Speech Communication 96 (2018) 129–141

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A perceptually motivated LP residual estimator in noisy and reverberant environments

Renhua Peng[a,c], Zheng-Hua Tan[b], Xiaodong Li[a,c], Chengshi Zheng[*,a,c]

[a] Key Laboratory of Noise and Vibration Control, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China
[b] Department of Electronic Systems, Aalborg University, Aalborg, 9220, Denmark
[c] University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing, 100049, China

ARTICLE INFO

Keywords:
Generalized singular value decomposition
Minimum mean square error
Auditory masking
Linear prediction residual
Speech dereverberation

ABSTRACT

Both reverberation and additive noise can degrade the quality of recorded speech and thus should be suppressed simultaneously. Previous studies have shown that the generalized singular value decomposition (GSVD) has the capability of suppressing the additive noise effectively, but it is not often applied for speech dereverberation since reverberation is considered to be convolutive as well as colored noise. Recently, we revealed that late reverberation is also additive and relatively white interference component in the linear prediction (LP) residual domain. To suppress both late reverberation and additive noise, we have proposed an optimal filter for LP residual estimator (LPRE) based on a constrained minimum mean square error (CMMSE) by using GSVD in single channel speech enhancement, where the algorithm is referred as CMMSE-GSVD-LPRE. Experimental results have shown a better performance of the CMMSE-GSVD-LPRE than spectral subtraction methods, but some residual noise and reverberation components are still audible and annoying. To solve this problem, this paper incorporates the masking properties of the human auditory system in the LP residual domain to further suppress these residual noise and reverberation components while reducing speech distortion at the same time. Various simulation experiments are conducted, and the results show an improved performance of the proposed algorithm. Experimental results with speech recorded in noisy and reverberant environments further confirm the effectiveness of the proposed algorithm in real-world environments.

1. Introduction

In hands-free communication systems, such as hearing aids, mobile phones and voice-controlled systems, it is well-known that both room reverberation and additive noise can significantly deteriorate the perceived quality and intelligibility of speech captured by a microphone in a closed room (Benesty and Makino, 2005; Naylor and Gaubitch, 2010; Loizou, 2013), especially when the desired talker is far away from the microphone. A listener who is sensorineurally impaired will have extra difficulty in perceiving and understanding the deteriorated speech (Bloom, 1980; Bloom and Cain, 1982). For automatic speech recognition (ASR) systems, it has been shown that late reverberation can degrade the performance of ASR severely (Sehr et al., 2010; Yoshioka et al., 2012). In order to obtain a satisfactory communication system both for human-to-human and human-to-machine interactions, speech dereverberation and noise reduction are fundamentally important. For the last half-century, many effective algorithms have been proposed to deal with reverberation (Lebart et al., 2001; Habets, 2005; Wu and Wang, 2006; Nakatani et al., 2008; Habets et al., 2009; Jeub et al., 2010), noise (Boll, 1979; Cohen and Berdugo, 2002; Martin, 2001; Cohen, 2003; Zheng et al., 2010; Gerkmann and Hendriks, 2012), or both (Jensen and Tan, 2015; Kun et al., 2015).

For noise reduction, the additive noise is often assumed to be uncorrelated with the source signal. One of the methods based on this assumption is spectral subtraction (SS) that was first proposed by Boll (1979), which is the most popular for its simplicity of implementation. The noise power spectral density (NPSD) estimation is the key step for this type of methods, and numerous state-of-the-art NPSD estimators have already been proposed in the literature (Cohen and Berdugo, 2002; Martin, 2001; Cohen, 2003; Zheng et al., 2010; Gerkmann and Hendriks, 2012). It's well-known that SS methods still suffer from the so-called '*musical noise*' problem, which is composed of tones at randomly distributed frequencies. Various algorithms have been proposed to reduce '*musical noise*', including the over-subtraction of noise and the introduction of a spectral floor (Berouti et al., 1979), the optimal minimum mean-square error (MMSE) estimation of the short-time spectral amplitude (Ephraim and Malah, 1984), and the incorporation of human auditory properties (Virag, 1999).

* Corresponding author at: Key Laboratory of Noise and Vibration Control, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China.
E-mail address: cszheng@mail.ioa.ac.cn (C. Zheng).

https://doi.org/10.1016/j.specom.2017.12.004
Received 5 July 2016; Received in revised form 13 August 2017; Accepted 4 December 2017
Available online 06 December 2017
0167-6393/

Since reverberation is considered as a source signal convoluting with a room impulse response (RIR), its characteristics are quite different from the additive noise. Speech dereverberation algorithms can be generally categorized into three main classes, such as inverse filtering methods (Miyoshi and Kaneda, 1988; Radlovic et al., 2000), cepstral subtraction (CS) methods (Bees et al., 1991; Subramaniam et al., 1996), and late reverberation suppression methods incorporating SS (Lebart et al., 2001; Wu and Wang, 2006; Kinoshita et al., 2009). Inverse filtering methods aim at deconvoluting the RIR and restoring the original source signal. However, inverse filtering methods are often sensitive to environmental noise (Neely and Allen, 1979), the fluctuation of the RIR (Mourjopoulos, 1985), and multiple desired speakers (Rotili et al., 2011). Different solutions have been proposed to cope with these problems (Neely and Allen, 1979; Tokuno et al., 1997; Hikichi et al., 2007), and most of them are based on regularization theory (Tokuno et al., 1997). Although inverse filtering methods can achieve satisfactory results, they are not suitable for real-time implementation due to their high computational load. To reduce the computation load, Bees et al. proposed to perform the deconvolution using cepstral analysis (Bees et al., 1991; Subramaniam et al., 1996). In the cepstral domain, the convolution operation is converted to an addition operation, and the deconvolution can be realized by CS. In contrast to inverse filtering and CS methods, many robust and practical approaches have been proposed to mitigate late reverberation only (Lebart et al., 2001; Wu and Wang, 2006; Kinoshita et al., 2009), as late reverberation has been shown to be the main reason for speech quality and recognition performance degradation (Naylor and Gaubitch, 2010). Generally, late reverberation is considered to be uncorrelated with early reverberation and the source signal. Therefore, these approaches aim at estimating the late reverberation spectral variance (LRSV) and then subtracting the estimated LRSV from the reverberant signal by using SS methods (Lebart et al., 2001). Examples of state-of-the-art LRSV estimators are the statistical model of RIR based methods (Habets et al., 2009), multiple-step linear prediction based methods (Kinoshita et al., 2009), and the smearing effect of late reverberation based methods (Wu and Wang, 2006). One can find other speech dereverberation and enhancement algorithms by temporal and spectral processing in Krishnamoorthy and Prasanna (2009); 2011).

Once the NPSD and the LRSV are estimated, SS methods are generally implemented to suppress additive noise and late reverberation simultaneously. In our previous work (Zheng et al., 2014), we investigated the signal subspace approach (SSA), that was originally proposed to suppress noise for noisy speech (Ephraim and Van Trees, 1995). The SSA method is based on the decomposition of the noisy signal space into two orthogonal subspaces called the noise subspace and the signal subspace. Signal enhancement is performed by removing the noise subspace, and then estimating the source signal from the remaining signal subspace. The signal decomposition can be achieved by the Karhunen–Loeve transformation (KLT) (Mittal and Phamdo, 2000; Rezayee and Gazor, 2001), the singular value decomposition (SVD) (Jensen et al., 1995), or the generalized singular value decomposition (GSVD) (Doclo and Moonen, 2002). Most of the traditional GSVD-based methods are proposed to reduce noise in the time domain directly (Doclo and Moonen, 2002; Yoshioka et al., 2009; Löllmann and Vary, 2009; Spriet et al., 2002), and others are applied to dereverberate the speech signal by estimating the RIR functions using multiple microphones (Gannot and Moonen, 2003). In Zheng et al. (2014), we proposed to apply the GSVD-based method for noise reduction and dereverberation in the LP residual domain, in which we show that both late reverberation and ambient noise are additive in the LP residual domain. A constrained MMSE LP residual estimator (LPRE) was introduced to suppress both late reverberation and additive noise at the same time by using GSVD, where the algorithm is referred as CMMSE-GSVD-LPRE algorithm. Although the CMMSE-GSVD-LPRE algorithm is superior to SS methods and the traditional GSVD-based methods, some residual noise and reverberation components are still perceivable under low signal-to-noise ratio (SNR) or low direct-to-reverberation ratio (DRR) regions, as the permissible residual noise and reverberation are not optimized (see Zheng et al., 2014 for details).

In this paper, we extend our work presented in Zheng et al. (2014) and propose to use the auditory masking properties to control the level of the residual noise and reverberation for single channel speech enhancement. Note that the auditory masking properties have already been well defined and studied in both time and frequency domains (Schroeder et al., 1979; Thiemann, 2001). Because the effect of frequency masking is much more dominant than that of time masking, we will focus on the frequency masking effect in this paper. However, the auditory masking threshold (AMT) in frequency domain can not be applied in signal subspace directly. To solve this problem, Jabloun and Champagne proposed a frequency domain to eigenvalue domain transformation (FET), which provides a way to calculate a perceptual upper bound for the residual noise (Jabloun and Champagne, 2002), and this was extended to the generalized singular value domain by Ju and Lee (2007). Whereas, FET can not be used in the LP residual domain either. Therefore, we need to study a new transformation to calculate the perceptual upper bound for the LP residual noise and reverberation in this paper.

The remainder of this paper is organized as follows. Section 2 formulates the problem and briefly introduces the Wiener optimal filtering and the GSVD. Section 3 presents the constrained MMSE LP estimator and the proposed perceptually motivated optimal filter, where the transformation from perceptual constrains to subspace values is reformulated in the linear prediction residual domain. Simulation and realistic experiments are given in Sections 4 and 5, respectively, with objective and subjective evaluation results. Finally, some conclusions are made in Section 6.

2. Problem formulation and GSVD-based optimal filtering

When placed at a certain distance from the talker in a closed room, the microphone not only acquires the direct sound, but also the reflected sounds, which are the delayed and modulated versions of the direct sound. Thus, reverberation can be modeled as the source signal convoluting with the RIR. Taking the environmental noise into consideration, the microphone signal is given by

$$\begin{aligned} x(n) &= s(n)*h(n) + v(n) \\ &= \sum_{i=0}^{L_h-1} s(n-i)h(i) + v(n) \\ &= y(n) + v(n) \end{aligned} \tag{1}$$

where $s(n)$ is the clean speech signal, $h(n)$ is the RIR from the talker to the microphone, which is modeled by a finite impulse response (FIR) filter with length L_h, and '*' is the convolution operator. It is assumed that $h(n)$ is time-invariant, and $v(n)$ is the additive noise. $y(n)$ is the reverberant speech signal without the additive noise.

Eq. (1) can be written in the vector multiplication form, which is

$$x(n) = \mathbf{s}_n^T\mathbf{h} + v(n) \tag{2}$$

where $\mathbf{s}_n = [s(n), s(n-1),...,s(n-L_h+1)]^T$ is the vector of the clean speech signal at time index n, and $\mathbf{h} = [h(0), h(1),...,h(L_h-1)]^T$ is the vector of the RIR coefficients, respectively. 'T' denotes the transpose operation.

