Multimedia Tools and Applications (2024) 83:45423–45441
https://doi.org/10.1007/s11042-023-17336-z

Iterative-processed multiband speech enhancement for suppressing musical sounds

Navneet Upadhyay[1]

Received: 7 March 2022 / Revised: 20 October 2022 / Accepted: 29 September 2023 /
Published online: 21 October 2023

Abstract
A multiband spectral subtraction (MBSS) processing step transforms background noise into annoying musical sounds. The paper proposes an iterative-processed multiband speech enhancement (IP-MBSE) post-processing method for suppressing musical sounds in enhanced speech recordings. In the proposed technique, the outturn of the MBSS processing is employed as an input for the subsequent iteration. The noise spectrum is estimated in each iteration, and the spectral subtraction is executed in each subband individually. The proposed method reduces musical sound even further by applying the estimated speech to the input and repeating the process. This procedure is repeated only a few times. The performance of the proposed technique, IP-MBSE, is measured using: (i) objective clarity measurements such as signal to noise ratio (SNR), segmental SNR (SegSNR), and perceptual evaluation of speech quality (PESQ), as well as (ii) subjective clarity metrics such as mean opinion score (MOS) and spectrogram at various SNR levels. The results of the IP-MBSE are compared with the conventional MBSS, and it is found that the IP-MBSE estimated speech is more pleasant for auditors.

Keywords Iteration number · Musical sound · Over-subtraction of spectral data · Subband · Speech enhancement

1 Introduction

In sectors, for example, speech recognition and speaker identification, speech is the notable and important kind of contact between humans and human to computer [1]. Due to numerous sorts of interferences, today's speech communication technologies are severely harmed, making direct listening difficult and causing inaccurate information transfer [2]. As a result, to achieve nearly transparent communication in applications like cell phones, one of the primary research undertakings in the speech processing field during the last few decades has been the enhancement of degraded speech. Speech enhancement's major purpose is to

✉ Navneet Upadhyay
nupadhyay@lnmiit.ac.in

[1] Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur 302 031, India

lower the distortions and to boost the perceptual characteristics of speech, such as clarity and intelligibility [3, 4]. In some cases, these two characteristics, clarity, and intelligibility, are unrelated. A speaker's exceptionally clear speech in a foreign language, for example, can be of great value to an auditor but has zero intelligibility. As a result, high-clarity speech could have little intelligibility, whereas low-clarity speech might have a lot of it [5].

The number of microphones used to collect speech data, which might be single, dual, or multiple, is used to classify speech enhancement systems. Even though multi-microphone speech enhancement outperforms single-microphone in terms of noise reduction [1], single-microphone speech enhancement remains a prevalent study theme due to its simplicity in design and processing. The single-microphone speech enhancement takes noisy data with only one microphone and does not provide any extra knowledge about the degradation or clear speech. [6] demonstrates that for speech clarity and intelligibility, short-term spectral magnitude (STSM) is more significant than phase shift.

Boll [7] pioneered spectral subtraction, an extensive single-microphone speech enhancement technique relying on the computation of STSM. The spectral subtraction method's main advantages are i) because of its ease (only noise spectrum estimation is required), and ii) its adaptability when it comes to changing the subtraction parameter. Spectral subtraction, despite its capacity to reduce background degradation, inserts musical sound into enhanced speech. The musical sound perception as twittering degrades the perceptual clarity of speech recordings. It may even be more disturbing than the interference before enhancement if it is too prominent.

In speech enhancement, the presence of musical sound is a key issue. Several speech enhancement approaches have been developed in previous decades to address improvements to the classical spectral subtraction method for counteracting musical sounds and enhancing speech clarity in noisy environments [8, 9]. To make musical sounds inaudible, over-subtraction and spectral flooring were recommended in [8]. [9] recommends employing a multiband model in the frequency domain to improve speech.

This study investigates the use of iterative-processed multiband speech enhancement (IP-MBSE) as a post-processing approach for musical sound suppression in enhanced speech recordings. The additive background degradation is converted to an unpleasant musical sound using a multiband spectral subtraction (MBSS) step. The MBSS processing step's outturn is used as the input for the following iteration in IP-MBSE. The musical sound is estimated again in every repetition, and the over-subtraction of spectral data is performed individually in each subband. This method is repeated only a few times. A tradeoff among the degradation level suppression, distorting speech, and musical sounds distinguishes the improved speech.

The rest of the paper is arranged as follows: The fundamentals of spectral subtraction [7], spectral over-subtraction (SOS) [8], and multiband spectral subtraction (MBSS) [9] are covered in Section 2. The proposed method for musical sound suppression, iterative-processing-based multiband speech enhancement, is described in Section 3 (IP-MBSE). The performance evaluation and experimental findings are presented in Section 4. The conclusion is addressed in Section 5.

2 The fundamentals of spectral subtraction

Spectral subtraction is a cost-effective method for successfully removing degradation from degraded speech. Boll [7] proposed the spectral subtraction technique, which can be utilized for speech enhancement and recognition.

In real-world conditions, additive noise degrades the speech signal [3, 7]. Background degradation is unrelated to clear speech and is known as additive noise. Degradations can be either stationary (for instance, white Gaussian) or non-stationary (for instance, colored). The speech signal that has been degraded by WGN is referred to as "noisy speech". The sum of clear speech and degradation can be used to represent the noisy signal mathematically [3, 7] as

$$y[n] = s[n] + d[n], \tag{1}$$

$y[n]$, $s[n]$, and $d[n]$ are the n^{th} samples of noisy speech, clean speech, and background degradation, respectively. Because the speech signal is non-stationary, it is usually broken into short-length frames for subsequent processing to render them stationary over time using the short-term Fourier transform (STFT). Equation (1) may now be expressed as [6, 7], with $Y_W(\omega)$, $D_W(\omega)$, and $S_W(\omega)$ denoting the STFT of the signals.

$$Y_W(\omega) = S_W(\omega) + D_W(\omega) \tag{2}$$

There are two segments of the spectral subtraction method. The noisy speech spectrum is subtracted from an average noise spectrum estimate in the first segment. This is referred to as the elementary subtraction step. To reduce the signal level in the silent zones, numerous changes are made in the second segment, including half-wave rectification (HWR), musical sound lessening, and speech distortion. Because phase distortion is not noticed by the human ear, the phase of noisy speech is kept constant throughout the process [6]. As a result, noisy speech's short-term spectral magnitude (STSM) is equal to the sum of clean speech's STSM and noise's STSM with a lack of phase shift information, and (2) can be represented as

$$\left|Y_W(\omega)\right| = \left|S_W(\omega)\right| + \left|D_W(\omega)\right| \tag{3}$$

Here

$$Y_w(\omega) = \left|Y_w(\omega)\right| \exp\left(j\varphi_y(\omega)\right),$$

$$S_w(\omega) = \left|S_w(\omega)\right| \exp\left(j\varphi_y(\omega)\right),$$

$D_w(\omega) = |D_w(\omega)|\exp(j\varphi_y(\omega))$ and $\varphi_y(\omega)$ is the phase-shift of the noisy signal. The spectrum of noisy speech is obtained by the product of $Y_w(\omega)$ by its conjugate $Y_w^*(\omega)$. As a result, (2) become

$$\left|Y_w(\omega)\right|^2 = \left|S_w(\omega)\right|^2 + \left|D_w(\omega)\right|^2 + S_w(\omega)D_w^*(\omega) + S_w^*(\omega)D_w(\omega) \tag{4}$$

