Reduction of Musical Residual Noise Using Hybrid-Mean Filter

Ching-Ta Lu[1], Kun-Fu Tseng[2] and Chih-Tsung Chen[2]

[1]*Department of Information Communication, Asia University, Taiwan, ROC*
[2]*Department of Multimedia and Game Science, Asia-Pacific Institute of Creativity, Taiwan, ROC*
Lucas1@ms26.hinet.net

Abstract

This study proposes a post-processor to reduce the effect of musical residual noise which is annoying to the human ear. Initially, a state-of-the-art speech enhancement algorithm is performed as the first stage to reduce background noise for noisy speech. Hence the enhanced speech is post-processed by a hybrid-mean filter to reduce the musical effect of residual noise. In the case of a vowel-like spectrum, directional-mean filtering is performed to slightly reduce the musical effect of residual noise, where the harmonic spectrum can be well maintained at an acceptable level. Conversely, block-mean filtering is performed to heavily reduce the spectral variation for noise-dominant spectra, enabling musical tones to be significantly smoothed. The musical effect of residual noise is therefore reduced. Finally, the pre-processed, the directional-mean filtered and the block-mean filtered spectra are fused according to speech-presence probability. Experimental results show that the proposed hybrid-mean filter can efficiently improve the performance of a speech enhancement system by reducing the musical effect of residual noise.

Keywords: *speech enhancement, spectral subtraction, musical residual noise, post-processing, hybrid-mean filter*

1. Introduction

Many speech enhancement algorithms have been proposed to reduce the background noise in noisy speech [1-8]. These algorithms attempted to efficiently remove the corruption noise, but the musical effect of residual noise is apparent in the enhanced speech. This musical noise is perceived as twittering and degrades the perceptual quality massively. If it is too prominent, it may be more disturbing than the inference before speech enhancement.

Recently, many studies attempted to suppress musical residual noise [3, 8-10]. Esch and Vary [10] proposed performing smoothing on the weighting gains for speech-pause and low SNR conditions, yielding the musical effect of residual noise being reduced. Jo and Yoo [3] considered a psycho-acoustically constrained and distortion minimized enhancement algorithm. This algorithm minimized speech distortion while the sum of speech distortion and residual noise was kept below the masking threshold.

Based on the above findings, how to efficiently remove the musical effect of residual noise is important for speech enhancement. In this paper, we employ a speech enhancement system to be the first stage for removing background noise; meanwhile, speech distortion should be maintained at a low level. The output signal is further processed by the proposed hybrid-mean (HM) filter which is motivated by that the adaptive median filter can efficiently remove impulse noise in image denoising [11, 12], yielding the musical effect of residual noise being efficiently reduced. An algorithm for estimating speech-presence probability [13] is employed and modified to classify the pre-processed spectrum as speech-dominant or noise-dominant.

In the case of speech-dominant spectrum, an eight-direction mean filter is performed to slightly reduce the musical effect of residual noise; meanwhile, the harmonic spectrum does not been seriously destroyed. When the value of speech-presence probability exceeds a high threshold, the spectrum is classified as a vowel. This spectrum is kept unchanged to maintain speech quality. Conversely, a block-mean filtering is performed to heavily reduce the spectral variation for noise-dominant spectra. Musical tones are then significantly smoothed, enabling the filtered speech to sound less annoying than the pre-processed speech. Finally, the pre-processed, the post-processed spectra (obtained either from directional-mean filter or the block-mean filter) are fused according to the speech-presence probability. If the value of speech-presence probability is high, the weight of pre-processed speech goes high. This enables the pre-processed to be preserved, resulting in less speech distortion in the post-processed speech. Conversely, the weight is high for (block or directional) mean filtered spectra, yielding the musical effect of residual noise being efficiently removed. Experimental results show that the proposed post processor can improve the performance of a speech enhancement system by efficiently removing the musical effect of residual noise, while speech distortion is not perceptible by the human ear. Accordingly, the post-processed speech sounds more comfortable than that without post-processed.

2. Proposed Speech Enhancement System

Initially, noisy speech is framed by a Hanning window, and then transformed into the frequency domain by fast Fourier transform (FFT). A minimum statistics algorithm [14] is employed to estimate the noise magnitude for each subband. Hence, this noise estimate is employed to adapt a speech enhancement system, enabling the background noise to be efficiently removed. Because the musical effect of residual noise is apparent in the pre-processed speech, the hybrid-mean (HM) filter is proposed to remove it. Noisy speech is utilized to estimate the pitch period. Hence, the robust harmonic spectra are searched for each frame. The number of robust harmonic is employed to adapt an algorithm for estimating speech-presence probability which will be applied to control the fusion weighting between the pre-processed and (directional or block) mean filtered signals. Each spectrum of pre-processed speech is analyzed to classify whether it is vowel-like. If the center spectrum of a local window is a vowel, the corresponding speech-presence probability would be large. The center spectrum is kept unchanged to maintain speech quality. If the value of speech-presence probability is less than a given threshold, the center spectrum is classified as vowel-like. A directional mean filter is employed to modify the magnitude of the center spectrum, yielding the musical effect of residual noise being slightly reduced. Conversely, the center spectrum is classified as noise-like when the value of speech-presence probability is equal to zero. A block-mean filtering is performed, enabling the center spectrum to be heavily smoothed. The musical effect of residual noise is then significantly reduced. Finally, the pre-processed, the directional-mean filtered, and the block-mean filtered spectra are fused according to the speech-presence probability. In turn, the inverse FFT is performed to achieve post-processed speech.