2.1. Optimal filtering

Here, we want to find a filter $w(n)$ with length L_w, such that the filtered microphone signal

$$\begin{aligned} \tilde{d}(n) &= x(n)*w(n) \\ &= \mathbf{x}_n^T\mathbf{w} \end{aligned} \tag{3}$$

is an estimation of the desired signal $d(n)$ or its delayed version, where $\mathbf{w} = [w(0), w(1), ..., w(L_w - 1)]^T$ is the vector of the filter coefficients, and $\mathbf{x}_n = [x(n), x(n-1), ..., x(n - L_w + 1)]^T$ is the vector of the microphone signal, which can be written as

$$\mathbf{x}_n = \mathscr{H}_s^T \mathbf{h} + \mathbf{v}_n \tag{4}$$

where $\mathbf{v}_n = [v(n), v(n-1), ..., v(n - L_w + 1)]^T$ is the vector of the noise signal. $\mathscr{H}_s \in \mathbb{R}^{L_h \times L_w}$ is the Hankel matrix of the clean speech signal, which is given by

$$\mathscr{H}_s = \left[\mathbf{s}_n, \mathbf{s}_{n-1}, ..., \mathbf{s}_{n-L_w+1}\right]. \tag{5}$$

The estimated error signal $e(n)$ is defined as

$$e(n) = d(n) - \widetilde{d}(n). \tag{6}$$

Minimizing the mean square error (MSE) of $e(n)$, i.e., $E\{e^2(n)\}$, where $E\{\cdot\}$ represents the expectation, leads to an optimal filtering problem. If the reverberant signal $y(n)$ is chosen as the desired signal $d(n)$, then it is a noise reduction problem, where only the additive noise will be suppressed. If $s(n)$ is chosen as the desired signal $d(n)$, then it is a dereverberation and denoising problem, where both the reverberation and the additive noise will be suppressed. Here we focus on the second problem, i.e., speech dereverberation and denoising in noisy environments. It's well-known that the optimal filter to minimize the MSE cost function is the Wiener filter (Kalman, 1963), and the optimal filter $\mathbf{w}_{\text{opt}}$ is given by

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{xd}} \tag{7}$$

where $\mathbf{R}_{xx} = E\{\mathbf{x}_n\mathbf{x}_n^T\} \in \mathbb{R}^{L_w \times L_w}$ and $\mathbf{r}_{xd} = E\{d(n)\mathbf{x}_n\} \in \mathbb{R}^{L_w \times 1}$ are the $L_w \times L_w$ dimensional auto-correlation matrix of the microphone signal, and the $L_w \times 1$ dimensional cross-correlation vector between the microphone signal and the desired signal, respectively.

It can be seen that if $M > > 1$ holds, $\mathbf{R}_{xx}$ and $\mathbf{r}_{xd}$ can be estimated by

$$\widehat{\mathbf{R}}_{xx} = \frac{1}{M}\mathscr{H}_x\mathscr{H}_x^T, \hat{\mathbf{r}}_{xd} = \frac{1}{M}\mathscr{H}_x\mathbf{d}_n \tag{8}$$

where $\mathscr{H}_x \in \mathbb{R}^{L_w \times M}$ is the $L_w \times M$ Hankel matrix of the microphone signal, which is given by

$$\mathscr{H}_x = [\mathbf{x}_n, \mathbf{x}_{n-1}, ..., \mathbf{x}_{n-M+1}], \tag{9}$$

and $\mathbf{d}_n = [d(n), d(n-1), ...d(n - M + 1)]^T$ is the vector of the desired signal.

2.2. GSVD

The reverberant signal $y(n)$ can be rewritten as

$$\begin{aligned} y(n) &= \sum_{i=0}^{D-1} s(n-i)h(i) + \sum_{i=D}^{L_h-1} s(n-i)h(i) \\ &= s_{\text{er}}(n) + s_{\text{lr}}(n) \end{aligned} \tag{10}$$

where $s_{\text{er}}(n)$ and $s_{\text{lr}}(n)$ are the early and the late reverberation signals, respectively. D defines the boundary of the early and the late reverberation, which corresponds to the time ranges from 40 to 80 ms (Naylor and Gaubitch, 2010). Note that $s_{\text{er}}(n)$ includes the direct sound and the early reflected sound. According to Eq. (10), late reverberation is also an additive component in the time domain. Further, it is assumed that both late reverberation and additive noise are uncorrelated with the early reverberation signal. Based on these assumptions, we can apply both SS algorithm and the GSVD-based optimal filtering algorithm for dereverberation and noise reduction.

According to Naylor and Gaubitch (2010), the early reflected sound has an effect on increasing the strength of the direct-path sound and thus generate a positive impact on the intelligibility of speech. Therefore, the early reflected sound should not be suppressed for the purpose of human listening and only late reverberation should be. Assume that we can roughly estimate late reverberation $s_{\text{lr}}(n)$ and additive noise $v(n)$, and denote

$$t(n) = s_{\text{lr}}(n) + v(n) \tag{11}$$

as the interference that needs to be suppressed. The detailed steps to estimate $s_{\text{lr}}(n)$ and $v(n)$ will be given in the next section. Then the microphone signal can be rewritten as

$$x(n) = d(n) + t(n) \tag{12}$$

where $d(n) = s_{\text{er}}(n)$ is considered as the desired signal in this paper.

To employ the GSVD-based algorithm, the interference signal $t(n)$ needs to be firstly constructed in the Hankel matrix form as the same dimension as $\mathscr{H}_x$,

$$\mathscr{H}_t = [\mathbf{t}_n, \mathbf{t}_{n-1}, ..., \mathbf{t}_{n-M+1}] \tag{13}$$

where $\mathbf{t}_n = [t(n), t(n-1), ...t(n - L_w + 1)]^T$ is the vector of the interference signal.

A nonsingular matrix $\mathbf{Q} \in \mathbb{R}^{M \times M}$ and two real matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{L_w \times M}$, whose columns are orthogonal vectors, can be found to transform both $\mathscr{H}_x$ and $\mathscr{H}_t$ into nonnegative, bounded diagonal matrices $\mathbf{C}$ and $\mathbf{B}$ simultaneously

$$\mathbf{U}^T\mathscr{H}_x\mathbf{Q} = \mathbf{C} \tag{14}$$

$$\mathbf{V}^T\mathscr{H}_t\mathbf{Q} = \mathbf{B} \tag{15}$$

subjected to

$$\mathbf{C}^T\mathbf{C} + \mathbf{B}^T\mathbf{B} = \mathbf{I}_M \tag{16}$$

where $\mathbf{I}_M \in \mathbb{R}^{M \times M}$ is an identity matrix, $\mathbf{C} = \text{diag}\{c_1, c_2, ..., c_M\}$ and $\mathbf{B} = \text{diag}\{b_1, b_2, ..., b_M\}$. The diagonal elements of $\mathbf{C}$ and $\mathbf{B}$ are arranged in descending and ascending order, respectively.

The auto-correlation matrix of the microphone signal $\mathbf{R}_{xx}^M$ can also be estimated by

$$\widehat{\mathbf{R}}_{xx}^M = \frac{1}{L_w}\mathscr{H}_x^T\mathscr{H}_x = \frac{1}{L_w}\mathbf{Q}^{-T}\mathbf{C}^2\mathbf{Q}^{-1} \tag{17}$$

under the condition that $L_w > > 1$, where $\mathbf{R}_{xx}^M \in \mathbb{R}^{M \times M}$.

It should be pointed out that both $\widehat{\mathbf{R}}_{xx}$ in Eq. (8) and $\widehat{\mathbf{R}}_{xx}^M$ in Eq. (17) are the estimated auto-correlation matrix of the microphone signal, but with different dimensions. Eq. (17) gives the auto-correlation matrix with dimensions $M \times M$. The auto-correlation matrix of the interference signal $\mathbf{R}_{tt}^M$ can be estimated in a similar way as

$$\widehat{\mathbf{R}}_{tt}{}^M = \frac{1}{L_w}\mathscr{H}_t^T\mathscr{H}_t = \frac{1}{L_w}\mathbf{Q}^{-T}\mathbf{B}^2\mathbf{Q}^{-1} \tag{18}$$

where $\mathbf{R}_{tt}^M \in \mathbb{R}^{M \times M}$.

GSVD-based optimal filtering is to find a transformation matrix $\mathbf{P} \in \mathbb{R}^{M \times M}$, which transforms the matrix $\mathscr{H}_x$ to the Hankel matrix of the estimated desired signal $\widehat{\mathscr{H}_d}$

$$\widehat{\mathscr{H}_d} = \mathscr{H}_x\mathbf{P}. \tag{19}$$

Under the assumption that the interference is additive and uncorrelated with the desired signal, the minimum variance estimation (MVE) algorithm (Van Huffel, 1993) gives $\mathbf{P}$ as

$$\mathbf{P} = \mathbf{Q}\left(\frac{\mathbf{C}^2 - \mathbf{B}^2}{\mathbf{C}^2}\right)\mathbf{Q}^{-1} \tag{20}$$

Substituting Eq. (20) into Eq. (19), we have

$$\widehat{\mathscr{H}_d} = \mathbf{U}\mathbf{C}'\mathbf{Q}^{-1} \tag{21}$$

where $\mathbf{C}' = \mathbf{C}^{-1}(\mathbf{C}^2 - \mathbf{B}^2)$.

In practice, the diagonal elements c_i' of matrix $\mathbf{C}'$ should be nonnegative, and c_i' is given by

$$c_i' = \max\{c_i^2 - b_i^2, 0\}/c_i, \; i = 1, 2, \; ..., M. \tag{22}$$

3. Optimal filtering in the LP residual domain

3.1. Constrained MMSE optimal filter in LP residual domain

The Hankel matrix of the error signal is given by

$$\begin{aligned}\mathscr{H}_e &= \mathscr{H}_d - \mathscr{H}_x\mathbf{P} \\ &= \mathscr{H}_d(\mathbf{I}_M - \mathbf{P}) - \mathscr{H}_t\mathbf{P}\end{aligned} \tag{23}$$

where $\mathscr{H}_e \in \mathbb{R}^{L_w \times M}$. The first term of the right hand side (RHS) of Eq. (23) is referred as the signal distortion, and the second term is referred as the residual interference. In fact, we want to minimize the MSE of the signal distortion while keeping the residual interference under a predefined threshold. Two linear constrained estimators have been proposed, namely time domain constrained (TDC) and spectral domain constrained (SDC) (Ephraim and Van Trees, 1995). The TDC estimator is a special case of the SDC estimator. Therefore, only the SDC estimator will be considered here. In this case, the optimization problem can be formulated as

$$\min_{\mathbf{P}} \mathrm{tr}\{\varepsilon_d \varepsilon_d^T\} \tag{24}$$

subjected to

$$E\{|\varepsilon_t^T \mathbf{q}_i|^2\} \le \alpha_i \sigma_t^2,\ i = 1, 2, \ldots, M \tag{25}$$

where $\mathrm{tr}\{\cdot\}$ is the matrix trace, $\boldsymbol{\varepsilon}_d$ and $\boldsymbol{\varepsilon}_t$ are the first column vector of $\mathscr{H}_d(\mathbf{I}_M - \mathbf{P})$ and $\mathscr{H}_t\mathbf{P}$, respectively. $\mathbf{q}_i$ is the ith column vector of $\mathbf{Q}^{-1}$, α_i is a suppression gain function, and σ_t^2 is the variance of the interference. The solution to this SDC MMSE optimal filtering problem is given by

$$\mathbf{P}_{\mathrm{opt}} = \mathbf{Q}\Delta\mathbf{Q}^{-1} \tag{26}$$

according to Ephraim and Van Trees (1995) under the assumption that the interference is additive, where $\Delta = \mathrm{diag}\{\delta_1, \delta_2, \ldots \delta_M\}$ is a diagonal matrix with elements