$D_w^*(\omega)$ and $S_w^*(\omega)$ are the conjugates of $D_w(\omega)$ and $S_w(\omega)$. The noisy spectrum, clean speech spectrum, and noise spectrum are denoted by $|Y_w(\omega)|^2$, $|S_w(\omega)|^2$, and $|D_w(\omega)|^2$, respectively. $|D_w(\omega)|^2$, $S_w(\omega)D_w^*(\omega)$ and $S_w^*(\omega)D_w(\omega)$cannot be obtained directly in (4), thus they are approximated as, $E\{|D_w(\omega)|^2\}$, $E\left\{S_w(\omega)D_w^*(\omega)\right\}$ and $E\left\{S_w^*(\omega)D_w(\omega)\right\}$, here $E\{.\}$ is the operator of ensemble averaging. The terms $E\left\{S_w(\omega)D_w^*(\omega)\right\}$ and $E\left\{S_w^*(\omega)D_w(\omega)\right\}$ fall to zero when the additive noise is regarded as zero-mean and orthogonal to speech [3]. As a result, (4) can be rephrased as

$$\left|\hat{S}_w(\omega)\right|^2 = \left|Y_w(\omega)\right|^2 - E\left\{\left|D_w(\omega)\right|^2\right\} = \left|Y_w(\omega)\right|^2 - \left|\hat{D}_w(\omega)\right|^2 \tag{5}$$

where $\left|\hat{S}_w(\omega)\right|^2$ and $|Y_w(\omega)|^2$ are the processed and the noisy speech short-term power spectrums, respectively. The average noise power, $\left|\hat{D}_w(\omega)\right|^2$, is calculated and adjusted during speech interruptions using voice activity detector (VAD) [7].

$$\left|\hat{D}_w(\omega)\right|^2 = \frac{1}{M}\sum_{i=0}^{M-1}\left|\hat{Y}_{SP}(\omega)\right|^2 \tag{6}$$

M denotes speech pauses number in consecutive frames.

The spectral subtraction method assumes that the speech signal has been corrupted by additive white Gaussian noise (WGN) with a flat spectrum, meaning that the degradation has affected the signal evenly across the spectrum. The subtraction step in this procedure must be done with caution to minimize speech distortion. Due to an erroneous estimation of the noise spectrum, the spectra obtained after the subtraction operation may have some negative values. Half-wave rectification (HWR, setting the negative regions to zero) or full-wave rectification (FWR, absolute value) are utilized because the spectrum of estimated speech can grow negative but not be negative. HWR is widely used, but it introduces distracting sounds into the estimated speech. FWR prevents the production of irritating sounds, but it is less effective at degradation suppression. As a result, the equation for spectral subtraction is given by

$$\left|\hat{S}_w(\omega)\right|^2 = \begin{cases} \left[\left|Y_w(\omega)\right|^2 - \left|\hat{D}_w(\omega)\right|^2\right] & \text{if } \left|Y_w(\omega)\right|^2 > \left|\hat{D}_w(\omega)\right|^2 \\ 0 & \text{else} \end{cases} \tag{7}$$

Because human perception is phase insensitive [6], the improved speech spectrum may be produced using the phase of the degraded speech, and the estimated speech can be reconstructed using the inverse STFT (ISTFT) of the enhanced spectrum using the phase of the degraded speech and the overlaps-add (OLA) approach, which can be represented as

$$\hat{s}_w[n] = \text{ISTFT}\left\{\left|\hat{S}_w(\omega)\right| \exp\left(j\varphi_y(\omega)\right)\right\} \tag{8}$$

The disadvantage of spectral subtraction is that it makes the enhancing procedure more complicated. According to (5), spectral subtraction's efficacy is strongly dependent on good noise estimation, which is further constrained by speech/pause detector performance. Musical sound and speech distortion are two primary challenges that develop when the noise estimate is not correct. The spectral over-subtraction of Berouti [8] is a variation of magnitude spectral subtraction [7].

2.1 Spectral over-subtraction (SOS)

To lessen musical sound and distortion, [8] presents a modified spectral subtraction. An over-subtraction factor and the noise spectral floor parameter are used in addition to the spectral subtraction [7] in this method [8]. The steps are as follows:

$$\left|\hat{S}_w(\omega)\right|^2 = \begin{cases} \left|Y_w(\omega)\right|^2 - \alpha\left|\hat{D}_w(\omega)\right|^2 & \text{if } \frac{\left|\hat{D}_w(\omega)\right|^2}{\left|Y_w(\omega)\right|^2} < \frac{1}{\alpha+\beta} \\ \beta\left|\hat{D}_w(\omega)\right|^2 & \text{else} \end{cases} \tag{9}$$

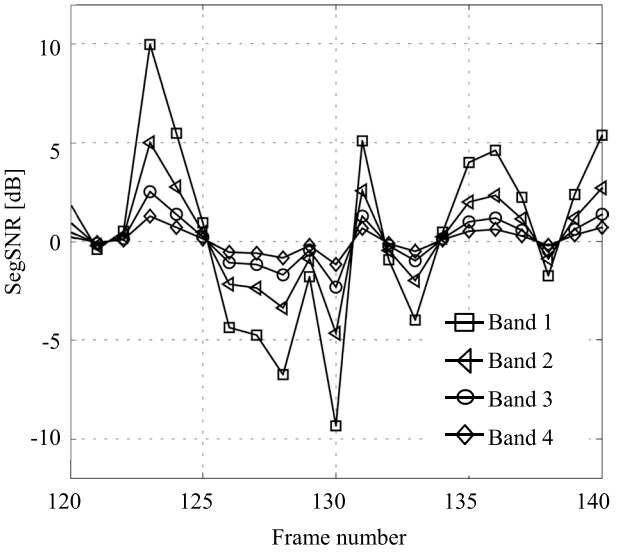

Fig. 1. SegSNR of four linearly spaced frequency subbands of degraded speech

with $\alpha \geq 1$ and $0 \leq \beta \ll 1$

The spectral floor prevents the final spectrum from falling below a predetermined minimum level instead of being set to zero, and the over-subtraction factor controls how much noise power is subtracted from noisy speech power in each frame. The *a-posterior* segmental SNR determines the over-subtraction factor. The following formula can be used to compute the over-subtraction factor

$$\alpha = \alpha_0 + (\mathrm{SNR})\left(\frac{\alpha_{\min} - \alpha_0}{\mathrm{SNR}_{\max}}\right) \tag{10}$$

The subtraction factor subtracts an overestimation of noise from the noisy spectrum in this approach, which assumes that noise has a uniform influence on the speech spectrum. As a result, different combinations of the over-subtraction factor α, and spectral floor

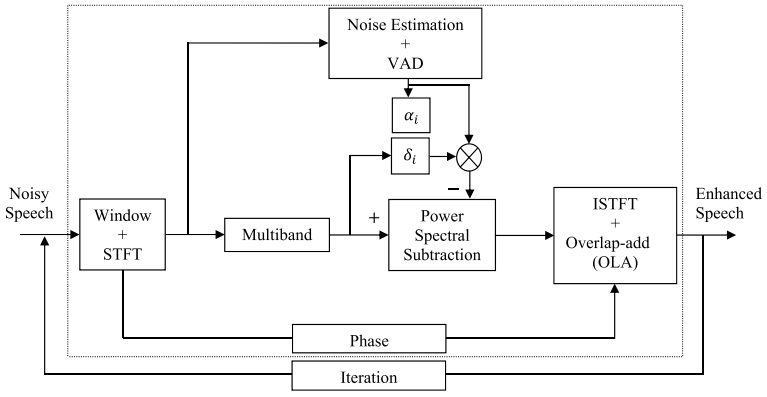

Fig. 2. Block diagram of iterative-processed multiband speech enhancement (IP-MBSE)

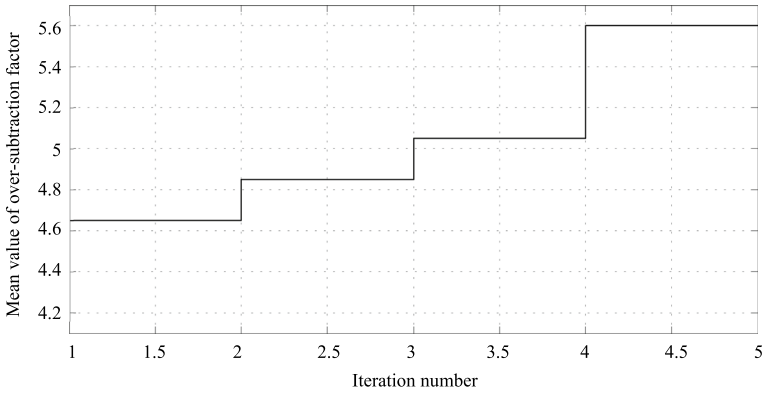

Fig. 3. Relation between the iteration number and the over-subtraction factor mean value

parameter β produce a tradeoff between the amount of leftover sound and the level of perceived musical sound for a balance of speech distortion and musical sound removal. When the parameter β is set to a high value, only a small amount of musical sound is audible; when β is set to a low value, the leftover sound is greatly reduced, but the musical sound becomes quite annoying. As a result, the appropriate value of α is set as per (10) and $\beta = 0.03$.