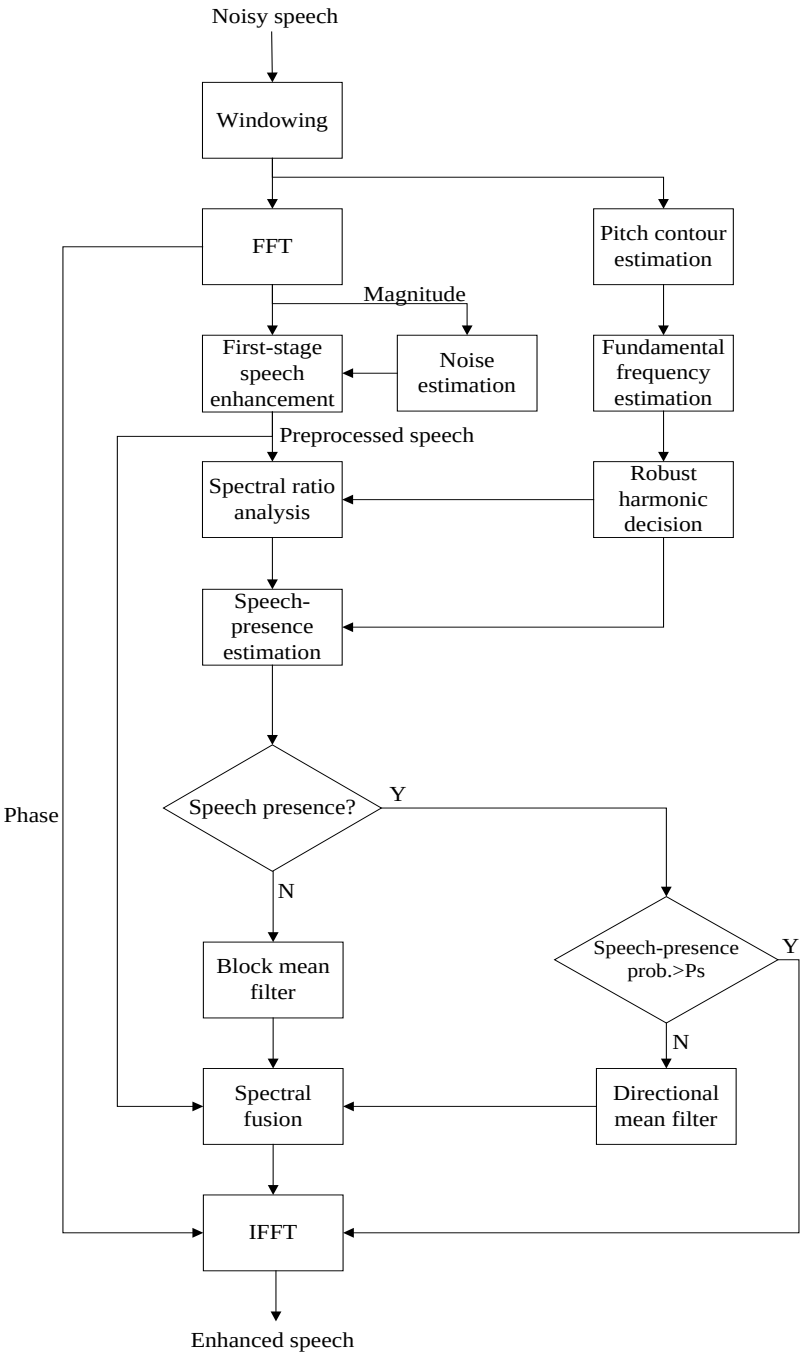


Figure 1. Block Diagram of Proposed Speech Enhancement System

2.1. Robust Harmonic Estimation

A harmonic spectrum distributes in the frequency ranges from 50 to 500 Hz. Low-pass filtering on noisy speech with cut-off frequency 500 Hz is performed to obtain a low-pass signal $\phi(n)$ which can be applied to accurately estimate the pitch period by

reducing the inference of high-frequency signals. In turn, we compute the auto-correlation function of the low-pass filtered signal $R_\phi(\tau)$, given as

$$R_\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}\phi(n)\cdot\phi(n+|\tau|) \tag{1}$$

where *N* denotes frame size.

In order to improve the accuracy for estimating the pitch period, an average magnitude difference function (AMDF)[15] is performed on the low-pass filtered signal $\phi(n)$, given as

$$AMDF(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|}|\phi(n)-\phi(n+|\tau|)| \tag{2}$$

In the position of pitch period, the value of AMDF is small, while the value of $R_\phi(\tau)$ given in (1) is large. The ratio of AMDF and $R_\phi(\tau)$ is enlarged, yielding the discriminability of pitch position increasing. It is beneficial to improve the accuracy in estimating the pitch period. A weighted autocorrelation function (WAC) can be defined as [15]

$$WAC(\tau) = \frac{R_\phi(\tau)}{AMDF(\tau)+\varepsilon} \tag{3}$$

where ε is a very small value to prevent the denominator being zero.

