Dong and Lee *EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:3
https://doi.org/10.1186/s13636-018-0126-8

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering

Huan-Yu Dong[1] and Chang-Myung Lee[1,2*]

Abstract

The speech intelligibility of indoor public address systems is degraded by reverberation and background noise. This paper proposes a preprocessing method that combines speech enhancement and inverse filtering to improve the speech intelligibility in such environments. An energy redistribution speech enhancement method was modified for use in reverberation conditions, and an auditory-model-based fast inverse filter was designed to achieve better dereverberation performance. An experiment was performed in various noisy, reverberant environments, and the test results verified the stability and effectiveness of the proposed method. In addition, a listening test was carried out to compare the performance of different algorithms subjectively. The objective and subjective evaluation results reveal that the speech intelligibility is significantly improved by the proposed method.

Keywords: Speech intelligibility, Speech enhancement, Inverse filtering, Auditory model, Dereverberation

1 Introduction

An indoor public address (I-PA) system is a sound amplification system that is widely used in auditoriums, classrooms, factories, and conference rooms. However, its speech intelligibility is often degraded due to near-end [1] reverberation and background noise [2]. Therefore, it is desirable to find an effective way to improve the speech intelligibility in such environments.

Reverberation is caused by wall reflections that distort the sound transmission channel [3], while background noise degrades the speech intelligibility through noise masking [4]. Thus, methods for improving speech intelligibility can be broadly classified into two categories. The first focuses on compensation for transmission channel distortion [3, 5–12], and the other category focuses on noise suppression and speech enhancement. In the first category, sound transmission in enclosed spaces is regarded as a linear time invariant (LTI) system [5, 6], so the output response of the system can be expressed as the convolution of the input signal and room impulse response (RIR). Therefore, the influence of reverberation can be eliminated by realizing the inverse of an RIR [3]. However, this inverse will be either unstable or acausal since the RIR is generally considered a nonminimum phase function [7].

In the first research on this problem [5], Neely realized a stable and causal inverse filter through decomposition of the RIR into the minimum phase and all-pass phase. This inverse filter can basically eliminate the distortion caused by wall reflections. An adaptive equalization (A-EQ) method [6] was later proposed to compensate for the distortion of the room frequency response. The equalizer could minimize the square errors between the target response and input signal adaptively, but the method was very sensitive to peaks and notches for the room responses.

Based on Neely's method, a new equalization method was proposed by combining a vector quantization method with an all-pole room transfer function (RTF) model to reduce the effects of the reverberation by using a lower equalizer order [7]. However, this approach is based on an approximation of the RTF, so the exact solution of the inverse filter cannot be obtained. Kirkeby and Nelson proposed a fast inverse filtering (FIF) method for designing single or multichannel sound reproduction systems [8–10]. This method uses the principles of least squares optimization to obtain a stable and causal inverse filter, as well as regularization, to

* Correspondence: cmlee@ulsan.ac.kr
[1]Department of Mechanical and Automotive Engineering, University of Ulsan, 93 Daehak-ro, Nam-Gu, Ulsan 44610, South Korea
[2]Lavrentyev Institute of Hydrodynamics, Siberian Brunch of Russian Academy of Sciences, 15 Lavrentyev Ave., Novosibirsk 630090, Russia

realize fast deconvolution. Although this method needs to use relatively long inverse filters, the algorithm has higher accuracy and fast deconvolution speed. Therefore, this algorithm has received much attention and is still used in current equalization methods.

Based on this algorithm, a warped domain equalization (W-EQ) method was proposed to improve the listening experience [12]. This method uses the bark scale, which is related to auditory perception and low-frequency response equalization and produces a better listening experience than other equalization methods [6, 13, 14]. However, the bark scale is not an auditory model and cannot simulate the frequency response characteristics of the basilar membrane in the cochlea. Moreover, these equalization methods do not account for the influence of background noise on speech intelligibility.

Increasing the playback level is one clear solution to improve the speech intelligibility in the event of background noise. However, it is impossible to increase the output level indefinitely due to the limited power output of loudspeakers and the pain-threshold pressure limitation of the ear [1]. In addition, in the case of I-PA systems, the listener is located in a noisy environment, and the noise reaches the ears without any possibility of intercepting it beforehand [15]. Therefore, a preprocessing speech enhancement method without increasing the output power would be more suitable for use with I-PA systems [16, 17].

An energy redistribution voiced/unvoiced (ERVU) method was proposed to improve intelligibility without increasing the output power [16]. The method redistributes more speech energy to the transient regions to reinforce speech signals. A perceptual distortion measure (PDM)-based speech enhancement (PDMSE) method was proposed [17] based on the ERVU method and the PDM algorithm [18]. Compared with the ERVU method, the PDMSE method can further improve speech quality without decreasing intelligibility. However, these methods do not consider the influence of reverberation on speech intelligibility.

In recent years, only a few studies have considered the effects of reverberation and background noise simultaneously [19–22]. Some methods just use the near-end speech enhancement method to reduce the influence of both reverberation and background noise [19, 20]. Other methods pre-compensate the output speech by obtaining the optimal solution of the established mathematical model to improve intelligibility [21, 22]. Crespo and Hendriks [21] proposed a multizone speech reinforcement method based on a general optimization framework. The signal model considered the influence of RTF on intelligibility in noisy environments, and the effectiveness of this approach was verified by simulation.

Hendriks et al. [22] proposed an approximated speech intelligibility index (ASII) method to improve the speech intelligibility in a single-zone scenario. Unlike the Multizone method [21], the ASII method uses a speech intelligibility index to establish a mathematical model that includes late reverberation and noise. The optimal solution of the mathematical model is used to preprocess the output speech to improve intelligibility. Although the Multizone and ASII methods could improve the speech intelligibility in noisy and reverberant environments, the distortion of the speech transmission channel and the auditory features of the human ear were not considered at the same time during the signal preprocessing. Therefore, the Multizone and ASII methods do not fundamentally compensate the distortion of the transmission channel, and the dereverberation performance is quite limited.

This paper proposes a new preprocessing method for improving speech intelligibility by a combination of the PDMSE method and the FIF method. The PDMSE method was modified for reverberant environments, and a new Gammatone (GT)-filter-based FIF method was designed to achieve better equalization and dereverberation performance. Compared with the A-EQ, W-EQ, and FIF equalization methods, the GT-filter-based FIF method can further decrease the distortion of the transmission channel. Compared with individual FIF and PDMSE methods, the improved combination method has better stability and higher speech quality. Furthermore, compared with the multizone and ASII methods, the combination method can significantly improve the speech intelligibility in different noisy and reverberant environments.

To validate the method, an experiment was performed in real environments with various noise and reverberation conditions. The speech transmission index, spectrogram, log-spectral distortion measure, short-time objective intelligibility measure, and modified rhyme test were used to compare the performance. The objective and subjective evaluation results illustrate that the method can effectively improve the speech intelligibility of I-PA systems in noisy and reverberant environments.

The remainder of the paper is organized as follows. Section 2 describes the algorithm in detail. Section 3 describes the experimental design and hardware setup. Section 4 presents the test results of the evaluations, and Section 5 concludes the paper.