$$\delta_i = \sqrt{\alpha_i},\ i = 1, 2, \ldots, M \tag{27}$$

and α_i is the noise suppression gain function. In Ephraim and Van Trees (1995), the authors proposed an aggressive noise suppression gain function for α_i, which is given by

$$\alpha_i = \exp\{-\gamma\sigma_t^2/c_i^2\},\ i = 1, 2, \ldots, M \tag{28}$$

where γ is an independent value which typically ranges from 1 to 5. In Zheng et al. (2014); Doclo and Moonen (2002), the following suppression gain function was used to avoid estimating the noise variance σ_t^2,

$$\alpha_i = \exp\{-\gamma b_i^2/c_i^2\},\ i = 1, 2, \ldots, M. \tag{29}$$

The major drawback of the approach above is that the LP coefficients (LPC), i.e., the structures of the enhanced speech are changed, and this will have a negative effect on speech quality, especially in remote speech communication based on LPC coding. To cope with this problem, we proposed to use the constrained MMSE optimal filtering in the LP residual domain (Zheng et al., 2014). By using the LP model, $x(n)$ can be given by

$$x(n) = \sum_{m=1}^{L_p} a_x^m x(n-m) + r_x(n) \tag{30}$$

where L_p is the order of the LP model, and a_x^m with $m = 1, 2, \cdots, L_p$ are the LP coefficients of $x(n)$. $r_x(n)$ is the LP residual of $x(n)$. It is assumed that the LP coefficients of the microphone signal is the same as the early reverberant signal (Gaubitch et al., 2003). Applying the same LP filtering process on each side of Eq. (12), we have

$$r_x(n) = r_d(n) + r_t(n) \tag{31}$$

where $r_d(n)$ and $r_t(n)$ are the LP residuals of the desired signal and the interference, respectively. It is obvious that $r_t(n)$ is additive, and the implementation of the constrained MMSE optimal filtering in the LP residual domain is straightforward. Meanwhile, the authors in Zheng et al. (2014) used a more aggressive suppression gain function, which is given by

$$\alpha_{i,\mathrm{LP}} = \exp\left\{-\frac{\gamma \cdot b_{i,\mathrm{LP}}^2}{\max\{c_{i,\mathrm{LP}}^2 - b_{i,\mathrm{LP}}^2, \sigma_{\min}^2\}}\right\} \tag{32}$$

where $b_{i,\ \mathrm{LP}}$, $c_{i,\ \mathrm{LP}}$ are the generalized singular values of the Hankel matrices $\mathscr{H}_t^{\mathrm{LP}}$ and $\mathscr{H}_x^{\mathrm{LP}}$, which are constructed by the residual signal $r_t(n)$ and $r_x(n)$, respectively. $\sigma_{\min}^2$ is a small positive value avoiding division by zero.

The estimated Hankel matrix of the desired signal in LP residual domain is given by

$$\begin{aligned}\mathscr{H}_d^{\mathrm{LP}} &= \mathscr{H}_x^{\mathrm{LP}}\mathbf{P}_{\mathrm{opt}}^{\mathrm{LP}} \\ &= \mathbf{U}_{\mathrm{LP}}\mathbf{C}'_{\mathrm{LP}}\mathbf{Q}_{\mathrm{LP}}^{-1}\end{aligned} \tag{33}$$

where $\mathbf{P}_{\mathrm{opt}}^{\mathrm{LP}}$ is the optimal transformation matrix in the LP residual domain, $\mathbf{U}_{\mathrm{LP}}$ and $\mathbf{Q}_{\mathrm{LP}}$ are the decomposed matrix of $\mathscr{H}_x^{\mathrm{LP}}$ and $\mathscr{H}_t^{\mathrm{LP}}$. $\mathbf{C'}_{\mathrm{LP}}$ is a diagonal matrix with diagonal elements $c'_{i,\mathrm{LP}} = \alpha_{i,\mathrm{LP}} \cdot c_{i,\mathrm{LP}}$, $i = 1, 2, \ldots, M$.

The estimated Hankel matrix $\mathscr{H}_d^{\mathrm{LP}}$ may not have the Hankel-form structure, and we can simply average the anti-diagonal elements of $\mathscr{H}_d^{\mathrm{LP}}$ to recover the Hankel-form structure. Once the LP residual of the desired signal is obtained, the desired signal can be reconstructed by using the inverse process of LP.

3.2. Perceptually constrained optimal filter in the LP residual domain

Under low SNR and DRR regions, the CMMSE-GSVD-LPRE

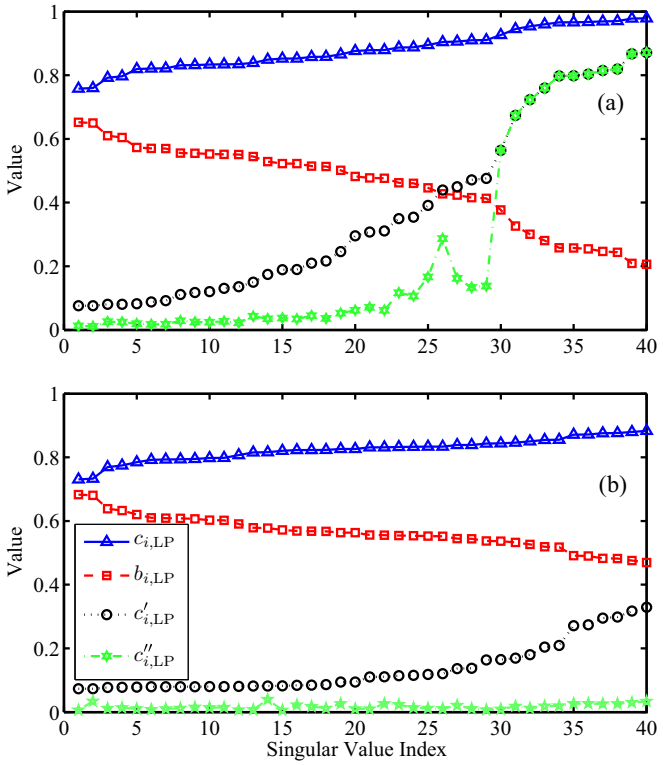

Fig. 1. Singular values of the speech signal, the interference signal, the CMMSE-GSVD-LPRE and the PCMMSE-GSVD-LPRE algorithm. (a) voice speech frame, (b) noise-only frame.

Table 1
Values of parameters used in the proposed algorithm.

$L_h = 512$	$L_s = 256$	$M = 40$	$L_w = 473$
$P = 20$	$\gamma = 2.5$	$f_s = 16$ kHz	$N = 512$

Table 2
Segmental SNR results of the six algorithms in reverberant and noisy environment.

SegSNR		White Gaussian noise					Babble noise					Factory noise				
SNR(dB)		−5	0	5	10	15	−5	0	5	10	15	−5	0	5	10	15
'anechoic room'	N	−7.25	−4.92	−2.27	0.35	2.68	−7.13	−4.78	−2.13	0.46	2.78	−6.78	−4.36	−1.76	0.82	3.09
	SS	−2.66	−0.85	0.70	1.97	2.94	−3.01	−1.42	0.07	1.45	2.60	−2.92	−1.32	0.29	1.61	2.72
	PSS	−2.38	0.05	2.03	3.70	5.00	−3.34	−1.28	0.72	2.50	4.08	−2.74	−0.75	1.17	2.84	4.27
	GSVD	−4.20	−1.67	0.80	3.00	4.78	−4.82	−2.44	−0.09	2.08	3.99	−4.18	−1.88	0.42	2.53	4.38
	PCGSVD	−4.16	−1.61	0.89	3.15	4.98	−4.63	−2.10	0.40	**2.61**	**4.43**	−4.11	−1.71	0.72	2.92	**4.75**
	CMMSE	−1.43	0.55	2.39	4.06	5.39	−2.82	−1.04	0.79	2.54	4.17	−2.46	−0.71	1.15	2.84	4.43
	PCMMSE	**−0.59**	**1.20**	**2.89**	**4.50**	**5.84**	**−2.77**	**−1.00**	**0.82**	2.55	4.19	**−2.19**	**−0.54**	**1.29**	**2.92**	4.50
	ΔPCMMSE	6.67	6.12	5.17	4.15	3.16	4.36	3.77	2.95	2.09	1.41	4.59	3.82	3.05	2.10	1.41
lecture room	R + N	−8.91	−7.98	−7.20	−6.66	−6.38	−8.83	−7.91	−7.18	−6.65	−6.34	−8.70	−7.79	−7.10	−6.59	−6.34
	SS	−5.05	−4.45	−4.18	−4.01	−3.97	−4.90	−4.36	−4.17	−4.03	−3.93	−4.74	−4.29	−4.12	−4.00	−3.91
	PSS	−5.30	−4.54	−4.20	−4.02	−3.99	−5.52	−4.72	−4.36	−4.15	−4.01	−5.13	−4.56	−4.24	−4.06	−4.01
	GSVD	−6.50	−5.62	−5.11	−4.79	−4.60	−6.69	−5.77	−5.24	−4.83	−4.51	−6.27	−5.49	−5.02	−4.67	−4.49
	PCGSVD	−6.50	−5.63	−5.11	−4.74	−4.43	−6.60	−5.59	−4.97	−4.41	−3.99	−6.26	−5.43	−4.86	−4.38	−4.00
	CMMSE	−4.22	−3.73	−3.49	−3.40	−3.40	−4.69	−4.04	−3.77	−3.59	−3.48	−4.29	−3.84	−3.59	−3.44	−3.43
	PCMMSE	**−3.15**	**−2.92**	**−2.88**	**−2.88**	**−2.93**	**−4.62**	**−3.98**	**−3.72**	**−3.54**	**−3.42**	**−3.87**	**−3.58**	**−3.38**	**−3.28**	**−3.29**
	ΔPCMMSE	5.76	5.06	4.31	3.78	3.46	4.21	3.93	3.47	3.11	2.91	4.83	4.21	3.71	3.31	3.05
meeting room	R + N	−9.08	−8.24	−7.42	−6.76	−6.30	−8.91	−8.15	−7.38	−6.72	−6.29	−8.93	−8.06	−7.28	−6.65	−6.23
	SS	−5.18	−4.61	−4.24	−4.01	−3.85	−5.09	−4.41	−4.15	−3.97	−3.81	−4.90	−4.42	−4.14	−3.96	−3.81
	PSS	−5.53	−4.84	−4.36	−4.12	−3.94	−5.76	−4.89	−4.46	−4.20	−4.02	−5.31	−4.74	−4.34	−4.14	−3.95
	GSVD	−6.78	−5.97	−5.38	−4.97	−4.66	−6.95	−6.03	−5.44	−4.99	−4.61	−6.54	−5.79	−5.23	−4.84	−4.52
	PCGSVD	−6.79	−5.98	−5.37	−4.93	−4.49	−6.84	−5.87	−5.12	−4.51	−4.00	−6.54	−5.74	−5.05	−4.50	−3.99
	CMMSE	−4.34	−3.91	−3.65	−3.50	−3.41	−4.91	−4.19	−3.83	−3.65	−3.53	−4.41	−3.92	−3.64	−3.50	−3.41
	PCMMSE	**−3.22**	**−3.06**	**−3.01**	**−2.97**	**−2.91**	**−4.84**	**−4.13**	**−3.79**	**−3.62**	**−3.50**	**−3.95**	**−3.57**	**−3.42**	**−3.36**	**−3.29**
	ΔPCMMSE	5.86	5.18	4.41	3.79	3.38	4.07	4.02	3.59	3.10	2.78	4.97	4.49	3.86	3.29	2.93
office room	R + N	−9.15	−8.33	−7.53	−6.94	−6.58	−9.09	−8.25	−7.51	−6.93	−6.63	−9.00	−8.16	−7.40	−6.86	−6.53
	SS	−5.24	−4.69	−4.40	−4.21	−4.06	−5.15	−4.56	−4.24	−4.15	−4.14	−5.01	−4.47	−4.28	−4.13	−4.01
	PSS	−5.57	−4.87	−4.48	−4.28	−4.14	−5.87	−5.03	−4.53	−4.36	−4.30	−5.39	−4.79	−4.48	−4.27	−4.15
	GSVD	−6.82	−6.02	−5.45	−5.11	−4.85	−7.05	−6.12	−5.51	−5.09	−4.89	−6.56	−5.82	−5.33	−4.97	−4.71
	PCGSVD	−6.83	−6.02	−5.45	−5.08	−4.72	−6.98	−5.97	−5.22	−4.70	−4.37	−6.56	−5.76	−5.17	−4.68	−4.27
	CMMSE	−4.39	−4.00	−3.76	−3.63	−3.54	−5.01	−4.30	−3.88	−3.76	−3.77	−4.52	−4.00	−3.74	−3.64	−3.54
	PCMMSE	**−3.26**	**−3.17**	**−3.13**	**−3.12**	**−3.07**	**−4.94**	**−4.23**	**−3.83**	**−3.72**	**−3.73**	**−4.08**	**−3.70**	**−3.52**	**−3.50**	**−3.43**
	ΔPCMMSE	5.89	5.16	4.40	3.83	3.51	4.15	4.02	3.68	3.21	2.90	4.92	4.45	3.88	3.36	3.10

