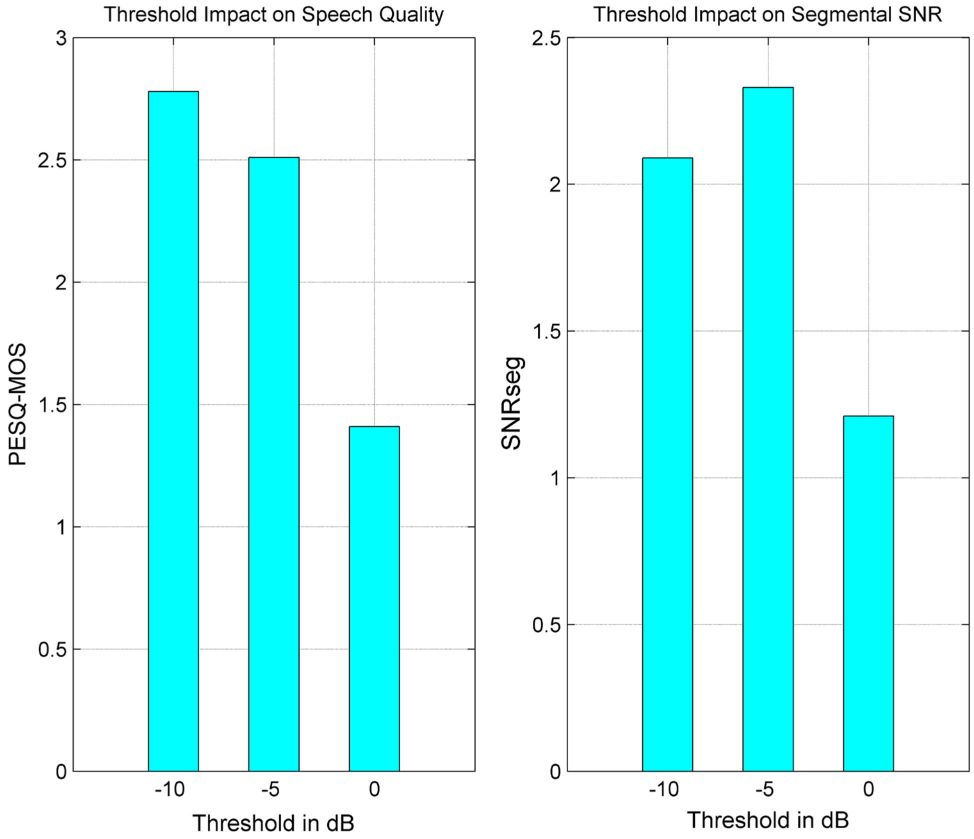


Fig. 4 Impact of threshold on PESQ–MOS score and SNRSEG scores

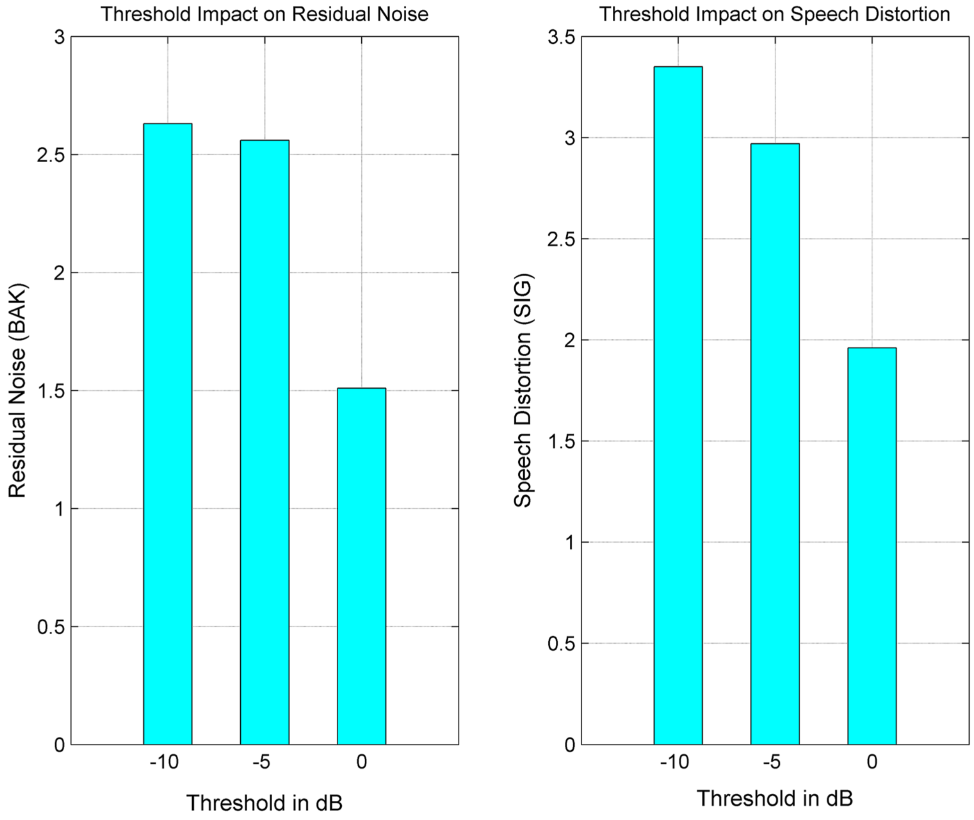


Fig. 5 Impact of threshold on residual noise and speech distortion

respectively. The composite measure indicates that low speech distortion, less residual noise and high quality speech was obtained with proposed system in all noise conditions.

5 Speech intelligibility measure

To measure the speech intelligibility, the normalized subband envelop correlation (NSEC) measure is used which is a good alternate to speech intelligibility index (SII) and speech transmission index (STI). Figure 2 shows the percentage intelligibility scores across all the SNR levels and noisy conditions. A significant improvement was reported with second stage in reference to the noisy and speech processed by first stage. Less improvement of first stage in low SNR was reported and the speech intelligibility was remained close to noisy speech. However, at higher SNR levels (0 and 5 dB), the improvements in intelligibility were significant. The percentage improvements in intelligibility with second stage in reference to noisy speech were remarkable in both babble and white noise, (i.e., 21.34% at −5 dB, 19.71% at 0 dB, and 14.79% at 5 dB) and (20.99% at −5 dB, 22.44% at 0 dB and 17.08% at 5 dB). The results in Fig. 2 show that the post-processing stage has remarkably improved speech intelligibility in the low SNR background conditions. The Fig. 3 shows the performance comparison in terms of the speech intelligibility (NSEC) among noisy speech, Spectral Subtraction, Weiner Filtering, first stage and second stage respectively. A remarkable improvement in the NSEC scores was observed with proposed system.