2 Proposed speech intelligibility improvement algorithm

The overall scheme of the proposed method is shown in Fig. 1. Initially, the input signal $s(n)$ is captured, and a time-frequency (TF) decomposition and GT filter are applied to obtain the short-term clean speech

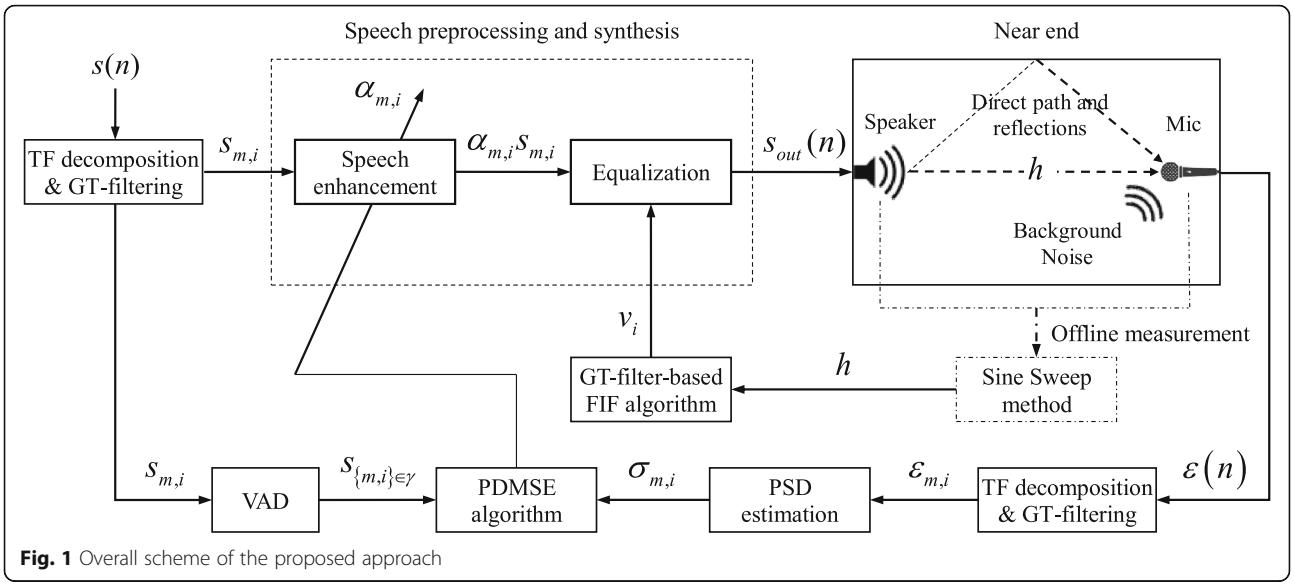

Fig. 1 Overall scheme of the proposed approach

frame $s_{m,i}$. $s_{m,i}$ is then sent to the voice activity detection (VAD) module and to the preprocessing and synthesis module. The VAD module is applied to obtain the positions of the active voice in speech signals and prepare the detection information for the PDMSE algorithm.

In the block for speech preprocessing and synthesis, a modified PDMSE method is used in the speech enhancement stage to increase the energy of transient speech. Next, a GT-filter-based FIF method is used in the equalization stage to pre-compensate the distortion of the transmission channel. The final preprocessing and synthesis signal $s_{\text{out}}(n)$ is used as an input for the loudspeaker to broadcast. The distortion signal $\varepsilon(n)$ is then recorded by a microphone, and TF decomposition and GT filtering are once again performed to obtain the short-term distortion frame $\varepsilon_{m,i}$.

The power spectral density (PSD) estimation module is next applied to estimate the energy of background noise. Finally, the gain function α is calculated by the PDMSE algorithm, and the inverse sub-filters v_i are obtained by the GT-filter-based FIF algorithm. Both parameters are used to adjust the preprocessing speech signal to obtain the best speech intelligibility. Furthermore, based on the method by Meng et al. [23], a sine sweep signal with a length of 10 s is used as an excitation signal to obtain the RIR in advance to calculate the inverse filter.

Three modules in the block diagram of Fig. 1 are mainly discussed in the next sections. Section 2.1 gives the description of the PDMSE algorithm module. Section 2.2 discusses the GT-filter-based FIF algorithm module, and Section 2.3 presents the block for speech preprocessing and synthesis.

2.1 Improved preprocessing speech enhancement

In the PDMSE algorithm, the PDM model plays an important role because it is more sensitive to transients than the spectral-only model. Furthermore, it can detect tiny differences between the input signal and the measured signal within a short time frame (20–40 ms). The PDM is a kind of TF decomposition method based on the spectro-temporal auditory model. The distortion measure $D(s,\varepsilon)$ can be described simply by summing all the individual short-term distortion frames $\varepsilon_{m,i}$ [17]:

$$D(s,\varepsilon)=\sum_{m,i} d\left(s_{m,i},\varepsilon_{m,i}\right), \tag{1}$$

where, $s_{m,i}$ is the clean speech passed through the GT filter, which can be represented as a convolution of the impulse response of the ith GT filter g_i and the mth short-term frame of clean speech s_m, that is, $s_{m,i}=s_m * g_i$. The GT filter is described in Eq. (8) of Section 2.2. Similar to the definition of $s_{m,i}$, the decomposed short-term distortion frames $\varepsilon_{m,i}$ can also be represented as $\varepsilon_{m,i}=\varepsilon_m * g_i$.

The target of the PDM model is to minimize $D(s,\varepsilon)$ in Eq. (1) under the constraint of constant energy of the modified speech. This is accomplished using a gain function α to adjust the input speech signal s. The Lagrange multiplier method is then used to establish a cost function:

$$J=\sum_{\{m,i\}\in\gamma} E\left[d\left(\alpha_{m,i}s_{m,i},\varepsilon_{m,i}\right)\right] +\lambda\left(\sum_{\{m,i\}\in\gamma}\left\|\alpha_{m,i}s_{m,i}\right\|^2-r\right), \tag{2}$$

where γ is the set of speech-active TF units

obtained from the VAD algorithm, and $\|\cdot\|$ represents normalization. λ denotes a Lagrange multiplier. $r = \sum_{\{m,i\}\in\gamma} \|\alpha_{m,i}s_{m,i}\|^2$ is related to the power constraint. The power constraint can be used to satisfy the constraints of the loudspeaker output power or to overcome hearing discomfort due to loud sounds [22].

By minimizing Eq. (2), the gain function α can be solved using the following equation:

$$\alpha_{m,i}^2 = \frac{r\beta_{m,i}^2}{\sum_{\{m',i'\}\in A} \beta_{m',i'}^2 \left\|s_{m',i'}\right\|^2}, \tag{3}$$

where

$$\beta_{m,i} = \left(\frac{E\left[d\left(s_{m,i},\varepsilon_{m,i}\right)\right]}{\left\|s_{m,i}\right\|^2}\right)^{1/4}. \tag{4}$$

The expected value $E[d(s_{m,\,i}, \varepsilon_{m,\,i})]$ in Eq. (4) can be expressed as follows:

$$E\left[d\left(s_{m,i},\varepsilon_{m,i}\right)\right] = \sum_n \frac{\left(E\left[\left|\varepsilon_{m,i}\right|^2\right] * h_s\right)(n)}{\left(\left|s_{m,i}\right|^{2*} h_s\right)(n)}, \tag{5}$$

where h_s is a smoothing low-pass filter. According to previous assumptions [17], the noise PSD within the frequency range of an auditory band is regarded as a "flat" spectrum, so the noise within an auditory band can be simply represented as $\varepsilon_{m,\,i} = (w_m N_{m,\,i}) * g_i$, where w_m and $N_{m,i}$ are the window function and zero mean, respectively.

Based on the central limit theorem, the stochastic process with variance can be represented as $E[N_{m,i}^2(n)] = \sigma_{m,i}^2, \forall n$. By combining this statistical model and the numerator of Eq. (5),

$$E\left[\left|\varepsilon_{m,i}\right|^2(n)\right] = \left(g_i^2 * w_m^2\right)(n)\sigma_{m,i}^2. \tag{6}$$

In the PDMSE method, $\sigma_{m,i}^2$ denotes the PSD estimation of noisy speech. The noise PSD estimation by Hendriks et al. [24] does not consider the influence of reverberation, resulting in overestimation of the noise PSD in noisy reverberant environments [25]. Therefore, in this modified version of the PDMSE method, the PSD estimation was modified based on Faraji and Hendriks' work [25] and made to be applicable in such environments. The average PSD within an auditory filter is then calculated as the PSD estimation results. As the final step, an exponential smoother for the gain function $\alpha_{m,\,i}$ is described by the following equation:

$$\widehat{\alpha}_{m,i} = (1\text{–}0.9)\alpha_{m,i} + 0.9\widehat{\alpha}_{m-1,i}, \tag{7}$$

which is used to prevent the generation of "music noise" during the signal processing.