Harmonic estimation can be performed by the fundamental frequency F_0 which can be obtained by the pitch period T_0 [16], given as

$$F_0 = N/T_0 \tag{4}$$

In the experiments, we find that the estimated fundamental frequency obtained by (4) suffers from underestimate. Thus we attempt to shift the location of fundamental frequency F_0 to that of the spectral peak for each segment. The shifted frequency F_0^* can be expressed as

$$F_0^* = \mathrm{F}_0 - F_0^{Bias} \tag{5}$$

where F_0^{Bias} denotes the offset from the fundamental frequency F_0 obtained by (4). It can be computed by

$$F_0^{Bias}(l) = \frac{1}{l_e - l_i}\cdot\sum_{m=l_i}^{l_e-1}F_0(m) - F_0'(m) \tag{6}$$

where l_i and l_e represent the starting and ending frames of the l^{th} segment. $F_0'(m)$ denotes the fundamental frequency with spectral peak.

Robust harmonic takes place on the multiple of fundamental frequencies, i.e., nF_0. The number of robust harmonic *K* can be decided by

$$K = \{k \mid |F_0^k - F_0^{k-1}| \le \delta_{F_0} \text{ and } |F_0^{k+1} - F_0^k| > \delta_{F_0}\} \tag{7}$$

where F_0^k denotes the frequency of k^{th} harmonic. δ_{F_0} is the frequency threshold of adjacent harmonic for deciding robust harmonic.

In (7), if the frequency offset between two adjacent harmonic varies largely, the harmonic structure may become weak. Thus the boundary of robust harmonic can be marked. The more the number of the robust harmonic is, the higher the probability of the speech-presence is. Accordingly, we can employ the number of robust harmonic to adapt an algorithm for estimating speech-presence probability.

2.2. Speech-Presence Probability

Speech presence can be determined by the ratio between the local energy of the noisy speech and its minimum within a specified time window. A speech-presence probability $p(m,\omega)$ can be computed by [13]

$$p(m,\omega) = \alpha_p \cdot p(m-1,\omega) + (1-\alpha_p) \cdot I(m,\omega) \tag{8}$$

where α_p (α_p = 0.2) is a smoothing parameter. $I(m,\omega)$ denotes an indicator function for speech-activity. It can be computed by

$$I(\omega,m) = \begin{cases} 1 & \text{, if } \gamma(m,\omega) > \delta_\gamma(m) \\ 0 & \text{, o.w.} \end{cases} \tag{9}$$

where $\delta_\gamma(m)$ is a speech-presence threshold for a power ratio $\gamma(m,\omega)$ (the ratio between the smoothed local power and the minimum power in a local segment).

In [13], the speech-presence threshold for the power ratio $\delta_\gamma(m)$ is set to a constant 5. Here we modify this threshold by adapting with the number of robust harmonic *K* given in (7). If a frame is vowel-like, the speech indicator $I(m,\omega)$ should approach unity. Thus a weak vowel frame can be classified as speech-presence frame. The ratio $\delta_\gamma(m)$ can be expressed by

$$\delta_\gamma(m) = \delta_{\max} - \frac{\delta_{\max} - \delta_{\min}}{2} \cdot K \tag{10}$$

where $\delta_{\max}$ and $\delta_{\min}$ are empirically chosen to 8 and 3, respectively. In order to prevent the threshold $\delta_\gamma(m)$ from being too small or negative, a lower bound for the threshold $\delta_\gamma(m)$ should be provided, given as $\delta_\gamma(m) = \max\{\delta_\gamma(m), \delta_{\min}\}$.

The value of speech-presence probability lies between 0 and 1 as shown in (8). We can employ it to control the fusion weighting for the pre-processed and the HM filtered spectra.

2.3. Hybrid-mean filter

The hybrid-mean filter is constituted of directional-mean and block-mean filters. Directional-mean filtering is performed when a frame has strong harmonic structure. The direction candidates are shown in Figure 1, where the center spectrum is denoted by a filled circle. The center spectrum is classified as vowel-like when the number of robust harmonic is great enough. In turn, we further check whether the center spectrum is a vowel by the speech-presence probability. If the value of speech-presence probability exceeds a given threshold, the center spectrum is classified as a vowel and kept unchanged to maintain speech quality. On the other hand, if the value of speech-

presence probability lies between 0.2 and 0.8, the center spectrum is classified as vowel-like and filtered by the directional mean filter, given as

$$M(m,\omega)=\begin{cases}\dfrac{a+d+e+h+\tilde{S}(m,\omega)}{5}, \text{if } i^*=1\\ \dfrac{a+b+g+h+\tilde{S}(m,\omega)}{5}, \text{if } i^*=2\\ \dfrac{b+g+\tilde{S}(m,\omega)}{3}, \text{if } i^*=3\\ \dfrac{b+c+f+g+\tilde{S}(m,\omega)}{4}, \text{if } i^*=4\\ \dfrac{c+d+e+f+\tilde{S}(m,\omega)}{5}, \text{if } i^*=5\\ \dfrac{d+e+\tilde{S}(m,\omega)}{3}, \text{if } i^*=6\\ \dfrac{a+h+\tilde{S}(m,\omega)}{3}, \text{if } i^*=7\\ \dfrac{c+f+\tilde{S}(m,\omega)}{3}, \text{if } i^*=8\end{cases} \tag{11}$$

where i^* denotes the optimum direction. $\tilde{S}(m,\omega)$ represents the pre-processed spectrum. The symbols *a-f* represent the pre-processed spectrum.