algorithm has no guidelines to control the amount of interference reduction and speech distortion, which may somewhat degrade the performance of the algorithm. Auditory masking is a well-known psychoacoustic property of the human auditory system that has been widely used in speech enhancement (Virag, 1999; Gustafsson et al., 1998), speech coding (Johnston, 1988; Painter and Spanias, 2000), etc. In this paper, we propose to use the auditory masking property to control the level of the residual interference. To make the residual interference imperceivable, the auditory masking threshold (AMT) curve is calculated first, and the perceptually based optimal filter is derived.

The power spectrum of the desired signal is required for evaluating the AMTs in the frequency domain, and this power spectrum was estimated by the Blackman–Tukey frequency estimation technique in Ju and Lee (2007). In this paper, we propose to estimate the power spectrum of the desired signal directly from the output of the CMMSE-GSVD-LPRE.

Suppose the output of the CMMSE-GSVD-LPRE is $\hat{d}(n)$, and the power spectrum $\hat{\Gamma}_d(w)$ is given by:

$$\hat{\Gamma}_d(w) = |\hat{D}(w)|^2 \tag{34}$$

where $\hat{D}(w)$ is the Fourier transform of $\hat{d}(n)$.

There are several steps involved in calculating the AMT curve using $\hat{\Gamma}_d(w)$ and we give a brief introduction of different calculation steps as follows (the detailed expression of these steps can be found in Johnston (1988) and Painter and Spanias (2000)):

- The power spectrum of the desired signal is partitioned into critical bands, and the energy in each critical band is summed up.
- The effects of masking across critical bands are calculated using the spreading function, which is taken from Schroeder et al. (1979)
- Subtract a relative threshold offset depending on the noise-like or tone-like nature of the spectrum structure. A relative threshold offset proposed in Sinha and Tewfik (1993) is used to simplify the calculation.
- Normalize the energy in each critical band, and include the absolute threshold information of hearing

The obtained AMT is denoted as $T(w)$ and is used to calculate the perceptually based singular value. The un-normalized auto-correlation matrix of the desired signal $\widehat{\mathbf{R}}_{dd}^{M} \in \mathbb{R}^{M\times M}$ is a Toeplitz matrix, which is given by

$$\widehat{\mathbf{R}}_{dd}^{M} = \begin{bmatrix} \hat{r}_{dd}(0) & \hat{r}_{dd}(1) & \cdots & \hat{r}_{dd}(M-1) \\ \hat{r}_{dd}(1) & \hat{r}_{dd}(0) & \cdots & \hat{r}_{dd}(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{dd}(M-1) & \hat{r}_{dd}(M-2) & \cdots & \hat{r}_{dd}(0) \end{bmatrix} \tag{35}$$

where $\hat{r}_{dd}(m)$, $m = 0, 1, ...,M-1$, is the un-normalized correlation function of $\hat{d}(n)$.

Suppose η_i and $\mathbf{g}_i \in \mathbb{R}^{M\times 1}$ are the ith eigenvalue and unit norm eigenvector of $\widehat{\mathbf{R}}_{dd}^{M}$, respectively. A well-known relationship between the eigenvalue and the power spectrum is given by Jabloun and Champagne (2002)

$$\eta_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{\Gamma}_d(w)\,|G_i(w)|^2 dw,\ i = 1, 2, ...,M \tag{36}$$

where $G_i(w)$ is the Fourier transform of $\mathbf{g}_i$. In the discrete domain, Eq. (36) can be computed by

$$\eta_i = \frac{1}{N}\sum_{k=0}^{N-1} \hat{\Gamma}_d(w_k)\,|G_i(w_k)|^2,\ i = 1, 2, ...,M \tag{37}$$

where $\hat{\Gamma}_d(w_k)$ and $G_i(w_k)$ are the Discrete Fourier transforms (DFT) of $\hat{r}_{dd}(m)$ and $g_i(m)$, respectively. $g_i(m)$ is the mth element in $\mathbf{g}_i$. $w_k = 2\pi k/N$, where N is the length of DFT. In practice, N should be

Table 3
PESQ results of the six algorithms in reverberant and noisy environment.

PESQ		White Gaussian noise					Babble noise					Factory noise				
SNR(dB)		−5	0	5	10	15	−5	0	5	10	15	−5	0	5	10	15
'anechoic room'	N	1.16	1.48	1.85	2.23	2.60	1.29	1.65	2.02	2.38	2.70	1.20	1.54	1.93	2.30	2.66
	SS	1.45	1.85	2.22	2.52	2.73	1.50	1.90	2.18	2.53	2.76	1.49	1.88	2.24	2.53	2.76
	PSS	1.61	2.08	2.45	2.75	3.01	1.41	1.85	2.24	2.63	2.96	1.48	1.89	2.30	2.65	2.97
	GSVD	1.40	1.81	2.21	2.60	2.96	1.39	1.79	2.16	2.52	2.85	1.37	1.76	2.18	2.56	2.90
	PCGSVD	1.46	1.87	2.28	2.67	2.98	1.43	1.84	2.22	2.62	2.99	1.43	1.83	2.26	2.68	3.05
	CMMSE	1.78	2.20	2.59	2.93	3.19	1.52	1.94	2.33	2.69	3.02	1.50	1.94	2.35	2.72	3.04
	PCMMSE	**1.78**	**2.23**	**2.60**	**2.94**	**3.23**	**1.70**	**2.09**	**2.45**	**2.79**	**3.10**	**1.68**	**2.08**	**2.46**	**2.83**	**3.14**
	ΔPCMMSE	0.62	0.75	0.75	0.71	0.63	0.41	0.44	0.43	0.41	0.39	0.48	0.54	0.53	0.53	0.48
lecture room	R + N	1.03	1.31	1.59	1.84	1.95	1.13	1.44	1.69	1.88	2.00	1.04	1.39	1.64	1.86	1.97
	SS	1.27	1.54	1.72	1.85	1.85	1.17	1.53	1.69	1.84	1.89	1.21	1.56	1.72	1.82	1.84
	PSS	1.34	1.67	1.86	1.98	1.99	1.17	1.55	1.79	1.93	2.01	1.27	1.57	1.81	1.95	1.97
	GSVD	1.28	1.61	1.88	2.04	2.08	1.19	1.54	1.81	1.96	2.06	1.25	1.57	1.84	1.98	2.03
	PCGSVD	1.31	1.65	1.90	2.06	2.06	1.20	1.57	1.82	1.95	2.06	1.27	1.62	1.86	1.99	2.03
	CMMSE	1.66	1.92	2.03	2.09	2.06	1.29	1.62	1.86	1.97	**2.06**	1.25	1.66	1.84	1.99	2.01
	PCMMSE	**1.73**	**1.97**	**2.06**	**2.11**	**2.10**	**1.40**	**1.70**	**1.89**	**2.01**	2.05	**1.35**	**1.72**	**1.93**	**2.03**	**2.04**
	ΔPCMMSE	0.69	0.66	0.47	0.26	0.14	0.26	0.26	0.20	0.13	0.05	0.31	0.33	0.29	0.18	0.08
meeting room	R + N	1.07	1.33	1.66	1.98	2.24	1.21	1.54	1.84	2.12	2.33	1.09	1.40	1.75	2.07	2.29
	SS	1.29	1.60	1.87	2.07	2.18	1.27	1.63	1.89	2.09	2.21	1.28	1.58	1.88	2.10	2.22
	PSS	1.38	1.73	2.03	2.21	2.32	1.26	1.65	1.93	2.18	2.32	1.24	1.63	1.95	2.19	2.32
	GSVD	1.31	1.65	1.98	2.24	2.40	1.24	1.66	1.91	2.19	2.36	1.23	1.60	1.96	2.22	2.38
	PCGSVD	1.36	1.68	2.02	2.26	2.42	1.29	1.69	1.93	2.20	**2.36**	1.27	1.64	2.00	2.25	**2.39**
	CMMSE	1.77	2.05	2.27	2.39	**2.45**	1.36	1.75	1.99	**2.22**	2.35	1.27	1.73	2.02	2.25	2.36
	PCMMSE	**1.83**	**2.09**	**2.28**	**2.40**	2.44	**1.44**	**1.81**	**2.06**	2.22	2.32	**1.41**	**1.82**	**2.09**	**2.27**	2.34
	ΔPCMMSE	0.76	0.75	0.62	0.43	0.20	0.23	0.27	0.22	0.10	−0.01	0.32	0.42	0.34	0.20	0.05
office room	R + N	1.05	1.32	1.62	1.91	2.13	1.20	1.50	1.80	2.04	2.15	1.12	1.39	1.71	1.98	2.16
	SS	1.23	1.54	1.74	1.96	2.04	1.31	1.64	1.85	1.97	2.00	1.27	1.56	1.81	1.96	2.05
	PSS	1.35	1.72	1.95	2.11	2.20	1.26	1.62	1.90	2.09	2.13	1.24	1.64	1.88	2.09	2.18
	GSVD	1.27	1.62	1.92	2.13	2.24	1.26	1.61	1.90	2.09	2.17	1.20	1.62	1.90	2.11	2.22
	PCGSVD	1.32	1.65	1.94	2.14	2.24	1.29	1.64	1.91	2.08	2.15	1.23	1.65	1.91	2.11	2.20
	CMMSE	1.71	1.97	2.13	2.21	2.25	1.34	1.71	1.98	2.12	**2.18**	1.28	1.72	1.94	2.14	2.20
	PCMMSE	**1.79**	**2.00**	**2.16**	**2.24**	**2.26**	**1.44**	**1.78**	**2.03**	**2.14**	2.18	**1.38**	**1.80**	**2.02**	**2.17**	**2.22**
	ΔPCMMSE	0.73	0.68	0.54	0.33	0.13	0.24	0.28	0.23	0.10	0.02	0.25	0.41	0.31	0.18	0.06

larger than M.