2.2 Improved fast inverse filtering

The FIF method is used to achieve an "inverse filter" (an equalizer) of the RIR. Taking into account the sensitivity of the human ear to different frequencies [26], a FIF method based on GT filters was designed to achieve suitable dereverberation and equalization performance for human auditory characteristics. In contrast to the 1/3 octave and the bark scale, the GT filter is a kind of auditory filter that can simulate the characteristics of the basilar membrane [18]. The central frequencies of the GT filter banks are distributed in a quasi-logarithmic form and are evenly distributed in the frequency range of the speech signal based on the equivalent rectangular bandwidth (ERB). The ERB is a measure used in psychoacoustics and approximates the bandwidths of the filters in human hearing [27]. The GT filter banks can be represented as follows in the form of an impulse response in the time domain [28]:

$$g(t) = ct^{n-1}e^{-2\pi bt}\cos(2\pi f_0 t + \phi), t > 0, \tag{8}$$

where f_0 is the central frequency of the GT filter banks and c is a constant for controlling the gain. n is the filter order, which is usually set as 4 to simulate the auditory response of human ears accurately [29]. ϕ is the phase of the filter, which can usually be ignored, and b is the decay factor, which can be obtained using the central frequency f_0 as follows:

$$b(f_0) = 1.019 \times (24.7 + 0.108 f_0). \tag{9}$$

For the single-input-single-output (SISO) system, the RIR between the loudspeaker and receiver point contains all the information of the sound transmission channel. The GT filter banks were used to decompose the RIR h to obtain the sub-filters, which are based on the auditory model; that is, $h_i = h * g_i$, where g_i denotes the ith GT filters and h_i denotes the decomposed ith sub-filters. In this process, a total amount of 40 sub-filters are decomposed in the frequency range of 125 to 8000 Hz. The Fast Fourier transform (FFT) is then performed on the decomposed ith sub-filters to obtain the ith frequency response $H_i(k)$ of these sub-filters.

Since each sub-filter contains N_h coefficients, the FFT's length is set to be equal to the next power of two from N_h in this algorithm. Because the human ear is not sensitive to the phase [30], only the room magnitude response $|H_i(k)|$ equalization is considered in the process of equalization. The ith frequency domain inverse filter $V_i(k)$ is the following:

$$V_i(k) = \frac{H_i^*(k)}{|H_i(k)|^2 + \beta}, \tag{10}$$

where $H_i^*(k)$ denotes the complex conjugate of $H_i(k)$ and β is a regularization index that is used to control the power output of the inverse filter [8].

The time-domain inverse sub-filters $v_i(k)$ are determined by computing the inverse FFT of the ith frequency domain inverse sub-filters $V_i(k)$. A "cyclic shift" of the inverse FFT is used to implement a modeling delay [7] to obtain causal and stable time-domain inverse sub-filters. Since finite impulse response (FIR) filters are used to replace the length of the "true" inverse sub-filters during the computation, the window function is used for $v_i(k)$ to suppress aliasing in the time domain.

2.3 Synthesis of preprocessing speech signals

In the speech enhancement stage, the signal is decomposed into 40 ERB-spaced filters between 125 and 8000 Hz. For accurate combination of the two preprocessing stages, the same decomposition is also performed in the equalization stage. The enhanced speech units $\alpha_{m,i}s_{m,i}$ and the inverse sub-filters v_i are obtained by speech enhancement and the GT-filter-based inverse filtering method, respectively. A block diagram of the preprocessing speech frame synthesis is illustrated in Fig. 2.

In the process of speech frame synthesis, the 40 decomposed and enhanced speech units and the 40 inverse sub-filters are reconstructed by sub-filter synthesis. These can be simply represented as $x_m = \sum_{i=1}^{40} \alpha_{m,i} s_{m,i}$ and $v = \sum_{i=1}^{40} v_i$, where x_m and v are the enhanced speech frame and the inverse filter, respectively. The FFT is then performed on x_m and v to realize the frequency domain transform. Therefore, the synthesized preprocessing speech frame of the frequency domain can be represented as $Y_m = X_m \times V$, and the inverse FFT is performed on Y_m to obtain the time-domain preprocessing speech frame y_m. Finally, the output speech $s_{\text{out}}(n)$ can be represented as follows by overlap addition of the preprocessing speech frames:

$$s_{\text{out}}(n) = \sum_{m=1}^{p} y_m(n). \tag{11}$$

Hanning analysis and synthesis windowing are used with 50% overlap.

3 Experiment implementation

A SISO audio system was established to simulate the I-PA system and applied in real environments to validate the proposed algorithm. To obtain data in different noisy reverberant environments, the experiments were performed using different rooms, types of noise, and signal-to-noise ratios (SNRs).

3.1 Experimental design

Four rooms with different reverberation times (RTs) were used to examine the influence of reverberation on speech intelligibility. The detailed parameters of the rooms are presented in Table 1. To study the influence of background noise, an additional omnidirectional loudspeaker was added near the position of the measuring microphone to simulate background noise. Four different types of background noise (white noise, factory noise-I, factory noise-II, and babble noise) were selected from the NOISE-92 database as the noise signals [31]. Each type of noise was divided into six different levels of SNR (– 10, – 5, 0, 5, 10, and 20 dB) to investigate the changes in speech intelligibility under different noise levels.

In this experiment, the volume of the loudspeaker was adjusted to keep the sound pressure level (SPL) of the listener position at 60 dB. To simplify the experimental system, it was assumed that the input speech from the far end is clear speech without any distortion. Furthermore, clear female speech with a sampling frequency of 16,000 Hz was randomly selected from the TIMIT database [32] as the input signal.

3.2 Hardware setup

Professional acoustic equipment was used to ensure the accuracy of the test results. Figure 3 shows the layout

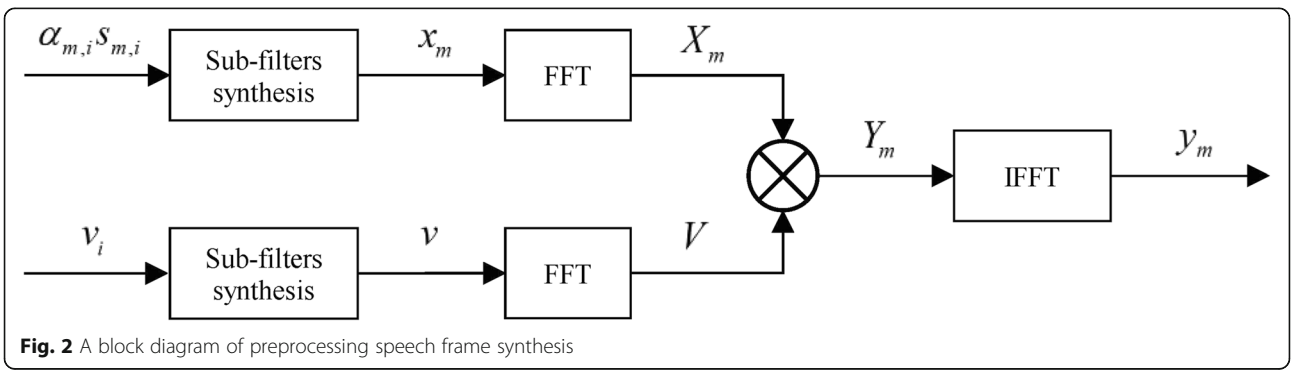

Fig. 2 A block diagram of preprocessing speech frame synthesis

Table 1 Information about the four test rooms

Room type	Room size (m)			Volume (m^3)	Temperature (°C)	T_{60} (s)
	Length	Width	Height			
Anechoic chamber	8.4	7.2	6.0	363	22.7	< 0.08
Small classroom	8.5	7.5	3.2	204	23.2	0.65
Large classroom	12.0	9.0	3.2	346	23.4	1.39
Hall	16.5	11.5	9.0	1708	22.8	3.57

and equipment used in the four different rooms. A BSWA MPA201 free-field microphone was connected to a SCIEN ADC 3241 professional sound card through a BNC connector cable. Two INTERM L-2400 power amplifiers were also connected to two Brüel & Kjær high-power omnidirectional sound sources through a Speakon connector cable. The measurement equipment used in the experiment has a flat response curve for a frequency range of 100 to 16,000 Hz. A cell phone was used as a noise generator to control the output of the noise source, and a laptop was used to manage the other equipment.