As shown in Figure 1, the optimum motion direction of the center spectrum should be selected among the eight candidate directions (1-8). The decision rule is to select the direction with the minimum spectral-distance. The spectral-distance measure $d^{(i)}(m,\omega)$ can be expressed by (12)-(19), given as [12]

$$d^{(1)}(m,\omega)=|d-h|+|a-e| \tag{12}$$

$$d^{(2)}(m,\omega)=|a-g|+|b-h| \tag{13}$$

$$d^{(3)}(m,\omega)=2\cdot|b-g| \tag{14}$$

$$d^{(4)}(m,\omega)=|b-f|+|c-g| \tag{15}$$

$$d^{(5)}(m,\omega)=|c-d|+|e-f| \tag{16}$$

$$d^{(6)}(m,\omega)=2\cdot|d-e| \tag{17}$$

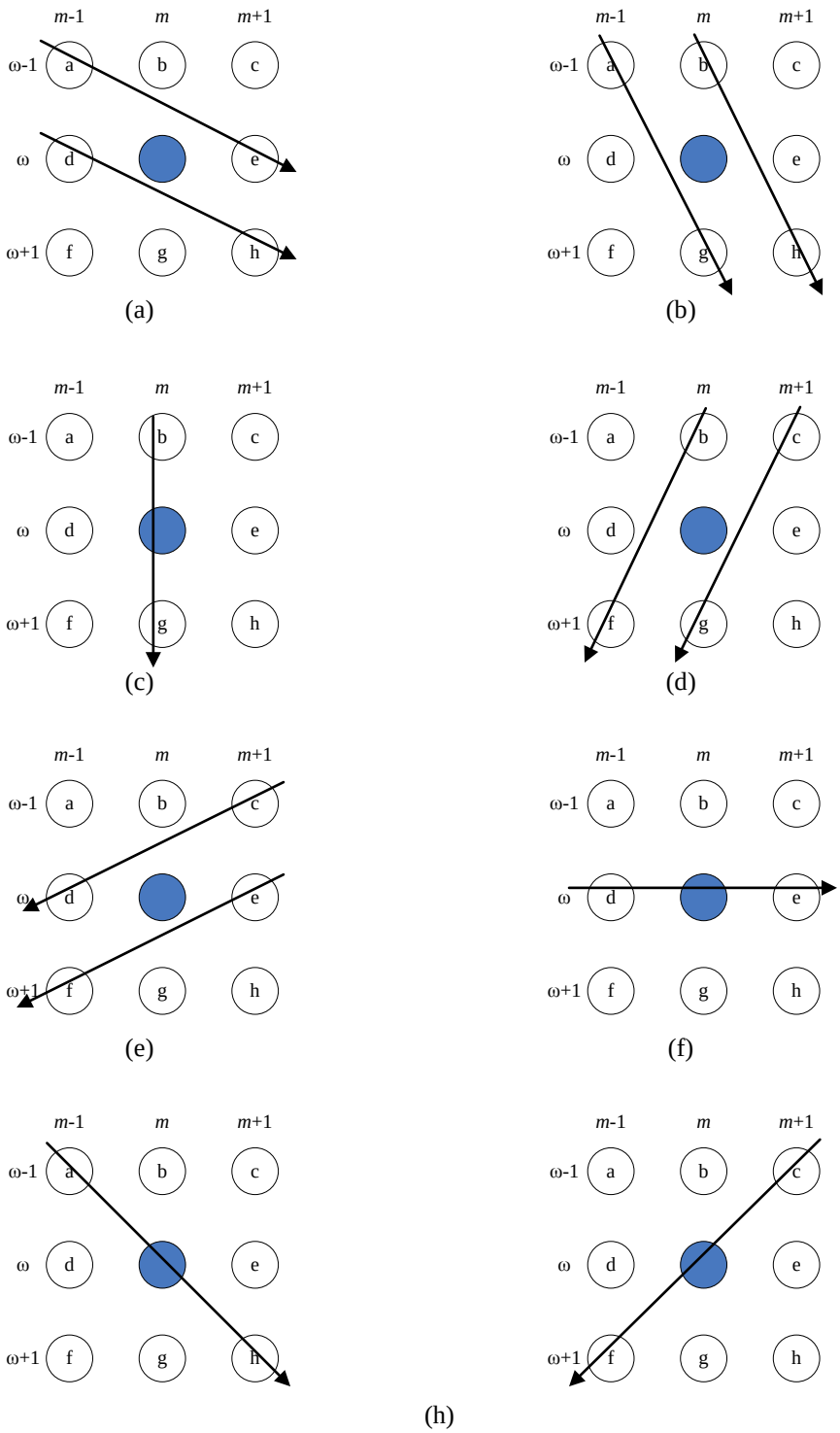


Figure 2. Motion Directions of the Center Spectrum, Directions (a) i^*=1, (b) i^*=2, (c) i^*=3, (d) i^*=4, (e) i^*=5, (f) i^*=6, (g) i^*=7, (h) i^*=8

$$d^{(7)}(m,\omega) = 2 \cdot |a - h| \tag{18}$$

$$d^{(8)}(m,\omega) = 2 \cdot |c - f| \tag{19}$$

The direction with the minimum spectral-distance given in (12)-(19) is declared as the optimum motion direction for the center spectrum. The optimum distance measure can be expressed as

$$d^{(i^*)}(m,\omega) = \min\left\{ d^{(i)}(m,\omega) \quad , 1 \le i \le 8 \right\} \tag{20}$$

The directional-mean filter can mitigate the fluctuation of random spectral peaks on the optimum direction, enabling the musical effect of residual noise to be reduced. In order to improve the performance in the reduction of musical tones, we further employ a block-mean filter to significantly smooth the variation of musical tones when a center spectrum is classified as noise-like. The larger the size of the window is, the greater the reduction of the spectral variation is. Increasing window size causes a quantity of speech distortion. Accordingly, we adopt the window size 3×3 to analyze and filter the pre-processed spectra.