Although this method reduces perceived musical sound, background noise remains, and enhanced speech is distorted.

2.2 Multiband spectral subtraction (MBSS)

In the real world, degradations have a different impact on the speech spectrum. A linear frequency spacing multiband approach to SOS is presented in [9]. The noisy spectrum is bifurcated into K ($K=4$) non-intersecting evenly spaced frequency subbands in this scheme, with spectral over-subtraction being applied independently in each subband. The over-subtraction factor for each subband is re-adjusted using the multiband spectral subtraction (MBSS) scheme. As a result, the estimation of the clean speech spectrum in the i^{th} subband is calculated to be as

$$\left|\hat{S}_i(\omega)\right|^2 = \begin{cases} \left[\left|Y_i(\omega)\right|^2 - \alpha_i \delta_i \left|\hat{D}_i(\omega)\right|^2\right], & \text{if } \left|\hat{S}_i(\omega)\right|^2 > \beta\left|Y_i(\omega)\right|^2 \\ \beta\left|Y_i(\omega)\right|^2 & \text{else} \end{cases} \tag{11}$$

where $k_i < \omega < k_{i+1}$.

The start and end limits of the i^{th} subband are represented by k_i and k_{i+1}. The α_i is the i^{th} subband-specific over-subtraction factor, which is a function of the segmental SNR (SegSNR) and allows some control over the noise subtraction level in each subband. The SegSNR_i is computed using spectral components from each subband i as

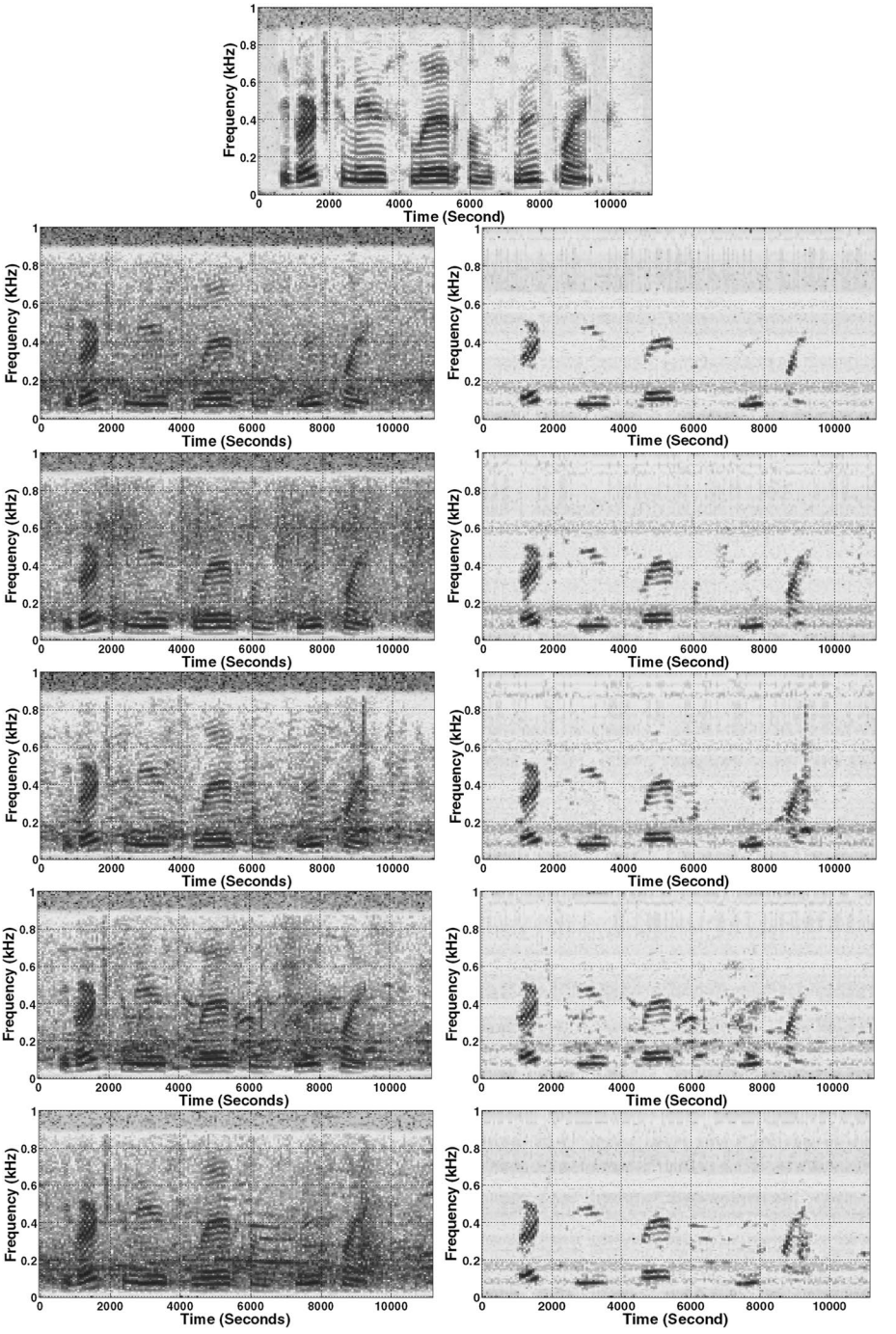

Fig. 4. Speech spectrograms of *sp1* utterance, "*The birch canoe slid on the smooth planks*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** (LEFT SIDE) speech degraded by Car, Train, Babble, Restaurant, Airport, Street, Exhibition, and White noise, respectively (5 dB SNR); **(c)** (RIGHT SIDE) corresponding enhanced speech

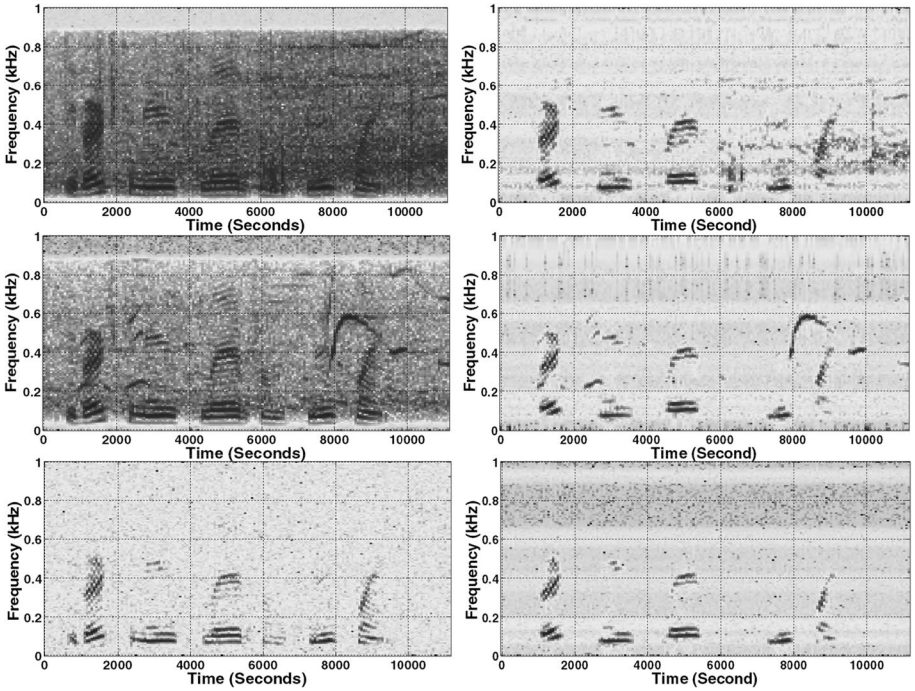

Fig. 4. (continued)

$$\text{SegSNR}_i\ (\text{dB}) = 10\log_{10}\left(\frac{\sum_{\omega=k_i}^{k_{i+1}}\left|Y_i(\omega)\right|^2}{\sum_{\omega=k_i}^{k_{i+1}}\left|\hat{D}_i(\omega)\right|^2}\right) \tag{12}$$