The speech recording and computing were performed using MATLAB software. For the hardware layout in each room, the distance between the sound source and measuring microphone was set between 3 and 5 m based on the different sizes of the rooms. The noise source was set up on a different side from the measuring microphone at a distance of 1.5 m. The sound source, noise source, and microphone were all installed on a tripod with a relative distance of 1.5 m from the floor. To ensure the consistency and validity of the listening test samples, 640 speech signals were selected from the MRT database [33].

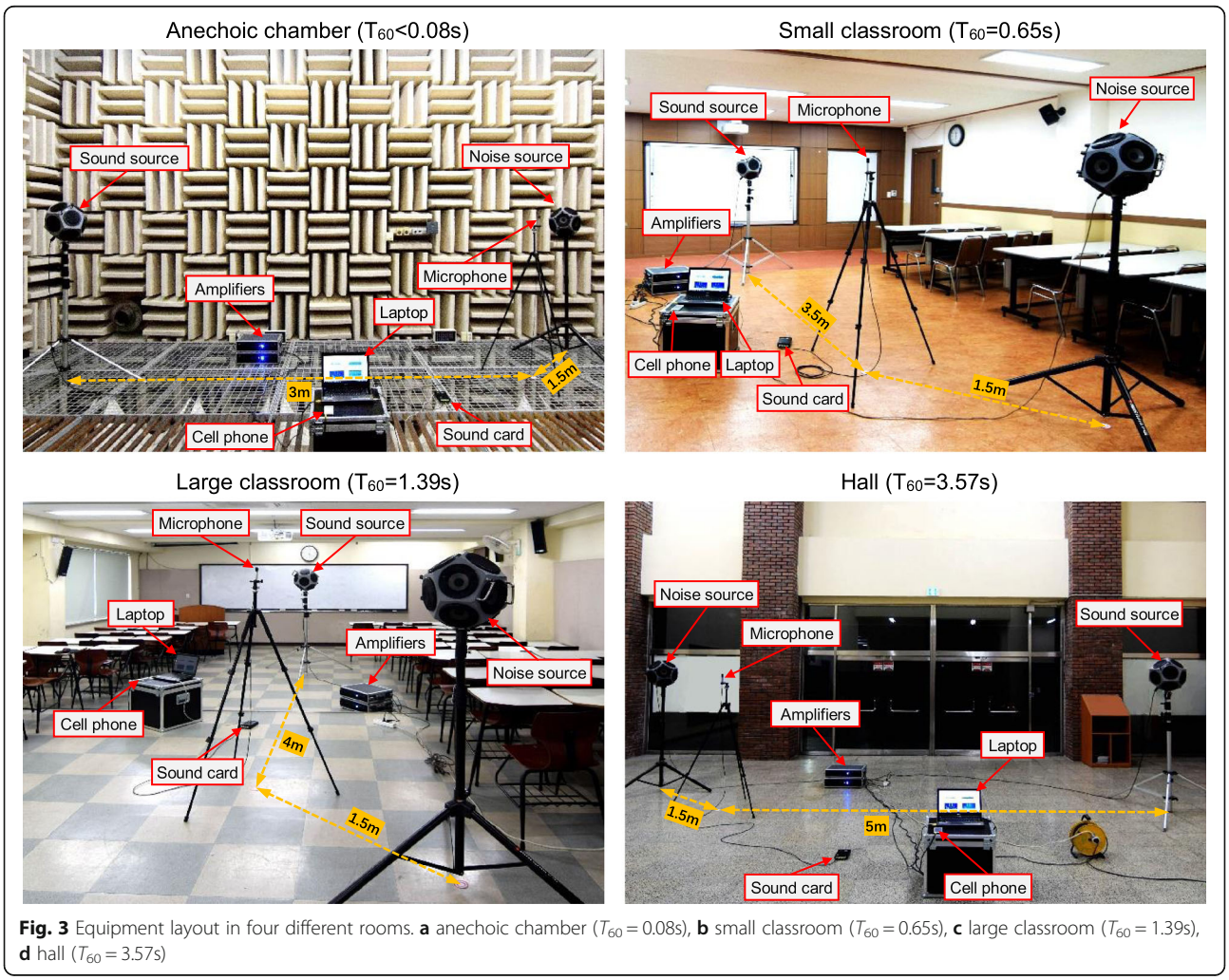

Fig. 3 Equipment layout in four different rooms. **a** anechoic chamber ($T_{60} = 0.08$s), **b** small classroom ($T_{60} = 0.65$s), **c** large classroom ($T_{60} = 1.39$s), **d** hall ($T_{60} = 3.57$s)

The signals were tested and saved in this experiment and later used as the test samples in the subjective evaluation.

4 Experimental results and discussion

A total of 576 conditions were tested in the real environments (4 noise types × 6 SNRs × 6 algorithms × 4 rooms). For conciseness, only the most representative experimental results are presented.

4.1 Objective results

Four kinds of measurements were performed to evaluate and compare the performance of the proposed method objectively. The speech transmission index was used to evaluate the dereverberation performance of the GT-filter-based FIF method. The spectrogram was used to visually display the changes of speech intelligibility before and after processing. The log-spectral distortion measure was used to compare the speech distortion of algorithms under different noise types, and the short-time objective intelligibility measure was used to predict and compare the changes of speech intelligibility of different algorithms.

4.1.1 Speech transmission index

The speech transmission index (STI) is a well-established objective measurement predictor that is used to evaluate the speech intelligibility of acoustic transmission channels [34]. It uses a series of complex calculations of RIR and values between 0 and 1 to represent the degree of speech intelligibility. Table 2 presents the subjective impression of the measured STI values [34]. The STI was used for comparison with previous results obtained with the FIF method [8], W-EQ method [12], and A-EQ method [6], as shown in Table 3.

The STI values with no equalized RIR decrease as RT increases. When RT increases to 3.57 s, the STI value decreases to 0.4, and the transmission channel is seriously distorted due to reverberation. However, after equalizing the RIR by the proposed algorithm, the STI values are significantly improved. Compared with the other methods in Table 3, it is clear that the STI values for the proposed method are always higher than those of the other equalization or dereverberation methods. These results prove that the auditory-model-based sub-filter equalization can further improve the speech intelligibility of the transmission channel under different RT conditions.

Table 2 Evaluation standards of STI values according to ICE 60268-16

STI value	Subjective intelligibility impression
0.75–1.00	Excellent
0.60–0.75	Good
0.45–0.60	Satisfactory
0.30–0.45	Poor
0.00–0.30	Very poor

Table 3 Comparison of STI values of different methods

Methods	R1	R2	R3	R4
No equalized RIR	0.99	0.74	0.56	0.40
FIF method	0.99	0.85	0.79	0.71
W-EQ method	1.00	0.88	0.80	0.73
A-EQ method	0.99	0.81	0.74	0.66
Improved FIF method	1.00	0.92	0.84	0.79

R1, R2, R3, and R4 represent the anechoic chamber ($T_{60} < 0.8$ s), small classroom ($T_{60} = 0.65$ s), large classroom ($T_{60} = 1.39$ s), and hall ($T_{60} = 3.57$ s), respectively

4.1.2 Spectrogram

A spectrogram is a visual representation of frequencies of a sound signal as it varies with time [35]. It uses the distribution of different colors on the image to observe the changes of the sound signal. The spectrogram was used to visually demonstrate how noise and reverberation degrade the speech intelligibility and to compare the differences in speech intelligibility before and after using the method. The results are shown in Fig. 4. Compared with Fig. 4a, the noisy speech in Fig. 4b is masked by the white noise at an SNR of − 5 dB, resulting in lost speech information over 2000 Hz. In Fig. 4c, it is clear that the speech signal becomes blurry because of reverberation. Smearing effects [36] also occurred at the end of each speech frame, resulting in a reduction in speech intelligibility.