3. Experimental Results

In the experiments, speech signals are Mandarin Chinese spoken by five female and five male speakers. Those speech signals are corrupted by various kinds of additive noise, such as white, F16-cockpit, factory, and helicopter-cockpit noise signals which are extracted from the Noisex-92 database. Three input average segmental SNR levels, including 0, 5, and 10dBs, are used to evaluate the performance of a speech enhancement system. A minimum statistics algorithm [14] is performed to estimate the power of noise for each frequency bin. This algorithm updates the noise estimate in both speech-activity and speech-pause regions, which fact represents the advantage of the minimum statistics approach. The following parameters are used in the experiments: (1) sampling frequency is 8 kHz; (2) the frame size is 256 with 50% overlap; (3) Hanning window is utilized; (4) total number of critical bands is 18, the center frequency and the corresponding bandwidth of each critical band can be found in [1].

Objective measures, including the average of segmental SNR improvement (Avg_SegSNR_Imp) and the perceptual evaluation of speech quality (PESQ) [17] are conducted to evaluate the performance of a speech enhancement system. Only the performance in speech-activity regions is evaluated. Speech spectrogram comparison is also performed. In order to evaluate the performance of the proposed system, a two-step-decision-directed algorithm [5] and the Virag method [1] are implemented as the first stage for comparisons. The proposed hybrid-mean filter is cascaded after the first stage to improve the performance by the more reduction of musical residual noise.

3.1. Noise Estimate

Noise estimator has been a major role on deciding the quality of a speech enhancement system. If the noise estimate is too low, residual noise increases. Conversely, if the level of noise estimate is too high, enhanced speech sounds would be muffled and intelligibility would be lost. The traditional voice activity detectors (VADs) are difficult to tune in non-stationary noise corruption. In addition, the voice activity detector (VAD) application to low SNR speech results often in clipped speech. Thus, the VAD cannot well estimate the noise level in non-stationary and low SNR environments.

Martin [14] proposed the minimum statistics algorithm to estimate the power of noise for each subband. The algorithm does not use the VAD, instead it tracks power minimum in each subband to decide the noise estimate. The minimum statistics noise tracking method is based on the observation that even during speech activity a short-term power density estimate of the noisy signal frequently decays to values which are representative of the noise level. This method rests on the fundamental assumption that during speech pause or within brief periods in between words and syllables, the speech energy is close or identical to zero. Thus, by tracking the minimum power within a finite window large enough to bridge high power speech segments, the noise floor can be estimated. Detailed procedure of the minimum statistics noise estimation algorithm can be found in [14].

Table 1. Comparison of SegSNR Improvement for the Enhanced Speech in Various Noise Corruption

noise type	SNR (dB)	Average SegSNR improvement			
		TSDD	TSDD +Post	Virag	Virag +Post
White	0	6.82	6.97	6.38	6.99
	5	4.79	4.92	4.90	5.43
	10	3.04	3.20	3.48	3.93
F16	0	4.99	5.10	5.09	5.35
	5	3.52	3.75	3.66	4.24
	10	2.32	2.58	2.39	3.18
Factory	0	4.71	4.83	4.64	5.01
	5	3.37	3.55	3.20	3.99
	10	2.23	2.48	1.97	3.00
Helicopter	0	6.75	7.21	6.44	7.12
	5	4.87	5.44	4.70	5.68
	10	3.24	3.87	3.19	4.28

3.2. Segmental SNR Improvement

The quantities of noise reduction, residual noise and speech distortion can be measured by the average segmental SNR improvement (Avg_SegSNR_Imp). The average of segmental SNR (Avg_SegSNR) of a test signal is evaluated according to clean speech $s(m,n)$, and the enhanced signal $\hat{s}(m,n)$. It can be expressed by

$$Avg\_\text{SegSNR} = \frac{1}{M}\sum_{m\in\{I\}} 10\cdot\log_{10}\left(\frac{\sum_{n=0}^{N-1}|s(m,n)|^2}{\sum_{n=0}^{N-1}|s(m,n)-\hat{s}(m,n)|^2}\right) \quad (21)$$

where $\{I\}$ represents a set of speech-activity frames. M and N denote the numbers of speech-activity frames and of samples per frame, respectively. m is frame index.

The Avg_SegSNR_Imp is computed by subtracting the Avg_SegSNR of noisy speech from that of enhanced speech. Table 1 presents the performance comparisons in terms of the average segmental SNR improvement. The larger the value of the Avg_SegSNR_Imp is, the better the quality of enhanced speech is. Observing the performance presented in the TSDD and the TSDD+Post, to cascade the proposed hybrid-mean filter after the TSDD method (TSDD+Post) can improve the performance of the pre-processed speech (TSDD). The performance of the Virag method (Virag) can be improved by the hybrid-mean filter (Virag+Post), too. Accordingly, the proposed hybrid-mean filter can improve the performance for a speech enhancement system in various noise corruptions. The major reason is attributed to the fact that the proposed method can remove the musical effect of residual noise; meanwhile, the harmonic structure of a vowel speech is well preserved.