Figure 1 depicts the implementation of four subbands with estimated SegSNR [9]. The noisy speech spectrum is divided into four frequency subbands: {60 Hz ~ 1000 kHz (Subband 1), 1 kHz ~ 2 kHz (Subband 2), 2 kHz ~ 3 kHz (Subband 3), and 3 kHz ~ 4 kHz (Subband 4)}. The figure shows that the SegSNR of the low-frequency bands (Subband 1) is significantly higher than that of the high-frequency subbands (Subband 4) [9].

The δ_i is a subband subtraction factor that may be modified independently for each frequency subband to tailor the noise removal procedure and gives more control over the noise subtraction level in each subband. Because the majority of the speech energy is held below 1 kHz, the values of δ_i [9] are empirically estimated and change as needed.

$$\delta_i = \begin{cases} 1 & f_i \le 1\ \text{kHz} \\ 2.5 & 1\ \text{kHz} < f_i \le \frac{f_s}{2} - 2\ \text{kHz} \\ 1.5 & f_i > \frac{f_s}{2} - 2\ \text{kHz} \end{cases} \tag{13}$$

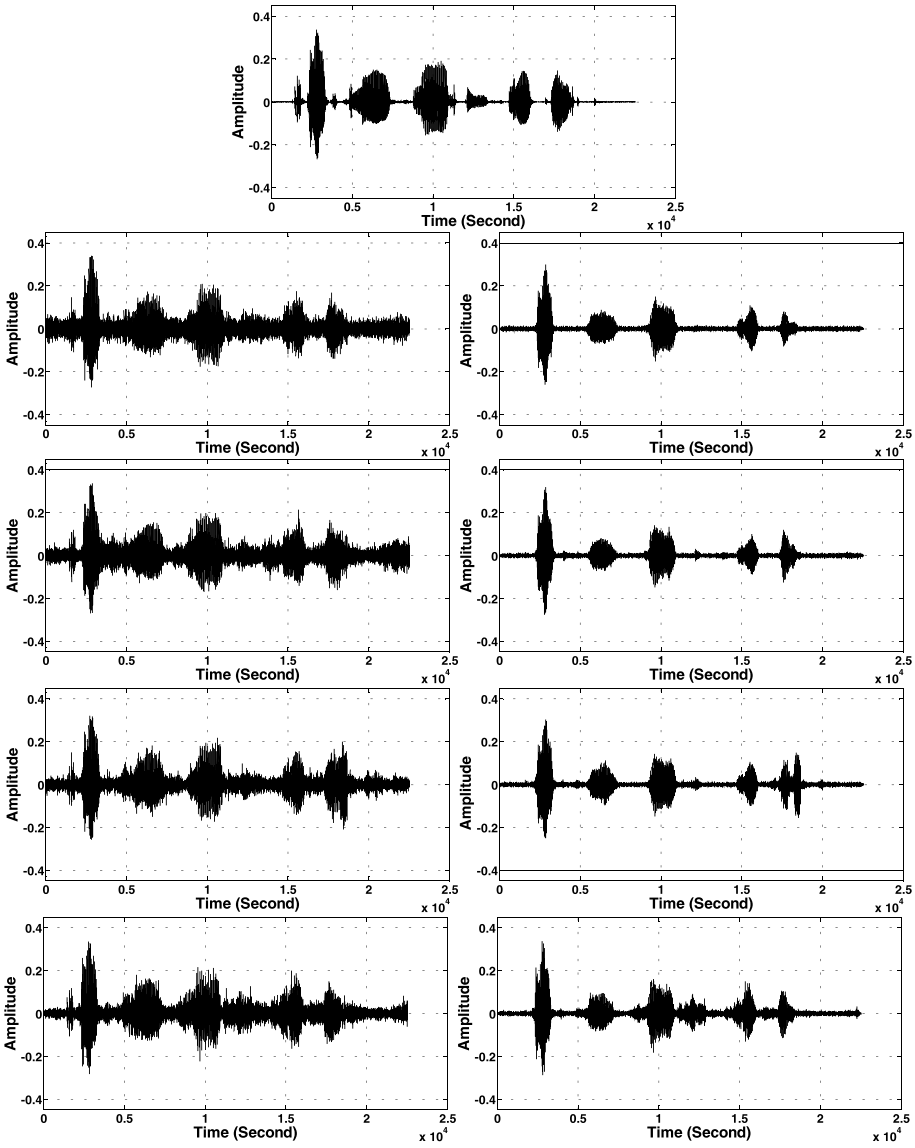

Fig. 5. Temporal waveforms of *sp1* utterance, "*The birch canoe slid on the smooth planks*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** (LEFT SIDE) speech degraded by Car, Train, Babble, Restaurant, Airport, Street, Exhibition, and White noise, respectively (5 dB SNR); **(c)** (RIGHT SIDE) corresponding enhanced speech

The higher frequency of the i^{th} subband is f_i, and the sampling frequency is f_s. Because the lower frequencies contain the majority of the speech energy, choosing the lower values of δ_i for the lower subbands minimizes speech distortion. Both the α_i and δ_i factors can be modified for each subband for different speech situations to boost speech clarity.

Because real noise is highly random, improving the MBSS for WGN reduction is required. However, MBSS outperforms the spectral subtraction method [7] and SOS [8].

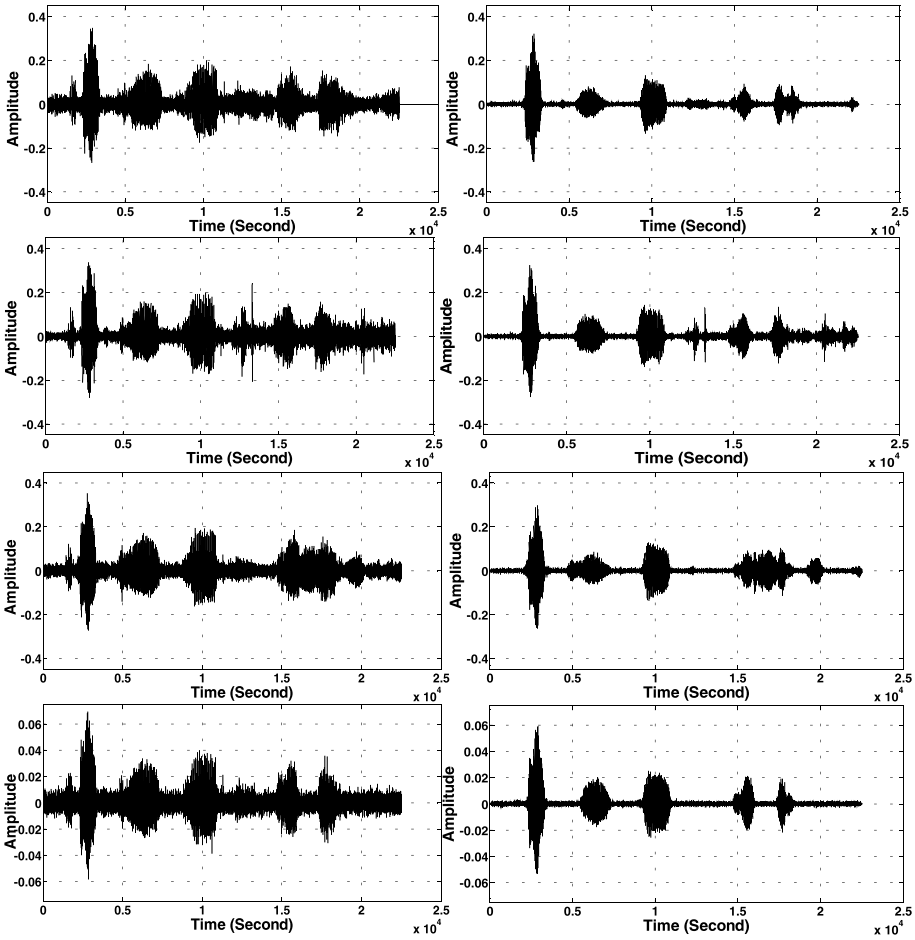

Fig. 5. (continued)