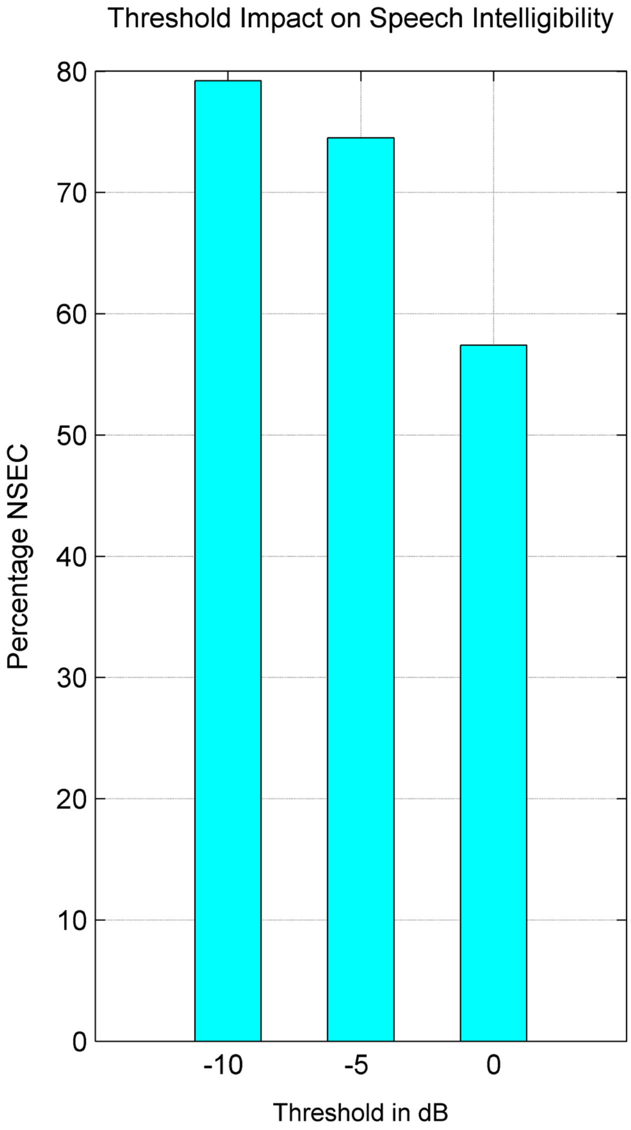


Fig. 6 Impact of threshold on speech intelligibility

6 Selection of threshold

For the best performance of the proposed system, the appropriate selection of threshold T value was mandatory. In a set of experiments, the influence of threshold was examined. The threshold was varied from −10 to 0 dB and the performance is measured in terms of PESQ, $\mathrm{SNR_{SEG}}$, SIG, BAK and NSEC (intelligibility) respectively. Figure 4, 5 and 6 sows the impact of threshold on speech quality ($\mathrm{PESQ} - \mathrm{SNR_{SEG}}$), the residual noise (BAK), speech distortion (SIG) and NSEC scores. In terms of the PESQ, a better performance was obtained when $T = -10$ dB while in terms of $\mathrm{SNR_{SEG}}$, the performance was significant at $T = -5$ dB. Similarly, in terms of SIG and BAK, the appropriate value of T was found to be $T = -10$ dB. A trade-off can be made for the selection of threshold T value for the proposed system. For speech quality, $T = -10$ dB is consistent while in terms of $\mathrm{SNR_{SEG}}$, $T = -5$ dB was a better choice. For that reason, the optimized value of T was varied according to measuring parameters. However, by observing the results, T value must be in between −10 to −5 dB.