3.3. Perceptual Evaluation of Speech Quality

The perceptual evaluation of speech quality (PESQ) measure, which has better correlation with subjective tests than the other objective measures, was selected as the ITU-T recommendation P.862 [17] to evaluate the speech quality of a test signal. In the computation of PESQ score for an enhanced speech signal (or a noisy speech signal), the clean and enhanced speech signals were initially level-equalized to a standard listening level, and then filtered by a filter with response similar to a standard telephone handset. The clean and enhanced speech signals were aligned in the time domain to correct the time delays between these two signals. Hence, these two signals were processed through an auditory transform, similar to that of perceptual speech quality measure (PSQM) to obtain the loudness spectra. The disturbance, obtained by computing the difference between the loudness spectra for the clean and the enhanced speech signals, was computed and averaged over time and frequency to produce the prediction of subjective mean opinion score. The detailed procedures for computing the PESQ score can be found in [17].

Table 2 presents the performance comparisons in terms of the PESQ. The maximal PESQ score corresponds to the best speech quality. We can find that the proposed hybrid-mean filter obtains higher PESQ scores than the TSDD and the Virag methods. It shows that the proposed hybrid-mean filter does not seriously deteriorate speech components while efficiently suppressing the musical effect of residual noise. These results are consistent with that in terms of average segmental SNR improvement shown in Table 1.

3.4. Waveforms

Figure 3 demonstrates an example of waveform plots for comparison. A speech signal uttered by a female speaker was corrupted by helicopter-cockpit noise with Avg_SegSNR = 0 dB. In Figure 3(c), the TSDD method can efficiently remove background noise, in particular in a speech-pause region. The quantity of residual noise of the proposed method (Figure 3(d)) is comparable to that of the TSDD method (Figure 3(c)). It is due to the proposed method aims at smoothing musical tones over successive frames and neighbor subbands, rather than to further suppress the magnitude of musical tones. Therefore, the proposed method will not cause additional speech deterioration when conducting hybrid-mean filtering.

Table 2. Comparisons of Perceptual Evaluation of Speech Quality (PESQ) for the Enhanced Speech in Various Noises

Noise type	SNR (dB)	PESQ Noisy	TSDD	TSDD +Post	Virag	Virag +Post
White	0	1.63	2.05	2.12	2.07	2.20
	5	1.91	2.36	2.43	2.45	2.57
	10	2.25	2.65	2.72	2.80	2.92
F16	0	1.84	2.18	2.24	2.29	2.34
	5	2.18	2.51	2.56	2.63	2.70
	10	2.53	2.81	2.86	2.97	3.03
Factory	0	1.81	1.97	2.06	2.21	2.21
	5	2.16	2.37	2.43	2.58	2.59
	10	2.52	2.71	2.78	2.93	2.95
Helicopter	0	2.02	2.43	2.52	2.55	2.67
	5	2.37	2.75	2.83	2.88	3.01
	10	2.73	3.05	3.11	3.16	3.30

Comparing the waveform plots of enhanced speech shown in Figures 3(e) and (f), the proposed hybrid-mean filter can slightly improve the Virag method by reducing the musical effect of residual noise during the speech-pause regions. A vowel with weak energy can be restored. It may be attributed to the adaptation of harmonic on the directional-mean filter, yielding a vowel with weak energy being restored by the neighbor vowel with stronger energy. Therefore, the proposed hybrid-mean filter does not suffer from the deterioration of speech when reducing the effect of residual noise.

3.5. Spectrograms

Objective measures cannot easily quantify the quantity of residual noise in the enhanced speech. Analyzing the time-frequency distribution of the enhanced speech and evaluating the structure of residual noise, are particularly important. The speech spectrograms are therefore observed, to yield more information about the residual noise and speech distortion.

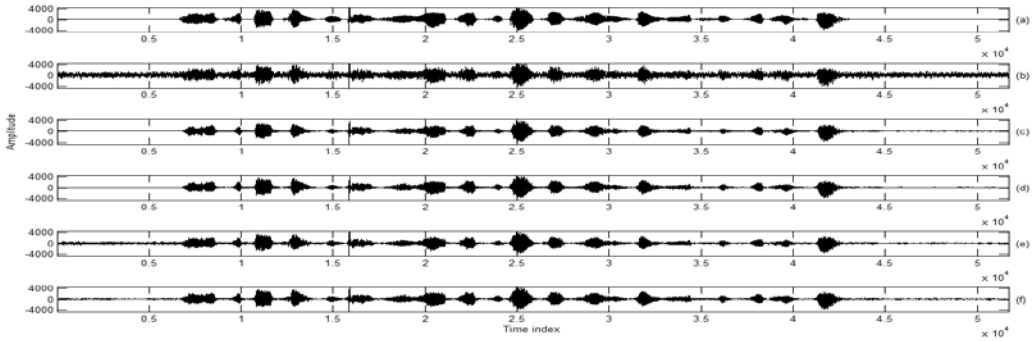
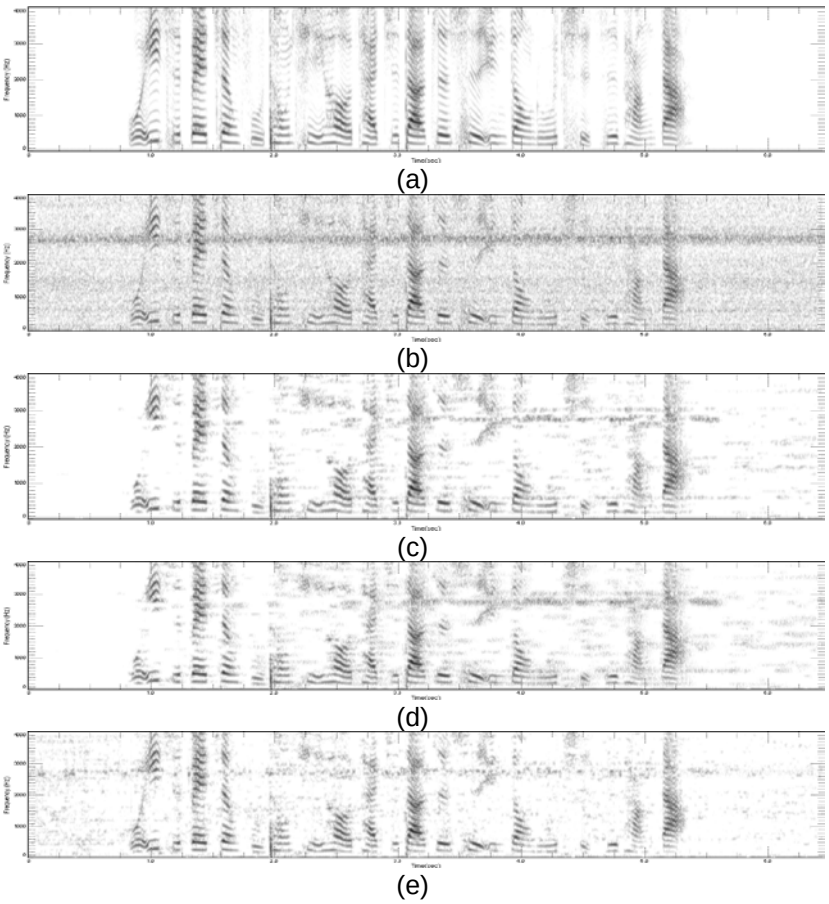