3 Iterative-processed multiband speech enhancement (IP-MBSE)

The additive background noise is converted to an annoying leftover sound with musical structure via the multiband spectral subtraction (MBSS) processing step. This paper proposes an iterative-processed multiband speech enhancement (IP-MBSE) post-processing method for suppressing musical sound in enhanced speech recordings. In the suggested method, the outturn of the MBSS processing step is fed into the subsequent iteration, which estimates the noise spectrum after each repetition (iteration) and performs spectral over-subtraction in each subband separately. By repeatedly applying the enhanced speech to the input and executing the operation, the proposed method reduces musical sound even further. This procedure is iterated only a few times because a higher iteration number distorts the signal, while a lower iteration number retains the musical sound. Because a higher iteration number distorts the signal, a lower iteration number retains the musical sound in the estimated speech. This process is iterated a few times.

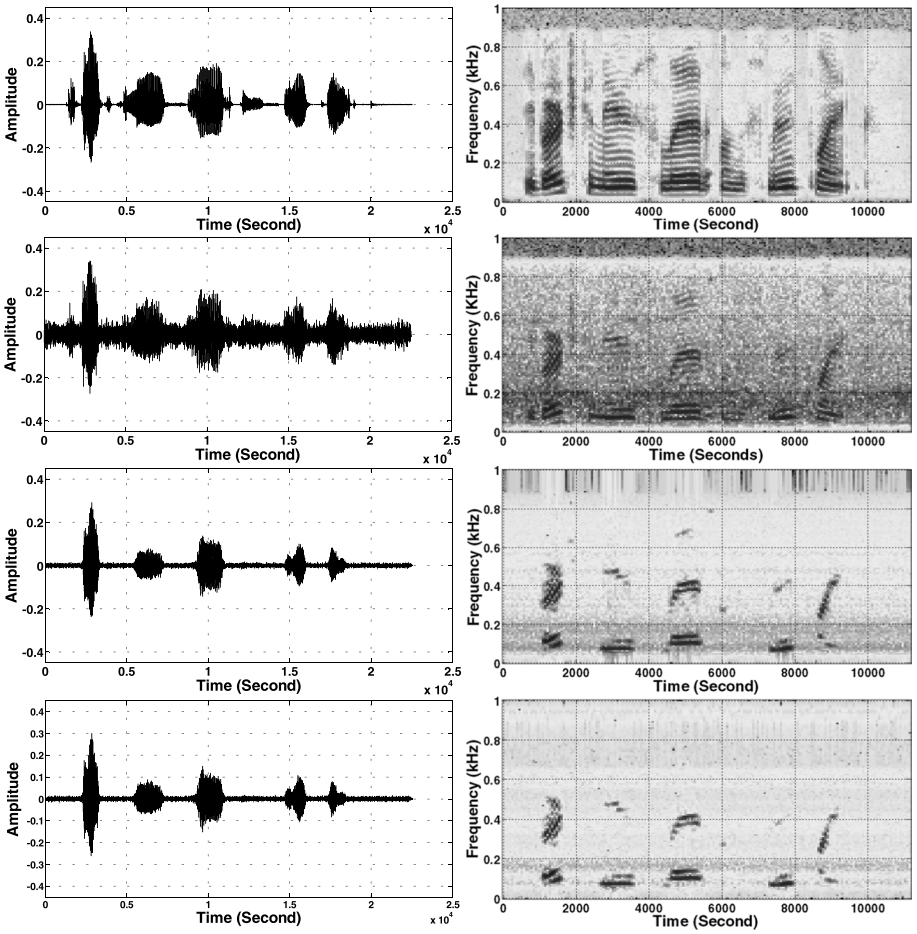

Fig. 6. Temporal waveforms and speech spectrogram of *sp1* utterance, "*The birch canoe slid on the smooth planks*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** noisy speech (degraded by Car noise at 5 dB SNR); **(c)** speech enhanced by MBSS (PESQ =1.78), and **(d)** speech enhanced by IP-MBSE (1.92)

Figure 2 depicts the block diagram of iterative-processed multiband speech enhancement (IP-MBSE). Iteration is used repeatedly to estimate speech as input to improve speech and eliminate musical sounds. As shown in Fig. 2, the additive background noise transforms into a musical sound after the first step of conventional MBSS. Assume the input signal is $y[n]$, and the enhanced speech obtained after the MBSS step is $\hat{s}[n]$. As a result, the MBSS reduces additive noise significantly. This noise reduction is associated with the presence of annoying musical structure sound in the enhanced speech $\hat{s}[n]$. By re-estimating, the remaining noise from each subband in each iteration is fed to the following iteration phase in IP-MBSE. As a result, the final enhanced outgoing speech signal can be obtained after a finite number of iterations.

The iterative technique is inspired by Wiener filtering, which is the noise reduction method [10–12]. As a result, if the noise estimation and MBSS procedures are considered filtering steps, the filter's outturn is employed not just for filter design but also for the iteration that follows. This filter may be adaptively renewed to enhance speech clarity and intelligibility by re-estimating leftover sound.