Figure 4d shows noisy reverberant speech degraded by white noise and reverberation simultaneously. The degraded speech loses the speech information over 2000 Hz, and the speech information of the remaining part is quite blurry. Figure 4e shows the speech signals obtained by the proposed method in a noisy reverberant environment. Compared with the noisy reverberant speech in Fig. 4d, the speech frames are independent of each other without smearing effects after applying the proposed method. Compared with the clean speech in Fig. 4a, the processed speech has not lost any important speech information. Therefore, the comparison results intuitively show that the proposed method can significantly improve the speech intelligibility in noisy reverberant environments.

4.1.3 Log-spectral distortion measure

The log-spectral distortion (LSD) is an established and straightforward speech distortion measure. It computes the difference of the root-mean-square (RMS) values between the clean speech and the test signal to show the extent of distortion of the test signal. The LSD can be used to evaluate the performance of various speech enhancement algorithms in a noisy environment and is moderately well suited for the assessment of dereverberation algorithms in

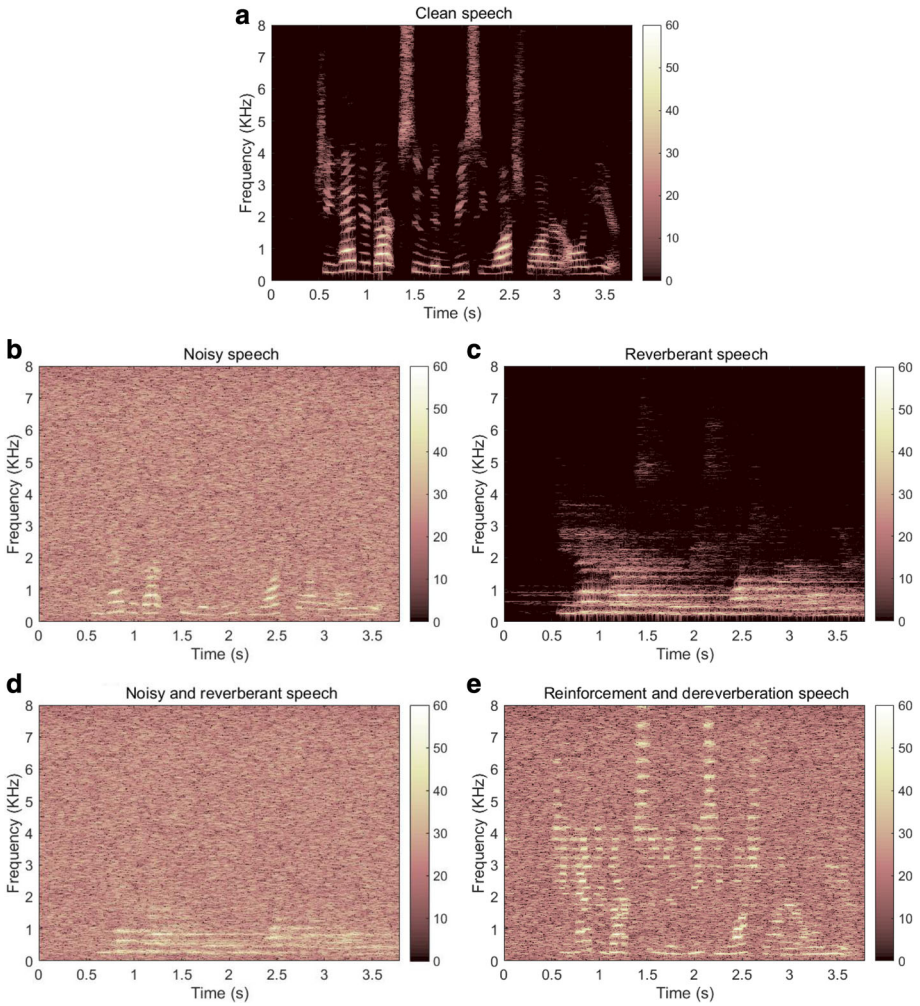

Fig. 4 Spectrogram comparison. **a** clean speech, **b** noisy speech (SNR = − 5 dB), **c** reverberant speech (T_{60} = 3.57 s), **d** noisy and reverberant speech (SNR = − 5 dB, T_{60} = 3.57 s), and **e** reinforcement and dereverberation speech (SNR = − 5 dB, T_{60} = 3.57 s)

cases of reverberation [37]. The LSD was used to measure the distortion of test signals obtained from the experiment and is defined as [38]:

$$\mathrm{LSD}(l) = \left(\frac{2}{N} \sum_{n=0}^{\frac{N}{2}-1} |L\{X(l,n)\} - L\{S(l,n)\}|^2 \right)^{\frac{1}{2}}, \quad (12)$$

where $X(l,n)$ and $S(l,n)$ are the FFT-based short-time spectra of the test speech signal and clean speech signal, respectively. l is the time frame, and n is the length of the FFT. Each of the frames is set to be 35 ms long, and Hamming analysis and synthesis windowing are used with 60% overlap. $L\{X(l,n)\} = \max\{20 \log_{10}(|X(l,n)|), \delta\}$ is the log spectrum confined to a dynamic range of about 50 dB ($\delta = \max_{l,n}\{20\log_{10}(|X(l,n)|)\} - 50$), and $L\{S(l,n)\}$ has a similar definition to $L\{X(l,n)\}$. The mean LSD is obtained by averaging over all frames.

In the LSD evaluation, four types of noise were considered to validate the proposed algorithm. For each type of noise, a total of 96 LSD test results were used, including four kinds of algorithms, as illustrated in Fig. 5. It is clear that the 3D plots for each type of noise have similar tendencies in that the LSD values of the FIF method [8], and the PDMSE method [17] have large fluctuations with changes in RT and SNR. However, the LSD values of the proposed method maintain a stable downward tendency. These results show that the individual FIF and PDMSE methods cannot reduce the speech distortion steadily in various SNR and RT conditions, in contrast to the proposed method.

4.1.4 Short-time objective intelligibility measure

A short-time objective intelligibility measure (STOI) is a method of obtaining intelligibility scores directly by analyzing the clean and processed signals [39]. It yields high