While in the LP residual domain, the perceptually based eigenvalue is calculated by replacing $\hat{\Gamma}_d(w_k)$ with the AMT, i.e. $T(w_k)$ and replacing $G_i(w_k)$ with $Q_i^{\mathrm{LP}}(w_k)$:

$$\eta_i^{\mathrm{AMT}} = \frac{1}{N}\sum_{k=0}^{N-1} T(w_k)\,|Q_i^{\mathrm{LP}}(w_k),\,|^2,\ i = 1,\ 2,\ \ldots,M \tag{38}$$

where $Q_i^{\mathrm{LP}}(w_k)$ is the DFT of $\mathbf{q}_i^{\mathrm{LP}}$, which is the ith column vector of $\mathbf{Q}_{\mathrm{LP}}^{-1}$ and $T(w_k) = T(w)|_{w=w_k}$

η_i^{AMT} are perceptually based eigenvalues and can not be implemented in GSVD-based approach directly, and should be transformed into the generalized singular value domain. The relationship between the eigenvalue and the generalized singular value is given by:

$$\Pi^2 = \Theta^T \Sigma \Theta \tag{39}$$

where $\Pi = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ is the transformed generalized singular value. $\Theta = \mathbf{G}^T\mathbf{Q}_{\mathrm{LP}}$ is the transformation matrix, where $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_M]$ and $\Sigma = \mathrm{diag}\{\eta_1^{\mathrm{AMT}}, \eta_2^{\mathrm{AMT}}, \ldots, \eta_M^{\mathrm{AMT}},\}$

So far, we have calculated the perceptually based singular value λ_i. Then, the estimated Hankel matrix of the desired signal in LP residual domain with perceptual constrain is given by

$$\mathcal{H}_d^{\mathrm{LP}} = \mathbf{U}_{\mathrm{LP}}\mathbf{C}_{\mathrm{LP}}^{'}\mathbf{Q}_{\mathrm{LP}}^{-1} \tag{40}$$

where $\mathbf{C}_{\mathrm{LP}}^{'} = \mathrm{diag}\{c_{1,\mathrm{LP}}^{'}, c_{2,\mathrm{LP}}^{'}, \ldots, c_{M,\mathrm{LP}}^{'}\}$ and the diagonal elements of $\mathbf{C''}_{\mathrm{LP}}$ are given by:

$$c_{i,\mathrm{LP}}^{''} = c_{i,\mathrm{LP}}^{'}\cdot\min\left\{1,\ \frac{\lambda_i}{b_{i,\mathrm{LP}}}\right\},\ i = 1,\ 2,\ \ldots,M. \tag{41}$$

The physical meaning of Eq. (41) is straightforward. When $b_{i,\ \mathrm{LP}}$ is larger than λ_i, which means that the interference components are perceptible, and the singular value obtained by CMMSE-GSVD-LPRE will be attenuated. When $b_{i,\ \mathrm{LP}}$ is smaller than λ_i, which means the interference components are imperceptible, and the singular value obtained by CMMSE-GSVD-LPRE will not be changed. Fig. 1 plots the singular values of the speech signal, the interference signal, the CMMSE-GSVD-LPRE output signal and the PCMMSE-GSVD-LPRE output signal for voiced and noise-only speech frames corrupted by white Gaussian noise at 10 dB SNR, with $M = 40$. One can get that some larger singular values of CMMSE-GSVD-LPRE algorithm are preserved unchanged in Fig. 1(a), while the smaller singular values, which are mainly contributed by perceptual residual interference, are further suppressed using the proposed PCMMSE-GSVD-LPRE algorithm. In noise-only frame, Fig. 1(b) reveals that all the singular values can be further suppressed using the proposed PCMMSE-GSVD-LPRE algorithm.

3.3. Real time implementation of the proposed algorithm

In this part, the detailed steps of the proposed algorithm are presented as follows:

1. *Framing and fast Fourier transform (FFT)*: The microphone signal $x(n)$ is segmented into frames with frame length L_h and frame shift L_s. Calculating the FFT of the lth frame to obtain $X(w_k, l)$, where $w_k = 2\pi k/L_h$, $k = 0, 1, \cdots, L_h - 1$.
2. *Estimating the NPSD and LRSV*: The unbiased MMSE NPSD estimator proposed in Gerkmann and Hendriks (2012) is used to estimate the NPSD, $\hat{\sigma}_n^2(w_k, l)$, and the simple and efficient method proposed in Wu and Wang (2006) is used to estimate the LRSV, $\hat{\sigma}_{s_{\mathrm{lr}}}^2(w_k, l)$ in this paper.
3. *Estimating the interference signal in time domain*: The power spectral of the interference $\hat{\sigma}_{\mathrm{int}}^2(w_k, l)$ is obtained by using Zheng et al. (2014, (15)), which is a weighted sum of NPSD and LRSV, i.e. $\hat{\sigma}_{\mathrm{int}}^2(w_k, l) = \xi\cdot\hat{\sigma}_n^2(w_k, l) + \beta\cdot\hat{\sigma}_{s_{\mathrm{lr}}}^2(w_k, l)$, where ξ, β are the weighting

Table 4
SRMRnorm results of the six algorithms in reverberant and noisy environment.

SRMRnorm		White Gaussian noise					Babble noise					Factory noise				
SNR(dB)		−5	0	5	10	15	−5	0	5	10	15	−5	0	5	10	15
'anechoic room'	N	1.01	1.80	2.83	3.54	3.86	1.54	2.34	3.23	3.71	3.93	1.41	2.30	3.16	3.69	3.92
	SS	2.23	3.15	3.77	4.11	4.18	2.84	3.48	3.80	4.07	4.09	2.47	3.36	3.82	4.07	4.19
	PSS	1.74	2.85	3.58	3.96	4.08	2.38	3.15	3.76	4.01	4.11	2.30	3.20	3.75	4.03	4.12
	GSVD	2.08	3.14	3.84	4.11	4.17	2.50	3.23	3.78	4.01	4.10	2.50	3.32	3.84	4.07	4.13
	PCGSVD	2.05	3.11	3.84	4.13	4.21	2.52	3.30	3.86	4.08	4.14	2.48	3.33	3.88	4.12	4.18
	CMMSE	2.54	3.58	4.06	4.21	4.24	3.02	3.60	3.98	4.11	**4.15**	2.82	3.59	3.98	4.16	4.19
	PCMMSE	**3.26**	**3.90**	**4.14**	**4.23**	**4.24**	**3.10**	**3.65**	**4.00**	**4.11**	4.15	**3.02**	**3.69**	**4.02**	**4.17**	**4.20**
	ΔPCMMSE	2.26	2.10	1.31	0.69	0.38	1.56	1.31	0.77	0.40	0.22	1.61	1.39	0.86	0.48	0.27
lecture room	R + N	0.81	1.15	1.62	1.95	2.09	1.14	1.51	1.85	2.07	2.17	1.11	1.47	1.82	2.06	2.16
	SS	1.72	2.04	2.24	2.39	2.43	2.04	2.30	2.43	2.49	2.51	2.05	2.30	2.39	2.46	2.50
	PSS	1.39	1.90	2.27	2.46	2.51	1.80	2.21	2.44	2.54	2.56	1.84	2.21	2.45	2.54	2.56
	GSVD	1.72	2.16	2.44	2.57	2.59	2.04	2.37	2.52	2.58	2.61	2.12	2.41	2.54	2.61	2.61
	PCGSVD	1.69	2.12	2.40	2.55	2.58	2.04	2.38	2.54	2.61	2.62	2.09	2.39	2.54	2.62	2.62
	CMMSE	1.82	2.24	2.53	2.67	2.70	2.29	2.63	2.73	2.75	2.75	2.32	2.60	2.73	2.77	2.75
	PCMMSE	**2.21**	**2.47**	**2.62**	**2.71**	**2.73**	**2.39**	**2.69**	**2.75**	**2.77**	**2.76**	**2.49**	**2.68**	**2.77**	**2.79**	**2.76**
	ΔPCMMSE	1.40	1.32	1.00	0.76	0.63	1.25	1.18	0.90	0.69	0.59	1.38	1.21	0.94	0.73	0.60
meeting room	R + N	0.91	1.39	2.05	2.51	2.74	1.27	1.80	2.34	2.69	2.83	1.22	1.73	2.31	2.66	2.81
	SS	1.97	2.40	2.67	2.84	2.92	2.30	2.66	2.91	3.01	3.06	2.18	2.60	2.84	2.94	3.01
	PSS	1.52	2.15	2.60	2.84	2.94	1.97	2.45	2.78	2.96	3.02	1.97	2.44	2.79	2.95	3.01
	GSVD	1.94	2.50	2.85	3.01	3.07	2.21	2.66	2.94	3.07	3.10	2.28	2.69	2.98	3.07	3.10
	PCGSVD	1.91	2.47	2.82	3.00	3.07	2.21	2.67	2.95	3.09	3.13	2.25	2.67	2.98	3.08	3.13
	CMMSE	2.03	2.58	2.89	3.06	3.13	2.52	2.91	3.11	3.19	3.20	2.40	2.80	3.09	3.18	3.20
	PCMMSE	**2.51**	**2.83**	**2.97**	**3.09**	**3.16**	**2.62**	**2.96**	**3.13**	**3.20**	**3.21**	**2.54**	**2.88**	**3.13**	**3.19**	**3.22**
	ΔPCMMSE	1.61	1.44	0.92	0.58	0.42	1.35	1.16	0.79	0.52	0.38	1.32	1.15	0.82	0.53	0.41
office room	R + N	0.85	1.32	1.91	2.33	2.56	1.25	1.75	2.24	2.54	2.63	1.19	1.70	2.22	2.51	2.64
	SS	1.90	2.33	2.54	2.74	2.86	2.31	2.69	2.88	2.99	2.94	2.22	2.60	2.79	2.93	2.94
	PSS	1.50	2.12	2.59	2.79	2.90	1.97	2.50	2.82	2.94	2.95	2.00	2.50	2.79	2.93	2.97
	GSVD	1.91	2.45	2.79	2.94	3.01	2.27	2.70	2.94	3.03	3.01	2.28	2.74	2.96	3.03	3.03
	PCGSVD	1.87	2.41	2.76	2.93	3.01	2.25	2.70	2.95	3.04	3.02	2.24	2.72	2.95	3.04	3.05
	CMMSE	2.00	2.56	2.87	3.02	3.11	2.58	3.00	3.16	3.19	3.16	2.44	2.89	3.09	3.18	3.18
	PCMMSE	**2.55**	**2.83**	**2.97**	**3.07**	**3.13**	**2.69**	**3.05**	**3.18**	**3.20**	**3.17**	**2.62**	**2.98**	**3.13**	**3.20**	**3.19**
	ΔPCMMSE	1.70	1.51	1.06	0.73	0.58	1.44	1.30	0.94	0.67	0.54	1.42	1.28	0.92	0.68	0.56

coefficients, and the time domain signal of the interference $\hat{t}(n)$ is calculated by using Zheng et al. (2014, (16)) with the overlap-add method frame by frame.