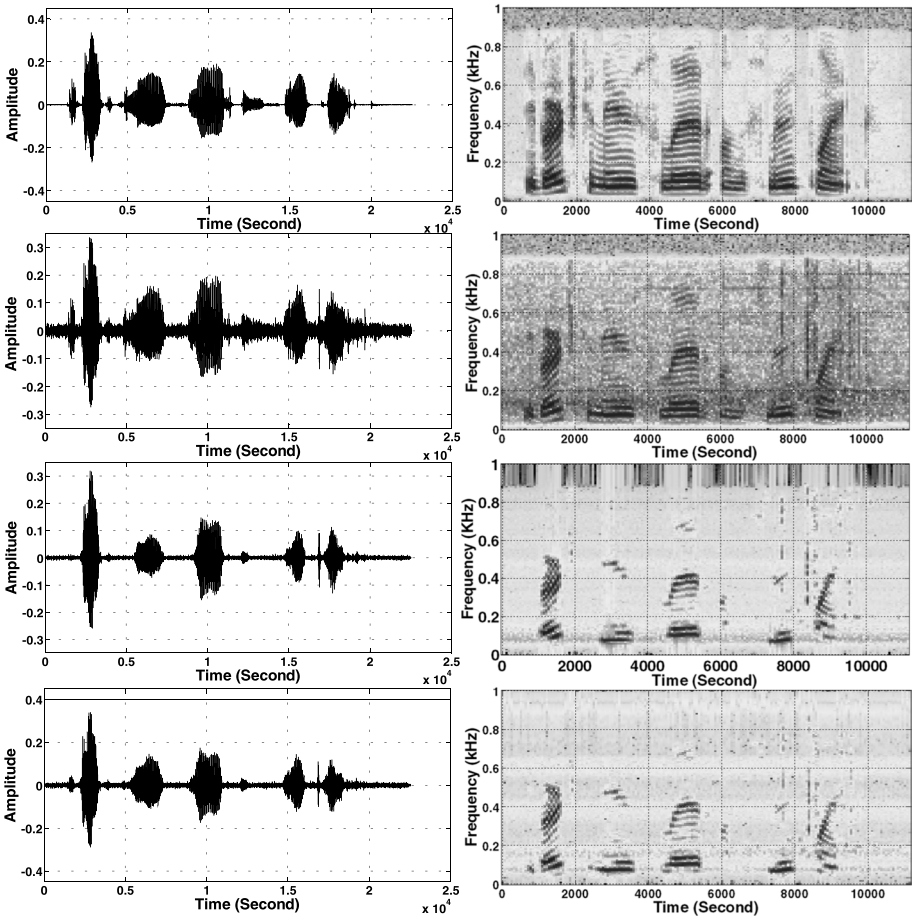

Fig. 7. Temporal waveforms and speech spectrogram of *sp1* utterance, "*The birch canoe slid on the smooth planks*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** noisy speech (degraded by Car noise at 10 dB SNR); **(c)** speech enhanced by MBSS (PESQ=2.03), and **(d)** speech enhanced by IP-MBSE (PESQ=2.15)

The noisy speech at the m^{th} iteration step, where m represents the iteration count, is expressed as

$$y[m,n] = s[m,n] + d[m,n] \tag{14}$$

$y[m,n]$, $s[m,n]$, and $d[m,n]$ are the n^{th} samples at m^{th} iteration of the degraded speech, clear speech, and interference, respectively. In MBSS processing, the m^{th} iteration step is calculated as

$$\left|\hat{S}_i(m,\omega)\right|^2 = \begin{cases} \left|Y_i(m,\omega)\right|^2 - \alpha_i \delta_i \left|\hat{D}_i(m,\omega)\right|^2 & \text{if } \left|\hat{S}_i(m,\omega)\right|^2 > \beta\left|Y_i(m,\omega)\right|^2 \\ \beta\left|Y_i(m,\omega)\right|^2 & \text{else} \end{cases} \tag{15}$$

where $k_i < \omega < k_{i+1}$

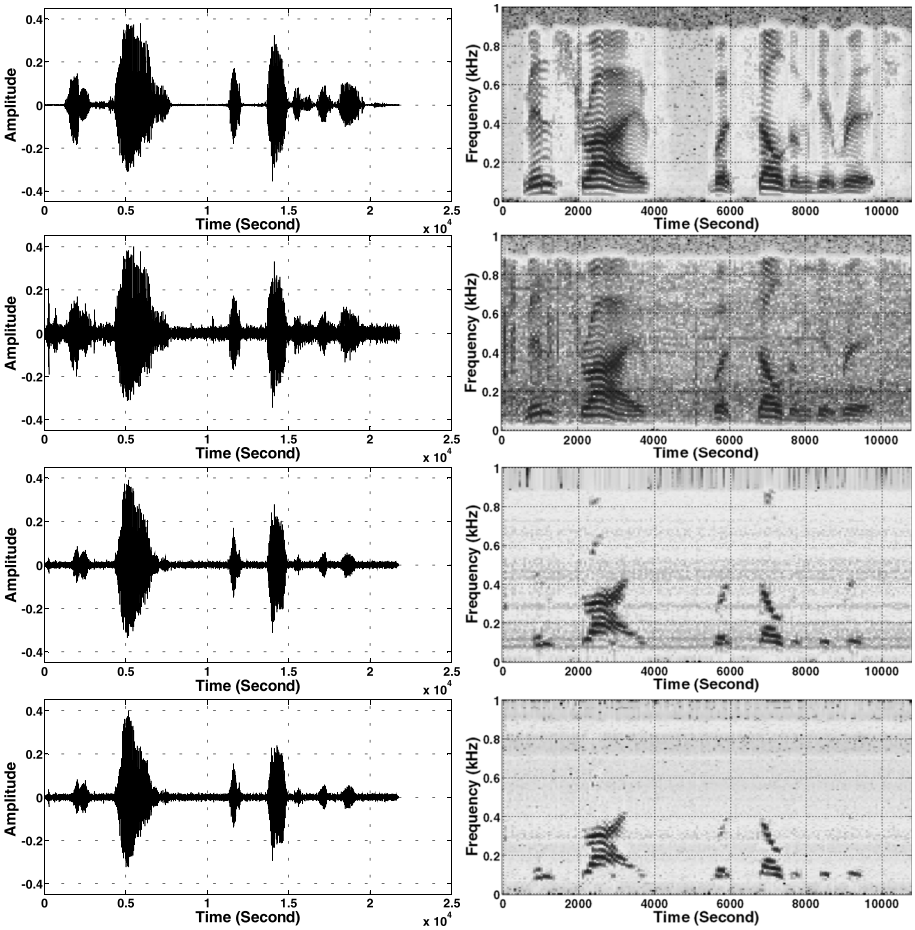

Fig. 8. Temporal waveforms and speech spectrograms of *sp6* utterance, "*Men strive but seldom get rich*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** noisy speech (speech degraded by Car noise at 10 dB SNR); **(c)** speech enhanced by MBSS (PESQ=2.16); and **(d)** speech enhanced by IP-MBSE (PESQ=2.27)

$$\left|\hat{S}_i(m+1,\omega)\right|^2 = \left|\hat{S}_i(m,\omega)\right|^2 \left|Y_i(m,\omega)\right|^2 \tag{16}$$

$\left|\hat{S}_i(m,\omega)\right|^2$, $|Y_i(m,\omega)|^2$, and $\left|\hat{D}_i(m,\omega)\right|^2$ represent the estimated speech, degraded speech, and estimated noise power in the i^{th} subband, respectively, at the m^{th} iteration step. After the m^{th} iteration, the outturn $\hat{S}_i(m,\omega)$ is used as the input in the $(m+1)^{\text{th}}$ iteration processing as