Figure 3. Example of Speech Signal Spoken by a Female Speaker. (a) Clean Speech, (b) Noisy Speech Corrupted by Helicopter-cockpit Noise with Average SegSNR = 0 dB, Enhanced Speech using (c) TSDD Method, (d) TSDD Method with Post Processing, (e) Virag Method, (f) Virag Method with Post Processing

Figure 4 presents the spectrogram comparisons for various speech enhancement methods. Speech signals were corrupted by F16-cockpit noise signal with Avg_SegSNR = 5 dB. Observing the spectrograms of enhanced speech during speech-pause regions shown in Figure 4(c), plenty of isolated spectral patches with strong energy exist in the enhanced speech for the TSDD method. After post-processed by the proposed hybrid-mean filter, these isolated patches tend to whiten (Figure 4(d)), *i.e.*, the isolated spectral patches of musical tones spread to the neighbor subbands and frames. This enables the musical tones to vary smooth in the successive frames and to sound less annoying than that produced by the TSDD method (Figure 4(c)). The musical effect of residual noise is then reduced. In Figure 4(e), there is a quantity of residual noise in the enhanced speech of Virag method. This noise is very annoying to the human ear and can be efficiently removed by the proposed hybrid-mean filter (Figure 4(f)). The major reason is attributed to residual noise being efficiently smoothed by the block-mean filter, enabling the isolated random spectral peaks to vary smoothly over successive frames and neighbor subbands. Accordingly, the musical effect of residual noise is efficiently reduced, while the harmonic structure of a vowel is well preserved. In addition, a muffled effect is absent in the hybrid-mean filtered speech. This results in the post-processed speech sounding more comfortable than the pre-processed one.

Based on the above discussion, the hybrid-mean filter can efficiently reduce the musical effect of residual noise and can adequately preserve the harmonic spectra of a vowel, yielding the post-processed speech sounding more comfortable than that without post-processing for the TSDD and the Virag methods. Accordingly, the proposed hybrid-mean filter can improve the performance of a speech enhancement system.



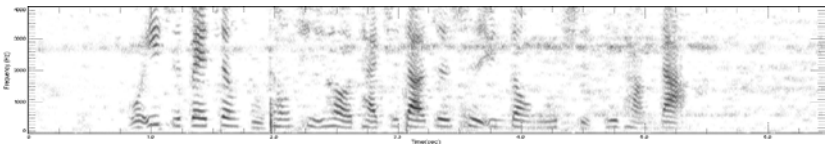
(a)

(b)

(c)

(d)

(e)



(f)

Figure 4. Spectrograms of Speech Spoken by a Female Speaker, (a) Clean Speech, (b) Noisy Speech Corrupted by F16-cockpit Noise with Average SegSNR = 5 dB, Enhanced Speech using (c) TSDD Method, (d) TSDD Method with Post Processing, (e) Virag Method, (f) Virag Method with Post Processing

4. Conclusions

Employing the hybrid-mean filter to post-process enhanced speech was proposed in this study. The major contribution is to significantly reduce the spectral variation of residual noise by block-mean filtering on the spectrogram of a noise-dominant region, and to slightly smooth the spectra of residual noise by directional-mean filtering in a speech-dominant region. Hence, the pre-processed and the hybrid (block or directional) mean filtered spectra are adequately fused according to speech-presence probability. It prevents the spectra in speech-dominant regions from being severely deteriorated by the proposed hybrid-mean filter. Experimental results show that the proposed post-processor can efficiently reduce the musical effect of residual noise for a speech enhancement system, yielding the post-processed speech sounding more comfortable than that without post-processing.

Acknowledgements

This research was supported by the National Science Council, Taiwan, under contract number NSC 101-2221-E-468-010.