4. *Linear prediction*: The LP coefficients a_x^m, $m = 0, 1, ..., P$, and the residual signal $r_x(n)$ of the microphone signal are calculated by using the Levinson–Durbin algorithm Cybenko (1980). The residual signal $\hat{r}_t(n)$ of the interference is obtained by using the same LP process.
5. *Constructing the Hankel matrix in LP residual domain*: The Hankel matrices $\mathscr{H}_x^{\mathrm{LP}}$, $\mathscr{H}_t^{\mathrm{LP}}$ with dimension $L_w \times M$ of the lth frame signal $r_x(n)$, $\hat{r}_t(n)$ are constructed, where L_w and M satisfy: $L_w + M = L_h + 1$.
6. *Generalized singular value decomposition*: Applying the generalized singular value decomposition to obtain the decomposed matrix $\mathbf{U}_{\mathrm{LP}}$, $\mathbf{V}_{\mathrm{LP}}$, $\mathbf{Q}_{\mathrm{LP}}$, $\mathbf{C}_{\mathrm{LP}}$ and $\mathbf{B}_{\mathrm{LP}}$.
7. *Estimating the LP residual of the desired signal*: The constrained MMSE GSVD-based optimal filter is obtained, and the estimated LP residual of the desired signal is calculated by using Eq. (33).
8. *Calculating the power spectral of the desired signal*: Using the first L_s elements in the first column vector of $\mathscr{H}_d^{\mathrm{LP}}$, and the LP coefficients of the $(l-1)$th frame to synthesize the estimated desired signal. Then the power spectral of the desired signal $\hat{\Gamma}_d(w)$ can be obtained.
9. *Calculating the AMT curve*: the AMT curve $T(w)$ is obtained by using the calculation steps described above.
10. *Projecting the AMT into the generalized singular value in LP residual domain*: By using Eqs. (38) and (39), the perceptually based singular value λ_i, $i = 0, 1, ..., M-1$ are obtained.
11. *Desired signal in LP residual domain*: The estimated Hankel matrix of the desired signal in LP residual domain with MMSE and percpetual constrains is obtained by using Eq. (40).
12. *Synthesizing the enhanced signal*: Using the first L_s elements in the first column vector of $\mathscr{H}_d^{\mathrm{LP}}$ in Eq. (40), and the LP coefficients of the $(l-1)$th frame to synthesize the enhanced desired signal.

4. Simulation experiments

In this section, the proposed PCMMSE-GSVD-LPRE is compared with the SS algorithm proposed in Cohen (2003). The proposed algorithm is also compared with the single channel speech enhancement algorithm based on masking properties of human auditory system proposed in Virag (1999), which is referred as PSS. The GSVD-based approach and the perceptually constrained GSVD-based approach, which were proposed in Jensen et al. (1995) and Ju and Lee (2007), respectively, are selected as two subspace algorithms for comparation. These two subspace algorithms are referred as GSVD and PCGSVD, respectively. Our previous work (Zheng et al., 2014), the CMMSE-GSVD-LPRE algorithm, is also compared with the proposed algorithm in this paper. In order to make a complete and fair comparison, we use the same NPSD and the same LPSV estimators for all of these algorithms, and extend these algorithms to suppress the interference signals estimated in this paper.

4.1. Experimental setup

The source signals consist of 100 male speech sentences and 100 female speech sentences taken from the TIMIT speech corpus (Garofolo et al., 1988). These source signals are convoluted with the recorded RIRs, which are taken from the Aachen Impulse Response Database (Jueb et al., 2009). Three types of room are selected, including the meeting room, the office room and the lecture room. For the meeting room, the office room and the lecture room, the distances are 145 cm, 100 cm and 225 cm, respectively. Meanwhile, the reverberation times (T_{60}) of these three types of room are about 340, 656 and 878 ms, respectively. Different kinds of noise, including babble noise, factory noise and white Gaussian noise, are added to the reverberant

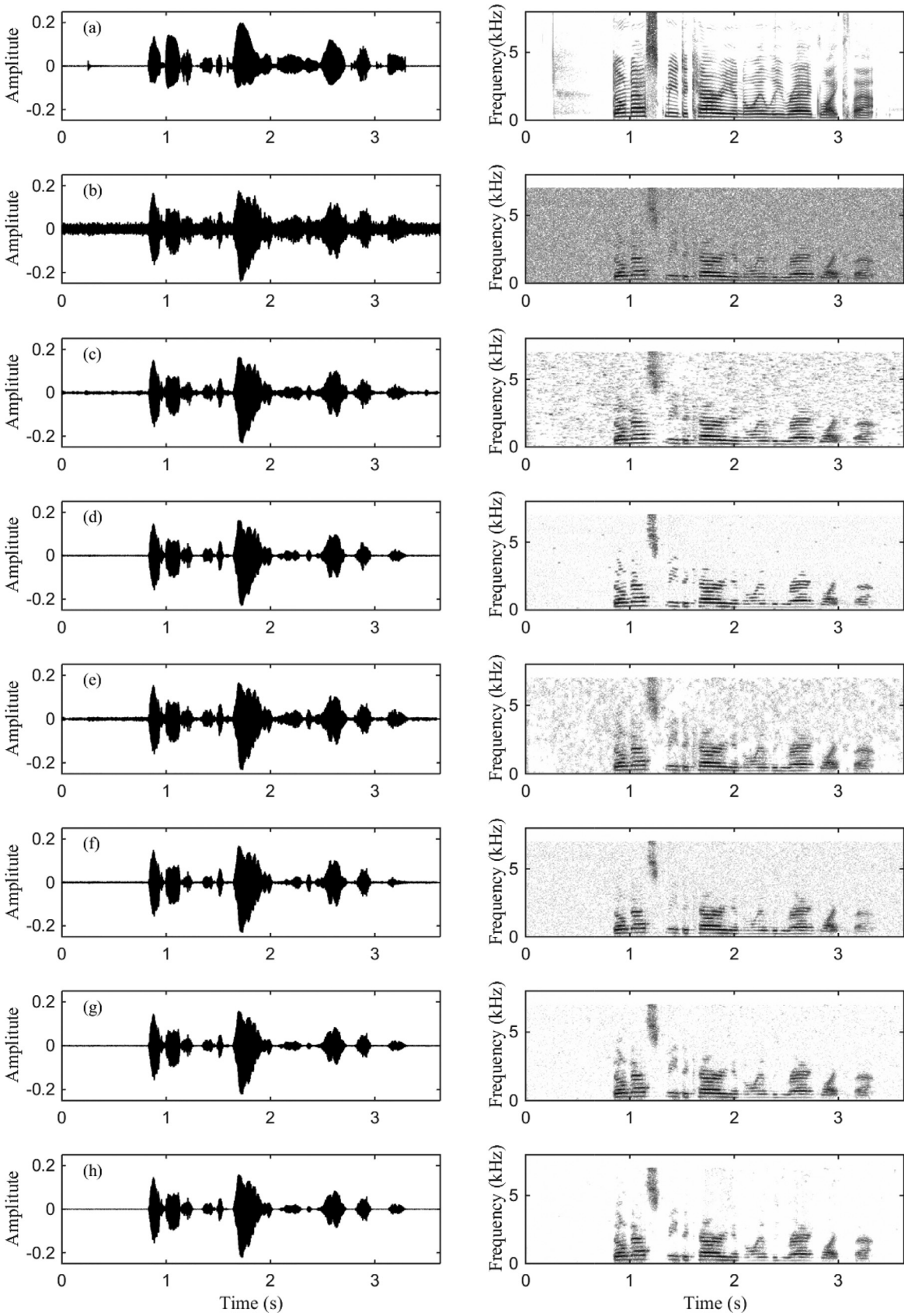

Fig. 2. Waveforms and spectrograms of: (a) clean speech, (b) the reverberant and noisy speech, (c) speech enhanced by the SS algorithm, (d) speech enhanced by the PSS algorithm, (e) speech enhanced by the GSVD algorithm, (f) speech enhanced by the PCGSVD algorithm, (g) speech enhanced by the CMMSE-GSVD-LPRE algorithm, (h) speech enhanced by the proposed PCMMSE-GSVD-LPRE algorithm.

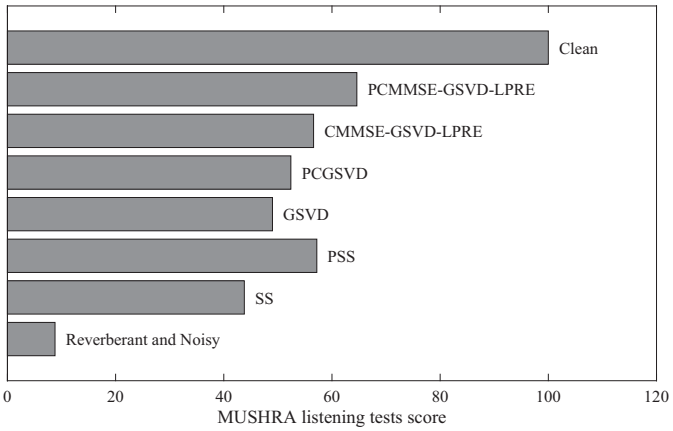

Fig. 3. Averaged MUSHRA Listening tests scores for simulation waveforms.

Table 5
Volume and reverberation time of the testing rooms.

	Room type	Volume (m^3)	T60 (s)
1	meeting room	103	0.49
2	lecture room	809	0.79
3	small reverberation room	135	1.98

signals using the filtering and noise adding tool (FaNT) (Hirsch, 2005). The input SNR ranges from − 5 dB to 15 dB with a 5 dB step size, where the noise signals are taken from the NOISEX-92 database (Varga and Steeneken, 1993). To evaluate the performance of these algorithms in noise-only environments without reverberation, these three kinds of noise are added to the clean source signals directly, where this special scenario is referred as anechoic room here.

The frame shift L_s and the frame length L_h are set to 256 and 512, respectively, which correspond to the quasi-stationary period of speech under sample frequency $f_s = 16$ kHz. M is set to 40, which can achieve a good balance between computational load and algorithm performance. P and γ are empirically chosen as 20 and 2.5, respectively. In reverberation rooms, both ξ and β are set to 1, while in anechoic room, ξ and β are set to 1 and 0, respectively. Typical values of the respective parameters for the proposed algorithm are summarized in Table 1.

4.2. Segmental SNR results

Segmental SNR is considered as a reasonable objective measure for speech enhancement, which is given by

$$\text{SegSNR}=\frac{10}{N_l}\sum_{l=0}^{N_l-1}\log_{10}\left(\frac{\sum_{n=lL_s}^{n=lL_s+L_h-1}s^2(n)}{\sum_{n=lL_s}^{n=lL_s+L_h-1}(\hat{s}(n)-s(n))^2}\right) \quad (42)$$

where N_l is the total number of the frames, and $\hat{s}(n)$ is the enhanced signal. Frames with SNRs above 35 dB do not reflect large perceptual differences and generally can be replaced with 35 dB in above equation. Likewise, during periods of silence, SNR values can become very negative since signal energies are small. These frames are set to a lower threshold, i.e. − 10 dB, instead (Quackenbush et al., 1988).

Segmental SNR results of the reverberant and noisy signals and that of the six algorithms are presented in Table 2, where the reverberant and noisy signals are denoted as 'R + N' and the noisy signals without reverberation are denoted as 'N' for abbreviation, respectively. The CMMSE-GSVD-LPRE algorithm and the PCMMSE-GSVD-LPRE algorithm proposed in this paper are denoted as 'CMMSE' and 'PCMMSE' for compact when no confusion arises, respectively. The segmental SNR improvement of the proposed algorithm is referred as ΔPCMMSE.

The segmental SNR results in Table 2 are the averaged values of all the 200 speech sentences. The proposed PCMMSE-GSVD-LPRE algorithm achieves the highest values of the segmental SNR over all competitive algorithms for white Gaussian noise in anechoic room. While for the babble noise and the factory noise, it has the best performance under low SNR conditions, i.e. SNR < 5 dB. For the reverberant and noisy signals, the proposed PCMMSE-GSVD-LPRE algorithm achieves the highest output segmental SNR values over all kinds of noise types and input SNR values evaluated in this paper. One can also get that the segmental SNR improvement of the proposed algorithm varies from both noise types and SNR values, where the largest segmental SNR improvement can be achieved for white Gaussian noise and the smallest segmental SNR improvement for babble.