7 Spectrogram analysis

In order to yield comprehensive information about residual noise and speech preservation capability of the proposed system, spectrogram analysis was performed. The Fig. 7 shows sample spectrograms for both stages of the proposed system. The speech utterance was degraded by babble noise at 0 dB SNR with PESQ = 1.51. By observing spectrograms of both stages in Fig. 7c, d, the second stage was better able to reduce the background noise and the speech contents were well preserved as compared to the

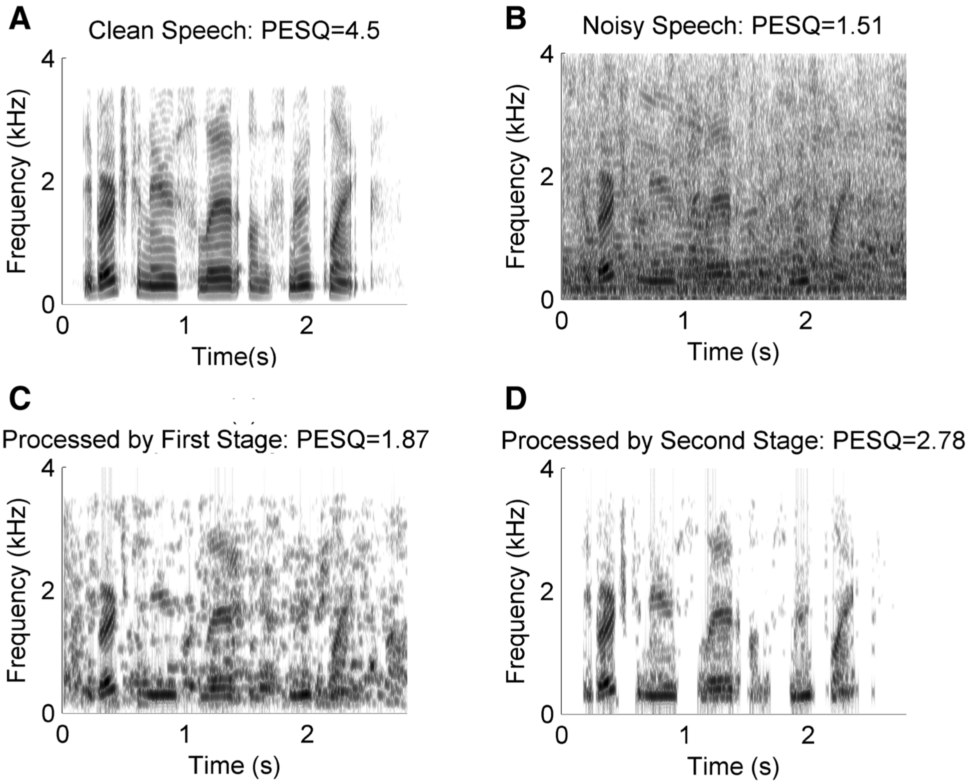


Fig. 7 Spectrogram analysis: **a** clean speech with PESQ = 4.5, **b** noisy speech with PESQ = 1.51, **c** speech processed by first stage with PESQ = 1.87, and **d** speech

first stage and the noisy speech respectively. The proposed noise reduction system performed exceptionally well by eliminating residual noise and also preserved the speech contents efficiently.