$$y[m+1,n] = \hat{s}[m,n] \tag{17}$$

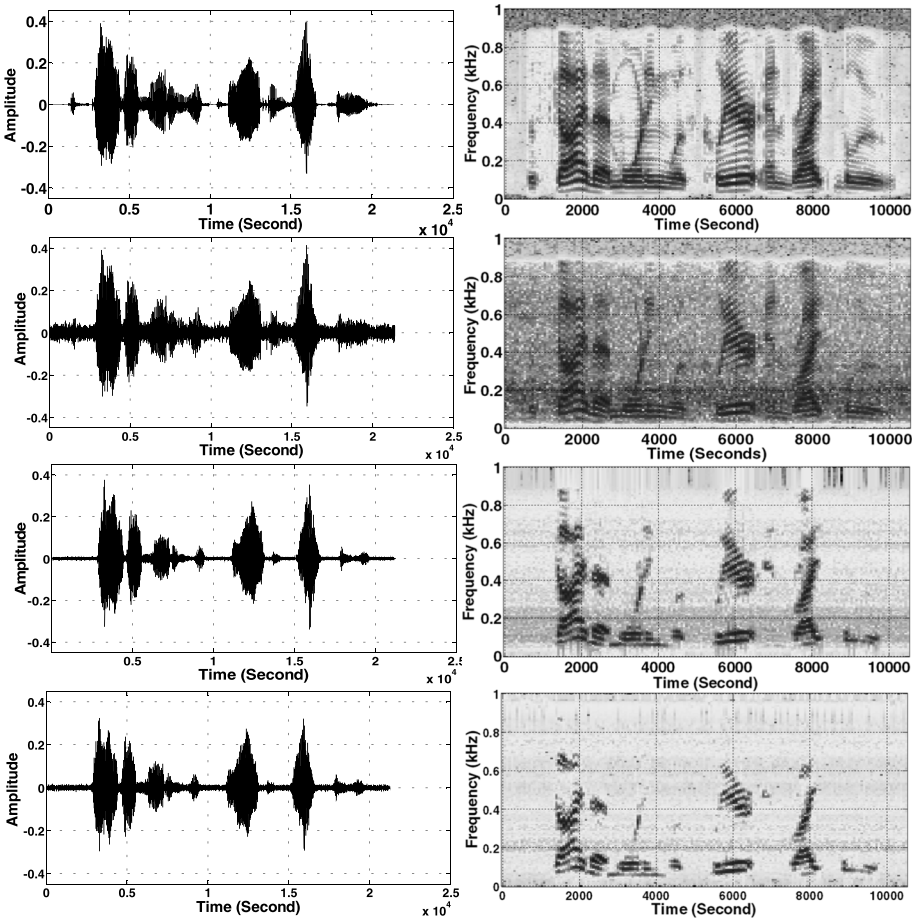

Fig. 9. Temporal waveforms and speech spectrograms of sp10 *sp10* utterance, "*The sky that morning was clear and bright blue*", by male speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** noisy speech (speech degraded by Car noise at 10 dB SNR); **(c)** speech enhanced by MBSS (PESQ=2.26); and **(d)** speech enhanced by IP-MBSE (PESQ=2.46)

The noise spectrum is estimated in IP-MBSE for each iteration based on the noise component that remains just after the preceding step's repetitive processing. The leftover noise component is the noise component of $y[m+1,n]$ that the MBSS was unable to suppress at the m^{th} iteration. Because each MBSS processing step reduces the amount of noise, increasing the iteration in this method reduces the quantity of leftover noise.

The number of iterations is a significant aspect of the IP-MBSE, and it affects speech enhancement performance [12]. At the end of each iteration, the SegSNR is proportional to the iteration number and grows as the iterations increase. Because the over-subtraction factor is affected by SegSNR, it increases as well. Figure 4 depicts the relationship between the iteration number and the mean value of the over-subtraction factor. The greater the number of iterations, as shown in Fig. 3, the better the speech enhancement performance with less musical sound.

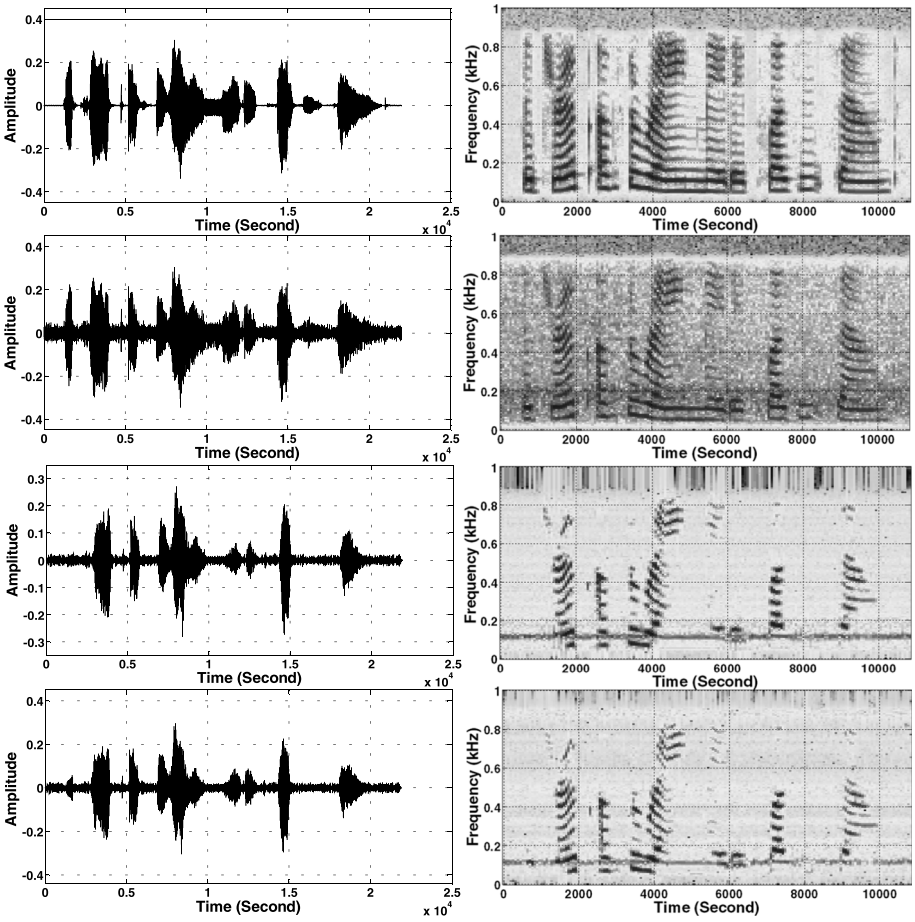

Fig. 10. Temporal waveforms and speech spectrograms of *sp12* utterance, "*The drip of the rain made a pleasant sound*", by female speaker from *NOIZEUS* corpus: **(a)** clean speech; **(b)** noisy speech (degraded by Car noise at 10 dB SNR); **(c)** speech enhanced by MBSS (PESQ=2.01); and **(d)** speech enhanced by IP-MBSE (PESQ=2.26)

4 Evaluation of performance and experimental results

The experimental findings and performance evaluation of the suggested methodology, IP-MBSE, and its compression with the conventional MBSS scheme are shown in this section. We took noisy speech samples (sampled at 8 kHz) from the NOIZEUS corpus speech database [13] for simulations. For the experiment, we employed four distinct utterances (three male speakers and a female speaker).

The time-frequency distributions of the background noises are varied, and they have varied effects on the speech signals. For the performance assessment of IP-MBSE, the utterances are degraded with seven different real noises and white Gaussian noise at various SNR levels ranging from 0 - 15 dB. The real-world noises are those of cars, trains, restaurants, babbles, airports, streets, and exhibitions.

Table 1. IP-MBSE objective evaluation and comparison in terms of SNR [dB] and SegSNR [dB]

Type of Noise	Enhancement methods	SNR [dB]				SegSNR [dB]			
		0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Car	MBSS	4.26	6.01	6.39	6.88	4.19	5.98	6.35	6.82
	IP-MBSE	4.50	6.11	6.46	6.94	4.46	6.10	6.44	6.90
Train	MBSS	3.47	5.82	7.41	7.19	3.42	5.75	7.38	7.17
	IP-MBSE	3.57	5.96	7.33	7.23	3.54	5.92	7.33	7.25
Restaurant	MBSS	2.15	4.60	5.73	6.46	2.10	4.54	5.66	6.45
	IP-MBSE	2.27	5.04	5.84	6.52	2.24	4.99	5.83	6.52
Babble	MBSS	2.27	4.64	6.45	5.92	2.21	4.63	6.42	5.87
	IP-MBSE	2.40	4.89	6.51	5.98	2.35	4.88	6.50	5.97
Airport	MBSS	3.61	4.81	6.26	5.57	3.52	4.76	6.23	5.50
	IP-MBSE	3.71	4.97	6.34	5.68	3.63	4.91	6.33	5.66
Street	MBSS	4.24	5.00	5.68	6.59	4.17	4.89	5.63	6.53
	IP-MBSE	4.42	5.56	5.72	6.66	4.39	5.38	5.68	6.63
Exhibition	MBSS	3.65	3.28	7.12	6.89	3.60	3.20	7.09	6.86
	IP-MBSE	3.92	3.34	7.12	6.91	3.91	3.27	7.11	6.89
White	MBSS	5.09	6.87	7.29	7.49	5.03	6.85	7.28	7.47
	IP-MBSE	5.25	6.86	7.25	7.46	5.23	6.86	7.26	7.46

The noisy utterance is separated into many frames for the experimental work, with a frame size of 256 and 50% overlap. The noisy signal is subjected to the Hamming window. The noise estimate is updated by averaging throughout the pause frames (20 frames). The noise power spectral density is calculated with a smoothing factor of 0.9.