4.3. Perceptual evaluation of speech quality (PESQ)

In this section, we use the PESQ recommended by ITU-T for speech quality assessment (ITU-2000, 2000) to compare the proposed algorithm with the five competing algorithms.

Table 3 gives the averaged PESQ scores over all the 200 speech sentences. It can be seen that the proposed PCMMSE-GSVD-LPRE algorithm achieves the best PESQ scores in anechoic room in all cases. While for the reverberant and noisy signals, the proposed algorithm still has the best PESQ scores in low input SNR conditions. In high input SNR conditions, the PCGSVD algorithm and the CMMSE-GSVD-LPRE algorithm have slightly higher PESQ scores, while the PESQ scores of the proposed algorithm are comparable with the best PESQ scores. Meanwhile, the PESQ scores of the proposed algorithm are higher than that of the CMMSE-GSVD-LPRE algorithm in most cases except in the meeting room with high input SNR, i.e. SNR ≥ 15 dB. The same as the segmental SNR improvement, the PESQ score improvement of the proposed algorithm are also highly correlated with noise types and input SNR values. For example, the PESQ scores of the proposed algorithm have the highest values for white Gaussian noise and the smallest values for babble.

4.4. SRMR results

In this section, we use the speech-to-reverberation modulation energy ratio (SRMR) (Falk et al., 2010) for speech quality and intelligibility assessment as a non-intrusive metric. This metric was used as one of the objective metrics in the REVERB Challenge (Kinoshita et al.,

Table 6
SRMRnorm results of the six algorithms in realistic environment.

SRMRnorm	Meeting room				Lecture room				Small reverberation room			
Distance(m)	0.5	1.0	2.0	4.0	0.5	1.0	2.0	4.0	0.5	1.0	2.0	4.0
R + N	3.02	2.74	2.30	2.11	3.97	3.47	3.18	2.83	1.94	1.15	0.85	0.84
SS	3.31	3.01	2.61	2.41	3.98	3.81	3.35	3.07	2.95	2.14	1.60	1.61
PSS	3.48	3.12	2.67	2.59	4.24	3.89	3.46	3.18	3.12	2.07	1.54	1.45
GSVD	3.44	3.10	2.67	2.61	4.29	4.04	3.51	3.15	3.05	2.04	1.44	1.45
PCGSVD	3.42	3.09	2.71	2.57	4.30	4.03	3.49	3.16	3.01	2.00	1.43	1.44
CMMSE	3.57	3.23	2.79	2.72	4.37	4.05	3.55	3.25	3.38	2.43	1.71	1.57
PCMMSE	**3.60**	**3.30**	**2.91**	**2.76**	**4.42**	**4.11**	**3.63**	**3.34**	**3.41**	**2.49**	**1.77**	**1.63**

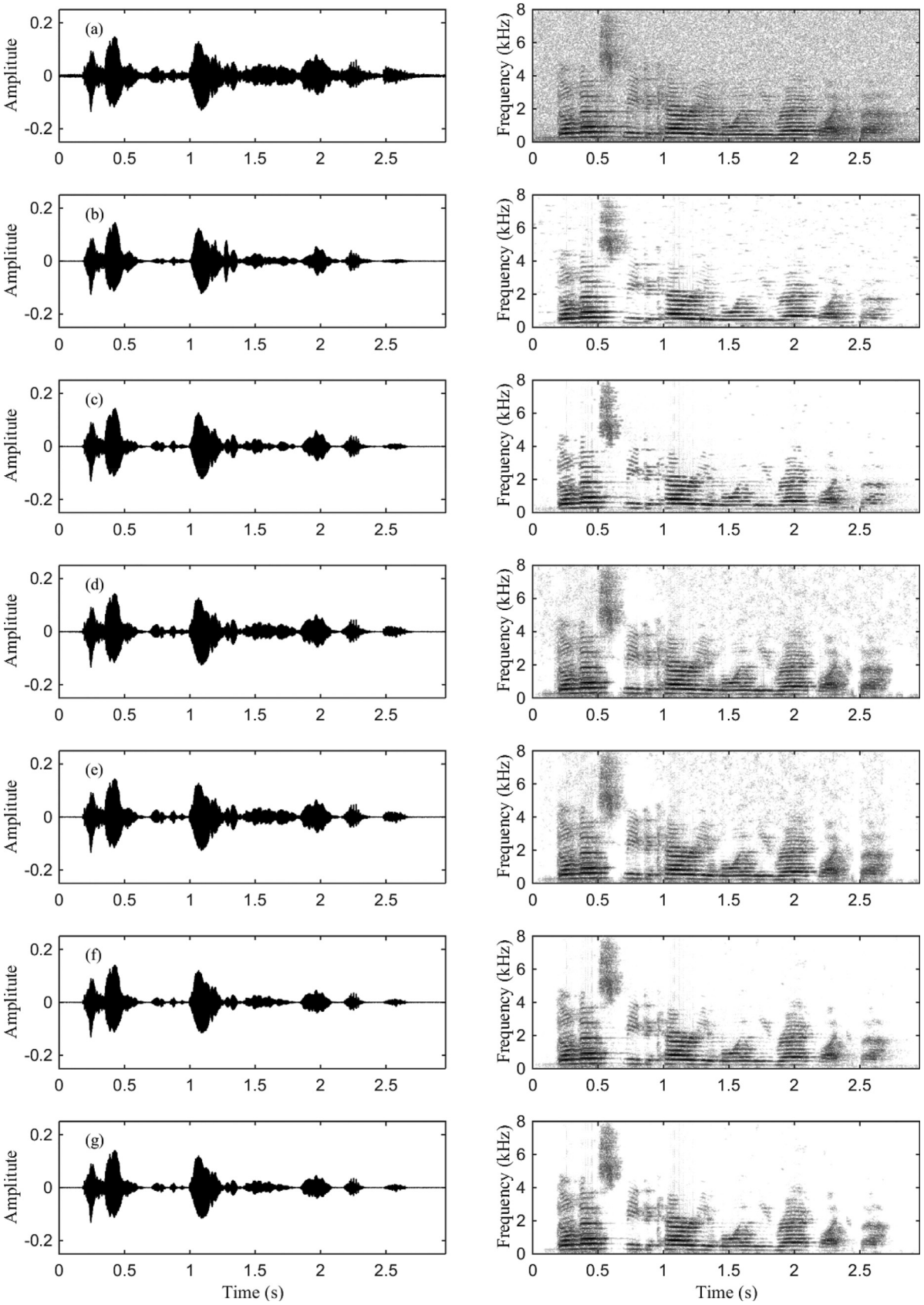

Fig. 4. Spectrograms of: (a) the recorded speech, (b) speech enhanced by the SS algorithm, (c) speech enhanced by the PSS algorithm, (d) speech enhanced by the GSVD algorithm, (e) speech enhanced by the PCGSVD algorithm, (f) speech enhanced by the CMMSE-GSVD-LPRE algorithm, (g) speech enhanced by the proposed PCMMSE-GSVD-LPRE algorithm.

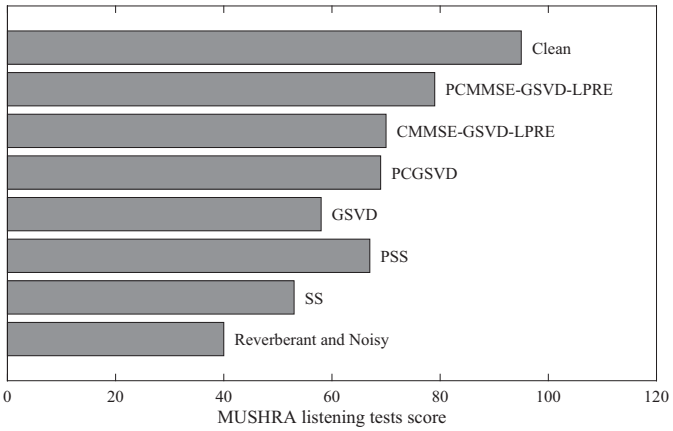

Fig. 5. Averaged MUSHRA Listening tests scores for recorded waveforms.

2013). Here, the updated version, i.e., SRMRnorm (Santos et al., 2014), is chosen to evaluate the proposed algorithm.

Table 4 gives the results of the averaged SRMRnorm scores over all the 200 speech sentences. It can be seen from Table 4 that the proposed PCMMSE-GSVD-LPRE algorithm is superior to all the other five competing algorithms under all kinds of testing conditions. One can also get that the SRMRnorm improvement of the proposed algorithm is higher in low SNR conditions than that in high SNR conditions.

4.5. Spectrogram comparison and MUSHRA listening test

Fig. 2 shows waveform and spectrogram of clean speech, reverberant and noisy speech, and speech enhanced by the six algorithms evaluated in this paper. The reverberant and noisy speech is generated by convoluting the clean speech with the RIR measured in the meeting room at distance 145 cm, and the white Gaussian noise is added to the reverberant speech with the input SNR value at 10 dB.

We can clearly see the improvement of the proposed algorithm over SS, GSVD and PCGSVD algorithm, which has lots of unwanted interferences in speech spectrograms. Although PSS algorithm and CMMSE-GSVD-LPRE algorithm achieve better performance than SS, GSVD, PCGSVD algorithm, PCMMSE-GSVD-LPRE has less *'musical noise'* than PSS and CMMSE-GSVD-LPRE, especially at the speech onsets.

To test the speech quality, MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) (Vincent, 2005) listening tests are conducted here. MUSHRA listening tests allow the comparison of high quality reference speech signals with several lower quality test speech signals. Here, we use the clean speech signals as the high quality reference ones, the reverberant and noisy signals and processed signals as the lower quality test speech signals. Listeners are asked to compare the high quality reference speech signals to several test speech signals sorted in random order, including the reference signal. Each subject is asked to assess the quality of each test sound (relative to the reference and other test sounds) by grading it on a quality scale between 0 and 100. Fig. 3 shows the averaged MUSHRA listening tests scores of ten listeners, where the proposed algorithm achieves the best listening performance.

5. Realistic experiments

In this section, we evaluate our algorithm in a realistic reverberant and noisy environment. Because the synchronized clean reference speech signal can not be acquired easily, we choose the non-intrusive metric, i.e. SRMRnorm and the spectrogram to measure these algorithms.

5.1. Experimental setup

The source signal is the same as the speech used in the simulation experiments in Section 4 and was played back by a HIVI-H4 active speaker. A 1/2″ microphone was used to record the speech in different rooms. Three kinds of rooms were tested, including a meeting room, a lecture room, and a small reverberation room. The volume and the reverberation time of each room can be found in Table 5. The position of the speaker was fixed throughout the recording, while the position of the microphone was moved such that the distances between the microphone and the speaker were 0.5 m, 1.0 m, 2.0 m and 4.0 m. A high-pass filtering was applied to the recordings before the dereverberation and denoising process to suppress the unwanted interference of alternating current (AC), where the cutoff frequency of the high-pass filter is 100 Hz.

5.2. SRMR results

Table 6 presents the results of the averaged SRMRnorm measurement for all the 200 recorded speech sentences and the corresponding enhanced speech sentences processed by the six algorithms evaluated in this paper. 'R + N' denotes the recorded reverberant and noisy speech. 'CMMSE' and 'PCMMSE' are shortening for the CMMSE-GSVD-LPRE algorithm and the proposed 'PCMMSE-GSVD-LPRE' algorithm.