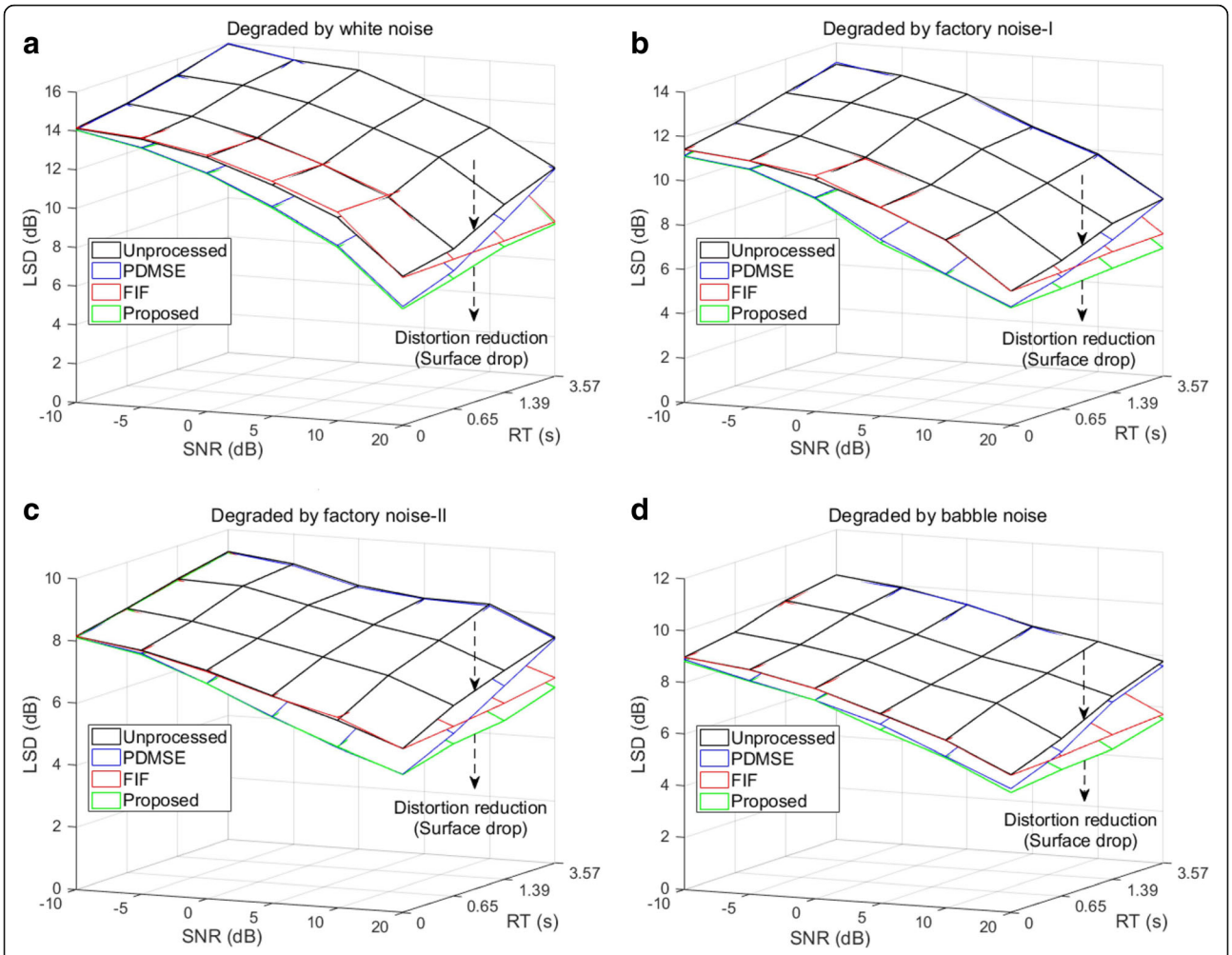

Fig. 5 Log-spectral distortion comparison of algorithms under the different noise types. **a** speech degraded by white noise, **b** speech degraded by factory noise-I, **c** speech degraded by factory noise-II, **d** speech degraded by babble noise

correlations with subjective listening results and is usually used to evaluate the intelligibility of denoised speech [40]. The objective speech intelligibility is more meaningful than the LSD measure for investigating the effectiveness of the proposed method. No unified objective intelligibility evaluation standards have been designed to predict distortions caused by additive noise and reverberation simultaneously. Nevertheless, we still attempted to use the STOI measure to predict the changes of speech intelligibility objectively.

The test data under factory noise-II conditions were selected for the intelligibility prediction of this part. Figure 6 shows that the STOI measures were monotonically decreased with increasing RT and decreasing SNR. Compared with the unprocessed speech, the other three methods significantly improved speech intelligibility under all test conditions. The performance of the Multizone approach [21] and ASII approach [22] was almost identical, and the improvement by the ASII approach was slightly higher than that of the Multizone approach under long reverberation conditions. However, compared with these two methods, the speech intelligibility was further improved by the proposed method.

There is no literature to support that the STOI measures can be used for the intelligibility evaluation of reverberant speech [21]. Therefore, the STOI measure was merely used to predict the intelligibility trends of different algorithms. However, compared with the results of the subjective listening test in Section 4.2, it is clear that the STOI prediction results have highly consistent trends with the subjective evaluation results. Therefore, the STOI prediction can be regarded as a meaningful reference result among the various objective evaluations.

4.2 Subjective results

The modified rhyme test (MRT) [33] was used for subjective and realistic evaluation of the speech intelligibility in a noise and reverberation environment.

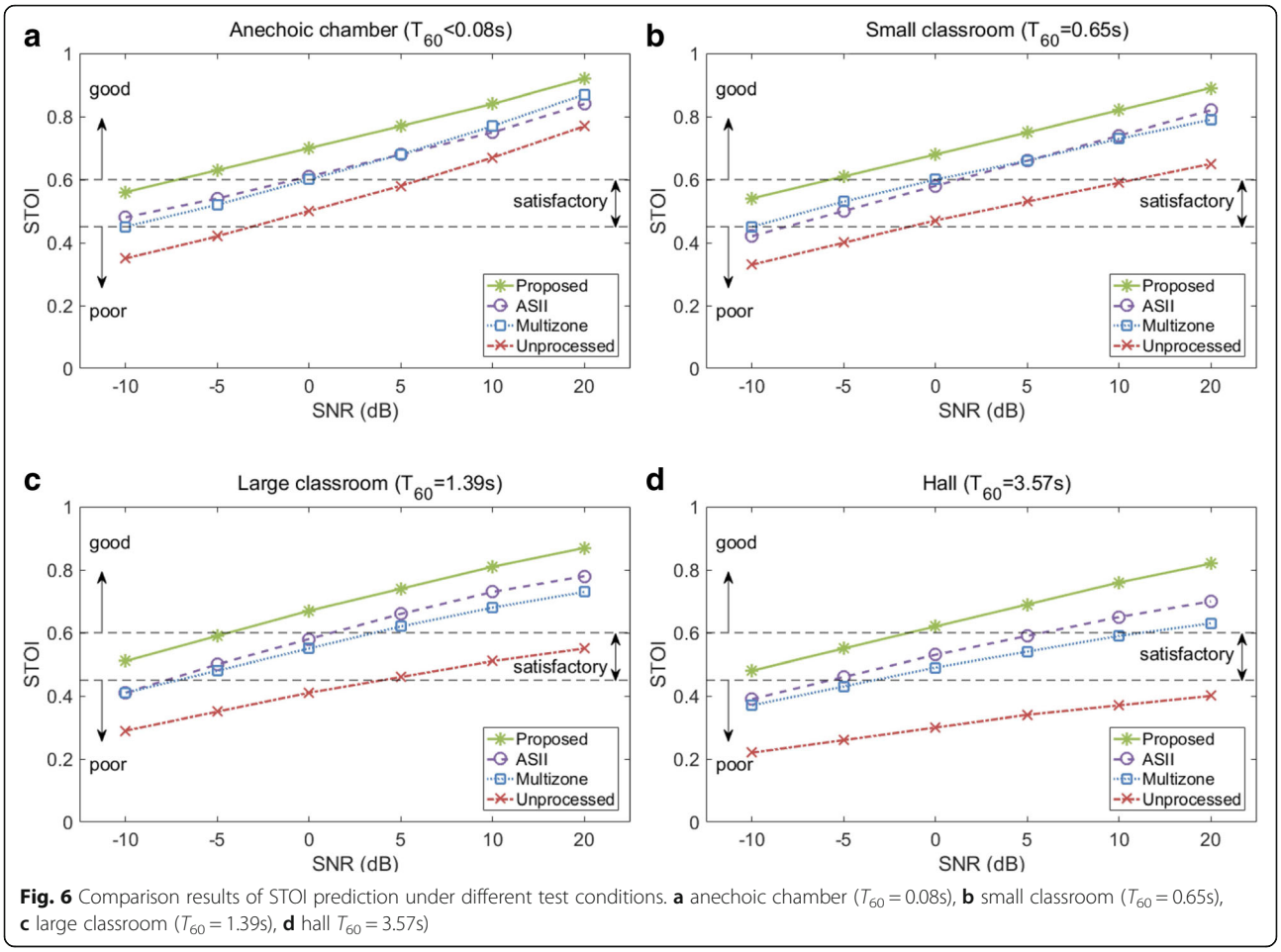

Fig. 6 Comparison results of STOI prediction under different test conditions. **a** anechoic chamber ($T_{60} = 0.08$s), **b** small classroom ($T_{60} = 0.65$s), **c** large classroom ($T_{60} = 1.39$s), **d** hall $T_{60} = 3.57$s)