8 Summary and conclusion

A two stage noise reduction system for reducing background noise using single-microphone recordings in very low signal-to-noise ratio (SNR) was proposed that is based on Wiener filtering and ideal binary masking. In first stage, the Wiener filtering with improved a priori SNR is applied to noisy speech for background noise reduction while in a post-processing second stage, the ideal binary mask is estimated in every time–frequency channel by using preprocessed first stage speech. The energy in every time–frequency channels was compared to a pre-selected threshold T to reduce the residual the background noise. All the time–frequency channels satisfying the constrained (threshold) were retained whereas all other time–frequency channels were attenuated. The PESQ was used to predict the mean opinion scores (MOS) and composite measure was used to predict the amount of residual noise (BAK) and speech distortion (SIG). All the measuring parameters indicated significant improvements with the proposed noise reduction system. The spectrogram analysis indicated low speech distortion and less residual noise was observed with proposed system. Moreover, significant improvement in speech intelligibility was also reported with the proposed noise reduction system.

References

Abd El-Fattah, M. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, S. M., & Al-Nuaimy, W., et al. (2014). Speech enhancement with an adaptive Wiener filter. *International Journal of Speech Technology, 17*(1), 53–64. doi:10.1007/s10772-013-9205-5.

Boldt, J. B., & Ellis, D. (2009). A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In *Proc. EUSIPCO'09, Glasgow, August 2009* (pp. 1849–1853).

Boldt, J. B., Kjems, U., Pedersen, M. S., Lunner, T., & Wang, D. (2008). Estimation of the ideal binary mask using directional systems. In *Proc. int. workshop acoust. echo and noise control* (pp. 1–4)

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. In *IEEE transactions on acoustics, speech, and signal processing, ASSP* (Vol. 27, pp. 113–120). doi:10.1109/TASSP.1979.1163209.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121. doi:10.1109/TASSP.1984.1164453.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. In *IEEE transactions on acoustics, speech, signal processing, ASSP* (Vol. 23, No. 2, pp. 443–445). doi:10.1109/TASSP.1985.1164550.

Hansen, J., & Pellom, B. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *Interna-*

tional Conference on Spoken Language Processing, 7(2819), 2822.

Hirsch, H., & Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *ISCA ITRW ASR2000, Paris*.

Hu, Y., & Loizou, P. (2007). Subjective evaluation and comparison of speech enhancement algorithms. *Speech Communication, 49*(7–8), 588–601. doi:10.1016/j.specom.2006.12.006.

ITU-T P.835. (2003). Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.

ITU-T Recommendation P.56. (1993). Objective measurement of active speech level.

Klatt, D. (1982). Prediction of perceived phonetic distance from critical band spectra. In *Proc. IEEE int. conf. acoust., speech, signal processing* (Vol. 7, pp. 1278–1281). doi:10.1109/ICASSP.1982.1171512.

Kitawaki, N., Nagabuchi, H., & Itoh, K. (1988). Objective quality evaluation for low bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications, 6*(2), 262–273. doi:10.1109/49.601.

Lim, J, & Oppenheim, A. V. (1978). All-pole modeling of degraded speech. In *IEEE trans. acoust., speech, signal proc., ASSP* (Vol. 26, No. 3, pp. 197–210). doi:10.1109/TASSP.1978.1163086.

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. Boca Raton, FL: CRC Press.

Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America, 126*(23), 1486–1494. doi:10.1121/1.3184603.

Quackenbush, S., Barnwell, T., & Clements, M. (1988). *Objective measures of speech quality*. Eaglewood Cliffs, NJ: Prentice-Hall.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, speech, and signal processing ICASSP*. doi:10.1109/ICASSP.2001.941023.

Saleem, N., Mustafa, E., Nawaz, A., & Khan, A. (2015a). Ideal binary masking for reducing convolutive noise. *International Journal of Speech Technology, 18*(4), 547–554. doi:10.1007/s10772-015-9298-0.

Saleem, N., Shafi, M., Mustafa, E., & Nawaz, A. (2015b). A novel binary mask estimation based on spectral subtraction gain-induced distortions for improved speech intelligibility and quality. *Technical Journal, UET, Taxila, 20*(4), 35–42.

Scalart, P., & Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. In *Proc. IEEE int. conf. acoust., speech, signal processing* (pp. 629–632). doi:10.1109/ICASSP.1996.543199.

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines* (pp. 181–197). doi:10.1007/0-387-22794-6_12.

Wang, D. (2008). Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification, 12*(4), 332–353. doi:10.1177/1084713808326455.