References

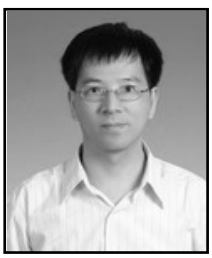
[1] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", IEEE Trans. Speech Audio Process, vol. 7, no. 2, **(1999)**, pp. 126-137.
[2] C. -T. Lu, "Enhancement of Single Channel Speech Using Perceptual-Decision-Directed Approach", Speech Commun., vol. 53, no. 4, **(2011)**, pp. 495-507.
[3] S. Jo and C. D. Yoo, "Psychoacoustically Constrained and Distortion Minimized Speech Enhancement", IEEE Trans. Audio Speech, Language Process, vol. 18, no. 8, **(2010)**, pp. 2099-2110.
[4] J. Ding, I. Y. Soon and C. K. Yeo, "Over-Attenuated Components Regeneration for Speech Enhancement", IEEE Trans. Audio Speech Language Process, vol. 18, no. 8, **(2010)**, pp. 2004-2014.
[5] C. Plapous, C. Marro and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Trans. Audio Speech Languge Process, vol. 14, no. 6, **(2006)**, pp. 2098-2108.
[6] V. J. Naveen, T. Prabakar, J. V. Suman and P. D. Pradeep, "Noise Suppression in Speech Signals Using Adaptive Algorithms", Int. J. Signal Process, Image Process and Pattern Recog., vol. 3, no. 3, **(2009)**, pp. 87-96.
[7] M. E. Hamid, S. Das, K. Hirose and M. K. I. Molla, "Speech Enhancement Using EMD Based Adaptive Soft-Thresholding (EMD-ADT)", Int. J. Signal Process, Image Process and Pattern Recog., vol. 5, no. 2, **(2012)**, pp. 1-16.
[8] C.-T. Lu and K.-F. Tseng, "A Gain Factor Adapted by Masking Property and SNR Variation for Speech Enhancement in Colored-Noise Corruptions", Computer Speech Language, vol. 24, no. 4, **(2010)**, pp. 632-647.
[9] C.-T. Lu, "Reduction of Musical Residual Noise for Speech Enhancement Using Masking Properties and Optimal Smoothing", Pattern. Recog. Lett., vol. 28, **(2007)**, pp. 1300-1306.
[10] T. Esch and P. Vary, "Efficient Musical Noise Suppression for Speech Enhancement Systems", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan, **(2009)** April 19-24, pp. 4409-4412.

[11] H. S. Yazdi and F. Homayouni, "Impulsive Noise Suppression of Images Using Adaptive Median Filter", Int. J. Signal Process, Image Process and Pattern Recog., vol. 3, no. 3, **(2009)**, pp. 1-12.
[12] P.-Y. Chen and C.-Y. Lien, "An Efficient Edge-Preserving Algorithm for Removal of Salt-and-Pepper Noise", IEEE Signal Process Lett., vol. 15, **(2008)**, pp. 833-836.
[13] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement", IEEE Signal Process Lett., vol. 9, no. 1, **(2002)**, pp. 12-15.
[14] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Trans. Speech Audio Process, vol. 9, no. 5, **(2001)**, pp. 504-512.
[15] T. Shimanura and H. Kobayashi, "Weighted Auto-Correlation for Pitch Extraction of Noisy Speech", IEEE Trans. Speech Audio Process, vol. 9, no. 7, **(2001)**, pp. 727-730.
[16] R. Cai, "An Automatic Pitch Detection Method Based on Multi-feature for Mandarin Speech", Int. J. Signal Process, Image Process and Pattern Recog., vol. 5, no. 4, **(2012)**, pp. 155-165.
[17] ITU-T. ITU-T P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs", Int. Telecommun, Union, series P, **(2001)**.

Authors

Ching-Ta Lu received both B.S. and M.S. degrees in Electronic Engineering from National Taiwan University of Science and Technology, Taipei, in 1991 and 1995, respectively; and his Ph. D. degree in Electrical Engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 2006. He had been with the Department of Electronic Engineering, Asia-Pacific Institute of Creativity, Miao Li, Taiwan (Aug. 1995- Feb. 2008). He had been the chair of the department in 2000 and 2006. Dr. Lu received the excellent teaching awards in 2006 and 2007, and the best tutor awards in 1986, 1987, 2005, 2006, 2007, in the Asia Pacific Institute of Creativity. He also received the best tutor awards in 2009 and 2012 in the Asia University. Currently, he is an associate professor of the Department of Information Communication, Asia University (since Feb. 2008). His current research interests include speech enhancement, image denoising, speech coding, and speech signal processing.

Kun-Fu Tseng received his Ph. D. degree in electronic engineering from National Defense University, Taiwan, in 1997. He then joined the faculty of the Electronic Engineering Department at Asia-pacific Institute of Creativity, Taiwan, and the department was reorganized and currently named as the Department of Multimedia and Game Science. His research areas focus on speech enhancement and high frequency analysis and simulation of IC package.



Chih-Tsung Chen received B.S. degree in Electronic Engineering from National Taiwan University of Science and Technology, Taipei, in 1989, respectively; and his M.S. degree in Electrical Engineering from Nation Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C., in 1993. Currently, he is a lecturer of the Department of Electronic Engineering, Asia-Pacific Institute of Creativity, Miaoli, Taiwan (since Aug. 1993). His current research interests include speech enhancement, image denoising, design of games, and Android Apps.