Table 2. The outcome of a noise reduction speech quality test

Type of Noise	Enhancement methods	PESQ Score				MOS Score			
		0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Car	MBSS	1.615	1.776	2.030	2.293	1.8	2.7	3.5	4.3
	IP-MBSE	1.693	1.915	2.147	2.489	2	2.8	3.6	4.1
Train	MBSS	1.608	1.886	1.850	2.166	2.6	3.3	3.7	4.2
	IP-MBSE	1.693	1.893	2.010	2.353	2.3	2.9	3.4	4.2
Restaurant	MBSS	1.697	1.885	2.039	2.295	1.8	2.7	3.5	4.0
	IP-MBSE	1.787	1.927	2.187	2.479	1.9	2.7	3.4	4.1
Babble	MBSS	1.665	1.907	2.134	2.237	1.6	2.7	3.6	4.2
	IP-MBSE	1.667	2.036	2.341	2.413	1.8	2.7	3.6	4.3
Airport	MBSS	1.774	1.953	2.161	2.263	1.8	2.8	3.6	4.2
	IP-MBSE	1.876	2.061	2.294	2.471	1.6	2.1	2.8	3.9
Street	MBSS	1.416	1.866	2.002	2.300	1.8	2.6	3.5	4.2
	IP-MBSE	1.614	1.956	2.190	2.501	2	2.7	3.5	4.2
Exhibition	MBSS	1.298	1.633	2.001	2.260	1.8	2.7	3.4	4
	IP-MBSE	1.379	1.782	2.102	2.420	1.9	2.6	3.8	4.4
White	MBSS	1.433	1.669	2.069	2.297	2.6	3.5	4.1	4.5
	IP-MBSE	1.602	1.901	2.235	2.474	2.9	3.6	4	4.4

The number of iterations has a big impact on IP-MBSE's speech enhancement performance. The relationship between iteration number and mean over-subtraction factor (α) is depicted in Fig. 3 to investigate the connection between speech enhancement performance and iteration number. It has been observed that α increases with the iterations, implying that the higher the number of iterations, the better the speech enhancement performance with less musical sound. Nevertheless, the waveforms and spectrograms in Figs. 4, 5, 6, 7, 8, 9, 10 show that increasing the iteration number reduces the speech component by some amount while effectively suppressing the musical sound. As a result, for the speech degraded by car noise, we fix iterations 2 to 3 while leaving the additional variables the same as in the reference MBSS step.

Both objective and subjective indicators have been used to assess IP-MBSE performance. SNR, SegSNR, and PESQ are objective metrics, while MOS and spectrograms are subjective metrics.

4.1 Objective evaluation

a). Signal-to-Noise Ratio (SNR): This is calculated by dividing an utterance's total signal energy by its total noise energy. The equation below is used to evaluate the SNR results of improved signals.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{L} s^2[n]}{\sum_{n=1}^{L}\{s[n]-\hat{s}[n]\}^2}\right) \tag{18}$$

n, L represents the sample index and the number of samples. $s[n]$,$\hat{s}[n]$ denotes the clean speech and improved speech. The summing is done across the length of the signal.

b). Segmental Signal-to-Noise Ratio (SegSNR): The average signal to noise energy ratio per frame is known as SegSNR, and it may be written as:

$$\mathrm{SegSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} s^2[n]}{\sum_{n=N_m}^{N_m+N-1}\{s[n]-\hat{s}[n]\}^2}\right) \tag{19}$$

M, N denotes the number of frames in a signal and the number of samples frames. The SegSNR is better correlated with perceptual clarity than the SNR. The greater SegSNR indicates the less distortions.

c). Perceptual Evaluation of Speech Quality (PESQ): The ITU-T recommends the PESQ for speech clarity assessment because it is an objective evaluation and predicts the subjective opinion score of a degraded speech sample [14]. In several testing situations, the PESQ is found to be highly linked with subjective tests [14].

4.2 Subjective evaluation – Mean Opinion Score (MOS)

A subjective evaluation is based on the judgment of others. The listening tests for our experimental review were conducted with five participants in a confined room wearing

headphones. For each test signal, each listener assigns a score ranging from one to 4.5. This score reflects their overall impression of the clarity of the speech, which includes musical sound and background noise, as well as speech distortion. These tests were conducted on a scale that corresponded to the MOS scale described in [3]. For each speaker, the following procedure is used: clear and noisy speech is played and replayed twice, and each signal is played and repeated twice.

At various SNR levels, Table 1 compares IP-MBSE to standard MBSS with respect to global SNR [dB] and SegSNR [dB]. For various forms of noise, the value of SNR and SegSNR for IP-MBSE is superior to MBSS.

The PESQ and MOS scores of IP-MBSE versus MBSS are shown in Table 2. The IP-MBSE outperforms traditional MBSS on the PESQ test for all noises except train and airport noise, while better speech generated by IP-MBSE exceeds MBSS on the MOS measure.

The time-wave patterns and spectrograms of clear, noisy, and enhanced speech signals are shown in Figs. 4, 5, 6, 7, 8, 9, 10. As seen in Figs. 4, 5, 6, 7, 8, 9, 10, the IP-MBSE decreases the musical structure of the leftover noise more than MBSS. As a result, IP-MBSE-affected speech is more pleasant to listen to, and musical sounds have a white character with acceptable distortion. This backs up the results of the SNR, SegSNR, and PESQ tests (Table 1), as well as listening tests (Table 2).

5 Conclusion

In this paper, we investigated an iterative-processed multiband speech enhancement (IP-MBSE) for the suppression of annoying musical sounds. The outturn of multiband spectral subtraction (MBSS) is fed into the proposed technique in subsequent iterations. The iteration number is crucial in IP-MBSE because a higher number distorts the signal while a lower number retains the musical sound in the estimated speech. As a result, only a few iterations are carried out. When IP-MBSE is compared to the conventional MBSS, it is found that IP-MBSE outperforms MBSS at low SNRs.

References

1. O'Shaughnessy D (2007) Speech Communications: Human and Machine, 2nd ed., Hyderabad, India: University Press (India) Pvt. Ltd.
2. Ephraim Y (1992) Statistical-model-based speech enhancement systems. in Proceedings IEEE 80(10):1526–1555
3. Loizou PC (2013) Speech Enhancement: Theory and Practice, IInd ed. Taylor and Francis
4. Ephraim Y, Ari HL, Roberts W (2006) A brief survey of speech enhancement, in Electrical Engineering Handbook, 3rd ed. Boca Raton, FL: CRC
5. Ephraim Y, Cohen I (2006) Recent advancements in speech enhancement, in The Electrical Engineering Handbook, CRC Press, ch. 5, pp. 12-26
6. Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. Proceedings IEEE 67:1586–1604
7. Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Transaction Acoustic, Speech, Signal Processing 27(2):113–120
8. Berouti M, Schwartz R, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise, in Proceedings Int. Conf. Acoustic, Speech, Signal Processing, Washington DC, 208-211
9. Kamath S, Loizou P (2002) A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in Proceedings Int. Conf. Acoustic, Speech, Signal Processing, Orlando, USA, May

10. Upadhyay N, Karmakar A (2012) Single channel speech enhancement utilizing iterative processing of multi-band spectral subtraction algorithm, in Proceedings IEEE Int. Conf. Power, Control and Embedded System, MNNIT Allahabad, India, Dec. 17-19, 196-201
11. Ogata S, Shimamura T (2001) Reinforced spectral subtraction method to enhance speech signal. in Proceedings Int. Conf. Electrical and Electronic Technology 1:242–245
12. Li S, Wang J-Q, Niu M, Jing X-J, Liu T (2010) "Iterative spectral subtraction method for millimeter-wave conducted speech enhancement," J. Biomedical Science and Engineering 3:187–192
13. A noisy speech corpus for assessment of speech enhancement algorithms. http://www.utdallas.edu/~loizou/speech/noizeus/
14. "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.