It can be seen that the proposed PCMMSE-GSVD-LPRE algorithm has the largest SRMRnorm values among the six algorithms in all kinds of rooms and distances.

5.3. Spectrogram comparison and MUSHRA listening test

Fig. 4 shows the spectrogram of the recorded speech, speech enhanced by the SS algorithm, the PSS algorithm, the GSVD, the PCGSVD, the CMMSE-GSVD-LPRE algorithm and the proposed PCMMSE-GSVD-LPRE algorithm, respectively. It can be seen that the proposed PCMMSE-GSVD-LPRE algorithm has a better dereverberation result, and the speech pauses are more distinct comparing to the speech processed by the competing algorithms. Meanwhile, the residual interferences of the proposed PCMMSE-GSVD-LPRE algorithm are suppressed more efficiently. The MUSHRA listening tests are also conducted for these speech signals recorded in realistic environments, where the playback signals are used as the high quality reference speech signals, the recorded signals and processed signals are used as the lower quality test sounds. Fig. 5 presents the MUSHRA listening scores averaged by ten listeners. One can find that the proposed algorithm also has the best speech quality in realistic environments.

6. Conclusion

In this paper, we have extended our previous work to suppress both noise and late reverberation in the LP residual domain, and introduced a perceptually constrained optimal filter, which can make the residual interference un-perceivable. To calculate the perception based upper bound for the residual noise and reverberation, we reformulated the frequency to generalized singular value transformation equation in the LP residual domain. Both objective measurements and subjective measurements are evaluated in simulation and realistic experiments to show the effectiveness of the proposed algorithm under different input SNR and reverberation time conditions. By introducing the auditory masking properties into the GSVD-based approach, the residual interference is constrained to a level which is imperceivable, which can significantly improve the quality of the processed speech in terms of segSNR, PESQ, and SRMR metrics. MUSHRA listening tests conducted for both simulated and realistic experiments also show the better speech quality of the proposed algorithm.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.specom.2017.12.004.

References

Bees, D., Blostein, M., Kabal, P., 1991. Reverberant speech enhancement using cepstral processing. Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on. IEEE, pp. 977–980.

Benesty, J., Makino, S., 2005. Speech Enhancement. Springer Science & Business Media.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. 4. IEEE, pp. 208–211.

Bloom, P., Cain, G., 1982. Evaluation of two-input speech dereverberation techniques. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82. 7. IEEE, pp. 164–167.

Bloom, P.J., 1980. Evaluation of a dereverberation process by normal and impaired listeners. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80. 5. IEEE, pp. 500–503.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. Acoust. Speech Signal Process. IEEE Trans. 27 (2), 113–120.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. Speech Audio Process. IEEE Trans. 11 (5), 466–475.

Cohen, I., Berdugo, B., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. Signal Process. Lett. IEEE 9 (1), 12–15.

Cybenko, G., 1980. The numerical stability of the Levinson-Durbin algorithm for toeplitz systems of equations. SIAM J. Sci. Stat. Comput. 1 (3), 303–319.

Doclo, S., Moonen, M., 2002. GSVD-based optimal filtering for single and multi-microphone speech enhancement. Signal Process. IEEE Trans. 50 (9), 2230–2244.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoust. Speech Signal Process. IEEE Trans. 32 (6), 1109–1121.

Ephraim, Y., Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. Speech Audio Process. IEEE Trans. 3 (4), 251–266.

Falk, T.H., Zheng, C., Chan, W.Y., 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. Audio, Speech, Lang. Process. IEEE Trans. 18 (7), 1766–1774.

Gannot, S., Moonen, M., 2003. Subspace methods for multimicrophone speech dereverberation. EURASIP J. Appl. Signal Process. 2003, 1074–1090.

Garofolo, J.S., et al., 1988. Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database. 107 National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Gaubitch, N.D., Naylor, P.A., Ward, D.B., 2003. On the use of linear prediction for dereverberation of speech. Proc. Int. Workshop Acoust. Echo Noise Control. 1. pp. 99–102.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. Audio, Speech, Lang. Process. IEEE Trans. 20 (4), 1383–1393.

Gustafsson, S., Jax, P., Vary, P., 1998. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. IEEE, pp. 397–400.

Habets, E., 2005. Multi-channel speech dereverberation based on a statistical model of late reverberation. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05). IEEE International Conference on. 4. IEEE, pp. iv–173.

Habets, E.A., Gannot, S., Cohen, I., 2009. Late reverberant spectral variance estimation based on a statistical model. Signal Process. Lett. IEEE 16 (9), 770–773.

Hikichi, T., Delcroix, M., Miyoshi, M., 2007. Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP J. Adv. Signal Process. 2007 (1), 1–12.

Hirsch, H. G., 2005. FaNT - Filtering and Noise Adding Tool. http://aurora.hsnr.de/download.html. V.pA4.

ITU-2000, 2000. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P. 862.

Jabloun, F., Champagne, B., 2002. A perceptual signal subspace approach for speech enhancement in colored noise. Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. 1. IEEE, pp. I–569.

Jensen, J., Tan, Z., 2015. Minimum mean-square error estimation of mel-frequency cepstral features - a theoretically consistent approach. Audio, Speech, Lang. Process. IEEE Trans. 23 (1), 186–197.

Jensen, S.H., Hansen, P.C., Hansen, S.D., Sørensen, J.A., De Moor, B., 1995. Reduction of general broad-band noise in speech by truncated QSVD: implementation aspects. SVD Signal Process. III 459.

Jeub, M., Schäfer, M., Esch, T., Vary, P., 2010. Model-based dereverberation preserving binaural cues. Audio, Speech, Lang. Process. IEEE Trans. 18 (7), 1732–1745.

Johnston, J.D., 1988. Transform coding of audio signals using perceptual noise criteria. Sel. Areas Commun. IEEE J. 6 (2), 314–323.

Ju, G.H., Lee, L.S., 2007. A perceptually constrained GSVD-based approach for enhancing speech corrupted by colored noise. Audio, Speech, Lang. Process. IEEE Trans. 15 (1), 119–134.

Jueb, M., Schafer, M., Vary, P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. IEEE, pp. S.1–4.

Kalman, R.E., 1963. New methods in Wiener filtering theory. Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, edited by JL Bogdanoff and F. Kozin, John Wiley & Sons, New York.

Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M., 2009. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. Audio, Speech, Lang. Process. IEEE Trans. 17 (4), 534–545.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech. Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, pp. 1–4.

Krishnamoorthy, P., Prasanna, S.R.M., 2009. Reverberant speech enhancement by temporal and spectral processing. Audio, Speech, Lang. Process. IEEE Trans. 17 (3), 253–266.

Krishnamoorthy, P., Prasanna, S.R.M., 2011. Enhancement of noisy speech by temporal and spectral processing. Speech Commun. 53 (2), 154–174.

Kun, H., Wang, Y., Wang, D., 2015. Learning spectral mapping for speech dereverberation. Acoustics, Speech, and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 186–197.

Lebart, K., Boucher, J.-M., Denbigh, P., 2001. A new method based on spectral subtraction for speech dereverberation. Acta Acustica United Acustica 87 (3), 359–366.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC press.

Löllmann, H.W., Vary, P., 2009. A blind speech enhancement algorithm for the suppression of late reverberation and noise. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, pp. 3989–3992.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. Speech Audio Process. IEEE Trans. 9 (5), 504–512.

Mittal, U., Phamdo, N., 2000. Signal/noise KLT based approach for enhancing speech degraded by colored noise. Speech Audio Process. IEEE Trans. 8 (2), 159–167.

Miyoshi, M., Kaneda, Y., 1988. Inverse filtering of room acoustics. Acoust. Speech Signal Process. IEEE Trans. 36 (2), 145–152.

Mourjopoulos, J., 1985. On the variation and invertibility of room impulse response functions. J. Sound Vib. 102 (2), 217–228.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H., 2008. Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, pp. 85–88.

Naylor, P.A., Gaubitch, N.D., 2010. Speech Dereverberation. Springer Science & Business Media.

Neely, S.T., Allen, J.B., 1979. Invertibility of a room impulse response. J. Acoust. Soc. Am. 66 (1), 165–169.

Painter, T., Spanias, A., 2000. Perceptual coding of digital audio. Proc. IEEE 88 (4), 451–515.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Prentice Hall.

Radlovic, B.D., Williamson, R.C., Kennedy, R.A., 2000. Equalization in an acoustic reverberant environment: robustness results. Speech Audio Process. IEEE Trans. 8 (3), 311–319.

Rezayee, A., Gazor, S., 2001. An adaptive KLT approach for speech enhancement. Speech Audio Process. IEEE Trans. 9 (2), 87–95.

Rotili, R., Principi, E., Squartini, S., Schuller, B., 2011. Real-time speech recognition in a multi-talker reverberated acoustic scenario. Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Springer, pp. 379–386.

Santos, J.F., Senoussaoui, M., Falk, T.H., 2014. An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation. International Workshop on Acoustic Signal Enhancement (IWAENC).

Schroeder, M.R., Atal, B.S., Hall, J., 1979. Optimizing digital speech coders by exploiting masking properties of the human ear. J. Acoust. Soc. Am. 66 (6), 1647–1652.

Sehr, A., Habets, E.A., Maas, R., Kellermann, W., 2010. Towards a better understanding of the effect of reverberation on speech recognition performance. Proc. IWAENC.

Sinha, D., Tewfik, A.H., 1993. Low bit rate transparent audio compression using adapted wavelets. Signal Process. IEEE Trans. 41 (12), 3463–3479.

Spriet, A., MooNEN, M., Wouters, J., 2002. A multi-channel subband generalized singular value decomposition approach to speech enhancement. Eur. Trans. Telecommun. 13 (2), 149–158.

Subramaniam, S., Petropulu, A.P., Wendt, C., 1996. Cepstrum-based deconvolution for speech dereverberation. Speech Audio Process. IEEE Trans. 4 (5), 392–396.

Thiemann, J., 2001. Acoustic noise suppression for speech signals using auditory masking effects. Department of Electrical and Computer Engineering, McGill University.

Tokuno, H., Kirkeby, O., Nelson, P.A., Hamada, H., 1997. Inverse filter of sound reproduction systems using regularization. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 80 (5), 809–820.

Van Huffel, S., 1993. Enhanced resolution based on minimum variance estimation and exponential data modeling. Signal Process. 33 (3), 333–355.

Varga, A., Steeneken, H.J., 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 Commun. 12 (3), 247–251.

Vincent, E., 2005. MUSHRAM: A MATLAB interface for MUSHRA listening tests. http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. Speech Audio Process. IEEE Trans. 7 (2), 126–137.

Wu, M., Wang, D., 2006. A two-stage algorithm for one-microphone reverberant speech enhancement. Audio, Speech, Lang. Process. IEEE Trans. 14 (3), 774–784.
Yoshioka, T., Nakatani, T., Miyoshi, M., 2009. Integrated speech enhancement method using noise suppression and dereverberation. Audio, Speech, Lang. Process. IEEE Trans. 17 (2), 231–246.
Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., 2012. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. Signal Process. Mag. IEEE 29 (6), 114–126.
Zheng, C., Peng, R., Li, J., Li, X., 2014. A constrained MMSE LP residual etimator for speech dereverberation in noisy environments. Signal Process. Lett. IEEE 21 (12), 1462–1466.
Zheng, C., Zhou, Y., Hu, X., Li, X., 2010. Speech enhancement based on the structure of noise power spectral density. Signal Processing Conference, 2010 18th European. IEEE, pp. 1519–1523.