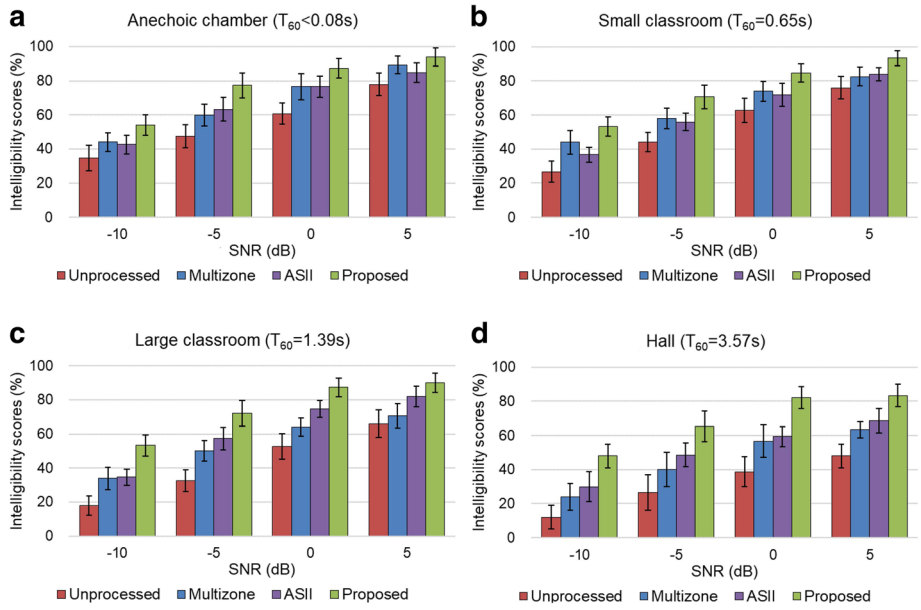

Fig. 7 Results of listening test under different RT and SNRconditions. **a** anechoic chamber ($T_{60} = 0.08$s), **b** small classroom ($T_{60} = 0.65$s), **c** large classroom ($T_{60} = 1.39$s), **d** hall $T_{60} = 3.57$s)

The MRT database contains a total of 2700 audio source files, including five males and four females reading 300 words. The 300 words read by each person are divided into 50 six-word groups of rhyming or similar-sounding English words, such as "same," "name," "game," "tame," "came," and "fame." Each word is a monosyllable of the form consonant-vowel-consonant (CVC), and the six words in each list differ in only the leading or trailing consonant. In this listening test, a total of 640 audio source files (4 RTs × 4 SNRs × 4 algorithms × 10 groups) were randomly selected from the database, modified using the four different of methods, and degraded by factory noise-II at SNRs of − 10, − 5, 0, and 5 dB in different reverberation conditions. This procedure was performed in the experiment described in Section 3, and the processed audio files were recorded by a laptop as test speech for the subjective evaluation.

Eighteen non-native English speakers (including 13 males and 5 females age 23 to 32) were invited to the listening test. All the listeners were knowledgeable of the English pronunciation and had no hearing impairments. Importantly, all of the listeners were Master's or Ph.D. students with a technical background in acoustics, and they were familiar with the basic concepts of reverberation and noise. The subjective tests were carried out in an anechoic chamber to prevent the effects of background noise and reverberation on the test speech. The same loudspeaker used in the experiment was also used in the listening test, and the volume of the loudspeaker was adjusted to keep the output SPL within the normal hearing range.

Before the listening test, some training samples were presented to the listeners to familiarize them with the test procedure. The audio files were played randomly for the different algorithms, RTs, and SNRs.

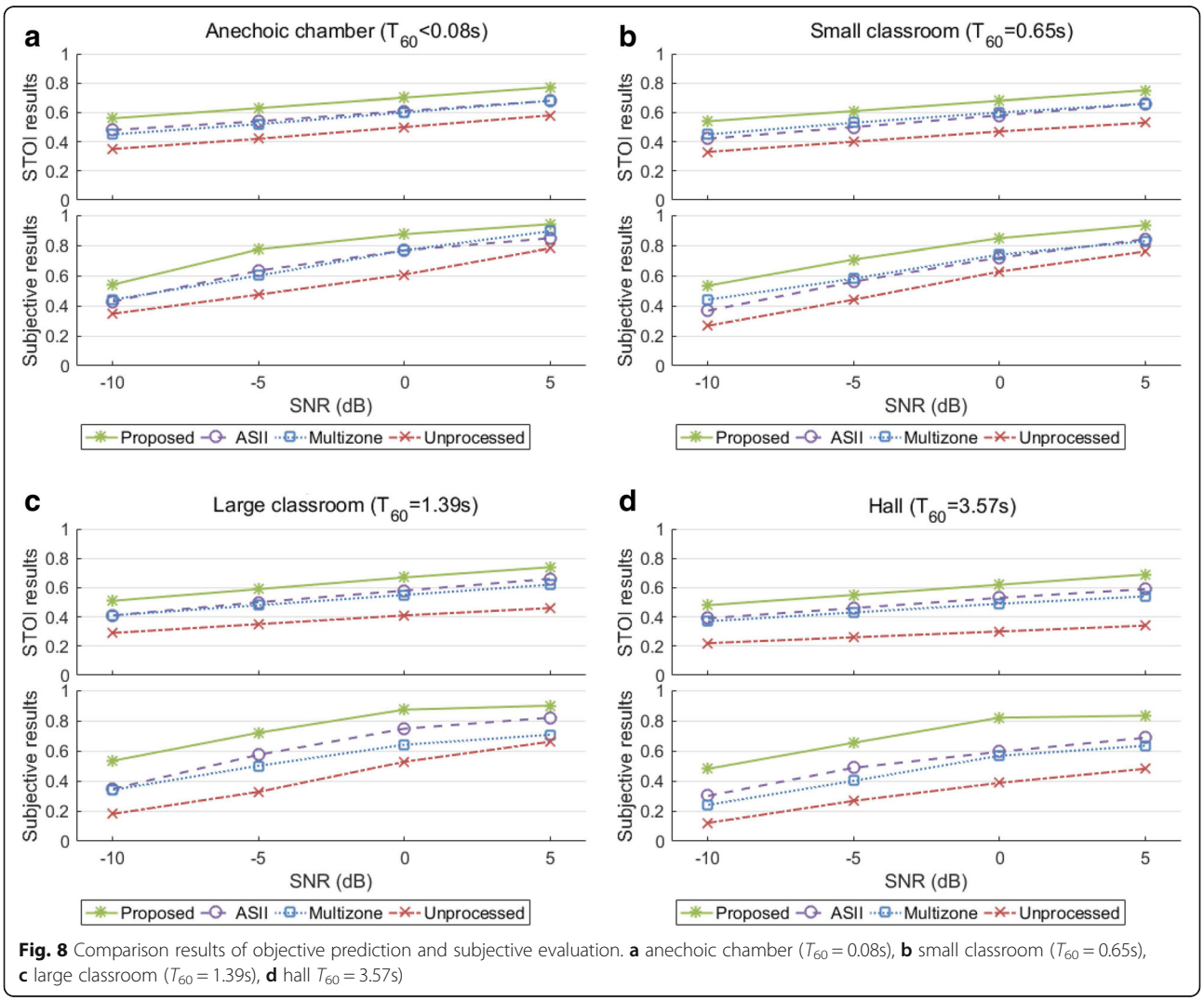

Fig. 8 Comparison results of objective prediction and subjective evaluation. **a** anechoic chamber ($T_{60} = 0.08$s), **b** small classroom ($T_{60} = 0.65$s), **c** large classroom ($T_{60} = 1.39$s), **d** hall $T_{60} = 3.57$s)

Each sentence was played only once, and the listener had 5 s to choose the right answer from a set of six alternative words on the response sheet. The intelligibility score of different algorithms under various SNR and RT conditions was obtained as the mean percentage of correct words.

To determine the statistical significance, the confidence intervals were calculated with a significance level of 0.05. Figure 7 shows the mean scores of the algorithms under the different test conditions and the corresponding confidence intervals as vertical colored blocks and vertical black lines, respectively. The intelligibility score of the proposed method has a significant improvement over the unprocessed speech under all test conditions compared with that of the Multizone approach [21] and ASII approach [22]. It is clear that the proposed method always has higher intelligibility scores than the other two approaches. However, the tendency of the intelligibility score of these two comparison approaches is not stable, so it is difficult to say which approach is better. In contrast, the proposed method can steadily and effectively improve the speech intelligibility in different noisy reverberant environments.

To observe the difference between the objective and subjective evaluation results, the STOI prediction results and the listening test results were compared at SNRs of − 10, − 5, 0, and 5 dB conditions, as illustrated in Fig. 8. The results of the two evaluations had slightly different numerical values under the same test conditions. However, it is important that the objective and subjective evaluation results of different algorithms showed quite similar trends under all test conditions.

5 Conclusions

A speech preprocessing method that combines the modified PDMSE and the improved FIF was proposed to improve the speech intelligibility of I-PA systems in noisy reverberant environments. The combination method reduces noise masking by means of speech enhancement and eliminates the influence of reverberation by means of transmission channel equalization. The experimental results showed that the speech intelligibility is significantly improved in noisy reverberant environments by the proposed method.

Compared with individual PDMSE and FIF methods, the combination method can stably reduce speech distortion under various noisy reverberant conditions. Furthermore, the subjective listening tests confirmed the validity and stability of the proposed method, and its mean intelligibility score was higher than those of state-of-the-art reference algorithms. Future work will focus on a method to obtain RIR in real time under noisy reverberant environments to realize real-time and steady improvement of speech intelligibility in variable room boundary conditions.

Acknowledgements
This work was supported by a 2017 grant from the Russian Science Foundation (Project No. 17-19-01389).

Authors' contributions
CML gave academic guidance to this research work and revised the manuscript. HYD designed the core methodology of this study, programmed the algorithms and carried out the experiments, and drafted the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 24 October 2017 Accepted: 24 April 2018
Published online: 23 May 2018

References

1. Taal, CH, Hendriks, RC, Heusdens, R (2012). *A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure*, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4061–4064).
2. Crespo, JB, & Hendriks, RC (2014). *Speech reinforcement in noisy reverberant environments using a perceptual distortion measure*, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 910–914).
3. Miyoshi, M, & Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Process*, *36*(2), 145–152.
4. Maganti, HK, & Matassoni, M. (2012). A perceptual masking approach for noise robust speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, *2012*(1), 29.
5. Neely, ST, & Allen, JB. (1979). Invertibility of a room impulse response. *The Journal of the Acoustical Society of America*, *66*(1), 165–169.
6. Elliott, SJ, & Nelson, PA. (1989). Multiple-point equalization in a room using adaptive digital filters. *Journal of the Audio Engineering Society*, *37*(11), 899–907.
7. Mourjopoulos, JN. (1994). Digital equalization of room acoustics. *Journal of the Audio Engineering Society*, *42*(11), 884–900.
8. Tokuno, H, Kirkeby, O, Nelson, PA, et al. (1997). Inverse filter of sound reproduction systems using regularization. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *80*(5), 809–820.
9. Kirkeby, O, Nelson, PA, Hamada, H, et al. (1998). Fast deconvolution of multichannel systems using regularization. *IEEE Transactions on Speech and Audio Processing*, *6*(2), 189–194.
10. Kirkeby, O, & Nelson, PA. (1999). Digital filter design for inversion problems in sound reproduction. *Journal of the Audio Engineering Society*, *47*(7/8), 583–595.
11. Radlovic, BD, & Kennedy, RA. (2000). Nonminimum-phase equalization and its subjective importance in room acoustics. *IEEE Transactions on Speech and Audio Processing*, *8*(6), 728–737.
12. Cecchi, S, Romoli, L, Carini, A, et al. (2014). A multichannel and multiple position adaptive room response equalizer in warped domain: real-time implementation and performance evaluation. *Applied Acoustics*, *82*, 28–37.
13. Mourjopoulos, J, Clarkson, P, Hammond, J (1982). *A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals*, *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)* (pp. 1858–1861).
14. Fuster, L, de Diego, M, Ferrer, M, et al. (2012). *A biased multichannel adaptive algorithm for room equalization*, *In Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (pp. 1344–1348).
15. B Sauert, P Vary, Improving speech intelligibility in noisy environments by near end listening enhancement. ITG-Fachbericht-Sprachkommunikation. (2006)
16. Skowronski, MD, & Harris, JG. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, *48*(5), 549–558.
17. Taal, CH, Hendriks, RC, Heusdens, R. (2014). Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Computer Speech & Language*, *28*(4), 858–872.
18. Taal, C, & Heusdens, R (2009). *A low-complexity spectro-temporal based perceptual model*, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 153–156).

19. Kusumoto, A, Arai, T, Kinoshita, K, et al. (2005). Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication*, *45*(2), 101–113.
20. Hodoshima, N, Arai, T, Kusumoto, A, et al. (2006). Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments. *The Journal of the Acoustical Society of America*, *119*(6), 4055–4064.
21. Crespo, JB, & Hendriks, RC. (2014). Multizone speech reinforcement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(1), 54–66.
22. Hendriks, RC, Crespo, JB, Jensen, J, et al. (2015). Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(5), 851–862.
23. Meng, Q, Sen, D, Wang, S, et al. (2008). *Impulse response measurement with sine sweeps and amplitude modulation schemes*, *IEEE 2nd International Conference on Signal Processing and Communication Systems(ICSPCS)* (pp. 1–5).
24. Hendriks, RC, Heusdens, R, Jensen, J (2010). *MMSE based noise PSD tracking with low complexity*, *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4266–4269).
25. Faraji, N, & Hendriks, RC (2012). *Noise power spectral density estimation for public address systems in noisy reverberant environments*, *In Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)* (pp. 1–4).
26. Peng, W, Ser, W, Zhang, M (2001). *Bark scale equalizer design using warped filter*, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3317–3320).
27. Moore, BC, & Glasberg, BR. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*(3), 750–753.
28. Maganti, HK, & Matassoni, M. (2014). Auditory processing-based features for improving speech recognition in adverse acoustic conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(1), 21.
29. Patterson, RD (1986). *Auditory filters and excitation patterns as representations of frequency resolution. Frequency selectivity in hearing*, (pp. 123–177).
30. Toole, FE (2000). *The acoustics and psychoacoustics of loudspeakers and rooms—the stereo past and the multichannel future*, *Audio Engineering Society Convention* (p. 109).
31. Varga, A, & Steeneken, HJ. (1993). Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251.
32. Zue, V, Seneff, S, Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, *9*(4), 351–356.
33. Miner, R, & Danhauer, JL. (1975). Modified rhyme test and synthetic sentence identification test scores of normal and hearing-impaired subjects listening in multitalker noise. *Journal of the American Audiology Society*, *2*(2), 61–67.
34. International Electrotechnical Commission. (2011). IEC 60268-16 Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index. Paris: International OECD Publishing.
35. Flanagan, JL (2013). *Speech analysis synthesis and perception*, (vol. 3). New York: Springer Science & Business Media p. 150.
36. Gomez, R, Nakamura, K, Mizumoto, T, et al. (2013). *Mitigating the effects of reverberation for effective human-robot interaction in the real world*, *13th IEEE International Conference on RAS, Humanoid Robots (Humanoids)* (pp. 177–182).
37. Habets, EAP. (2007). *Single-and multi-microphone speech dereverberation using spectral enhancement*. Dissertation Abstracts International, *68*(04).
38. Naylor, PA, & Gaubitch, ND (2010). *Speech dereverberation*, (p. 40). New York: Springer Science & Business Media.
39. Taal, CH, Hendriks, RC, Heusdens, R, et al. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136.
40. Loizou, PC (2013). *Speech enhancement: theory and practice*, (2nd ed., pp. 552–567). Boca Raton: CRC Press.