Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Speech Enhancement with Dual-path Multi-Channel Linear Prediction Filter and Multi-norm Beamforming

Chengyuan Qin[1], *Wenmeng Xiong*[1,*], *Jing Zhou*[2], *Maoshen Jia*[1], *Changchun Bao*[1]

[1]Speech and Audio Signal Processing Laboratory, Beijing University of Technology, China
[2]Institute of AI (TeleAI), China telecom, China

qinchengyuan@emails.bjut.edu.cn, wenmeng.xiong@bjut.edu.cn, zhouj100@chinatelecom.cn, jiamaoshen@bjut.edu.cn, baochch@bjut.edu.cn

Abstract

In this paper, we propose a speech enhancement method using dual-path Multi-Channel Linear Prediction (MCLP) filters and multi-norm beamforming. Specifically, the MCLP part in the proposed method is designed with dual-path filters in both time and frequency dimensions. For the beamforming part, we minimize the power of the microphone array output as well as the l_1 norm of the denoised signals while preserving source signals from the target directions. An efficient method to select the prediction orders in the dual-path filters is also proposed, which is robust for signals with different reverberation time (T_{60}) values and can be applied to other MCLP-based methods. Evaluations demonstrate that our proposed method outperforms the baseline methods for speech enhancement, particularly in high reverberation scenarios.

Index Terms: Speech enhancement, microphone array, dual-path multi-channel linear prediction, beamforming, sparsity

1. Introduction

Speech enhancement technique is of great importance for numerous applications such as automatic speech recognition, human-machine interaction, and smart home devices[1, 2, 3]. Conventional speech enhancement methods, including spectral subtraction [4], wiener filtering [5], and subspace-based methods [6], have been widely investigated and proven to be effective. However, their performance degrades severely in real-world applications where noise and reverberation are present.

To address the challenge of reverberation, numerous methods have been proposed based on the MCLP filter [7][8]. In [9], the Generalized Weight Prediction Error (GWPE) method reduces the late reverberation by minimizing the temporal Hadamard-Fischer (HF) mutual correlation of speech signals. In addition to dereverberation, methods based on beamforming [10] have been proven effective for denoising. Plenty of work [11, 12, 13] has been proposed for the integration of the MCLP-based methods and the beamforming-based methods for simultaneous dereverberation and denoising. For example, in [11, 12], Minimum Variance Distortionless Response (MVDR) beamforming and WPE are employed in a cascade framework. In [13] and [14], the MVDR beamforming and Minimum-Power Distortionless Response (MPDR) beamforming are combined with WPE in a unified joint optimization problem, respectively.

Traditional MCLP-based speech enhancement methods remove late reverberations by estimating temporal filters at each frequency bin of the time-frequency (TF) domain microphone signals. However, numerous deep-learning based speech enhancement techniques, including DPRNN [15], FullSubNet [16], TF-GridNet [17], SpatialNet [18], and DasFormer [19], have demonstrated significantly superior performance using dual-path cross-narrow band architectures. In these work, frequential dependencies as well as temporal correlations have been leveraged for a more comprehensive modeling by neural networks to learn the mappings between the inputs and the learning targets.

In this paper, we propose a speech enhancement method using dual-path MCLP filters and multi-norm beamforming. Specifically, for the MCLP part, exploiting both temporal correlations and frequential dependencies of signals in TF domain, dual-path filters in both temporal and frequency dimensions are designed to remove the late reverberations comprehensively. The l_1 norm constraint is incorporated into the cost functions of both dual-path MCLP and multi-norm beamforming because information-bearing signals, such as speech signals, exhibit sparsity in their short-time Fourier Transform (STFT) coefficients. Additionally, we propose an efficient method to determine the prediction orders in the temporal and frequential filters of the MCLP part. By calculating the Pearson correlation coefficients between signals on a single microphone as a function of different time or frequency lags and selecting an appropriate threshold, the corresponding time or frequency lag can be considered as the approximate optimal prediction orders. This method is robust for signals with varying reverberation time (T_{60}) values and can be applied to other MCLP-based methods. Experiments demonstrate the advantages of our proposed method over the baseline methods.

2. Preliminaries

2.1. Microphone array signal model

Considering that Q far-field wideband acoustic sources impinge on M microphones in a noisy and reverberant room. The signals received at the microphone array in the TF domain are approximately formulated as a K order convolution between the STFT of the room impulse response (RIR) $\mathbf{h}_q(n,\omega)$ and the STFT signal $s_q(n,\omega)$ of q^{th} source along the time frame axis for each frequency bin [20], $n \in \{1,...,N\}$ and $\omega \in \{1,...,\Omega\}$ denote the time frame and frequency bin indices, respectively. N and Ω denote the maximum time frame and frequency bin indices, respectively. The TF domain microphone array signals can be given as:

$$\mathbf{y}(n,\omega) = \sum_{q=1}^{Q} \sum_{k'=0}^{K} \mathbf{h}_q(k',\omega) s_q(n-k',\omega) + \mathbf{n}(n,\omega), \quad (1)$$

where $\mathbf{n}(n,\omega)$ denotes the additive noise at the microphone array. The microphone array signals $\mathbf{y}(n,\omega)$ can be reformulated

* Corresponding author

10.21437/Interspeech.2025-1502

as the sum of the early reverberant component, the late reverberant component and the noise:

$$\mathbf{y}(n,\omega) = \mathbf{y}_e(n,\omega) + \mathbf{y}_l(n,\omega) + \mathbf{n}(n,\omega). \quad (2)$$

where $\mathbf{y}_e(n,\omega) = \sum_{q=1}^{Q}\sum_{k'=0}^{k_e}\mathbf{h}_q(k',\omega)s_q(n-k',\omega)$ denotes the early reverberant component which contains the direct-path signals and few early reflections, $\mathbf{y}_l(n,\omega) = \sum_{q=1}^{Q}\sum_{k'=k_e}^{K}\mathbf{h}_q(k',\omega)s_q(n-k',\omega)$ is the late reverberant component which contains all other reflections. k_e is the convolutional order that separates the early reverberant component and the late reverberant component.

When $k' = 0$, the RIR $\mathbf{h}_q(0,\omega)$ is also referred to the steering vector of the q^{th} source from direction θ_q to the microphone array, defined as $\mathbf{a}(\theta_q,\omega) = [1, e^{-j\omega\Delta t_1(\theta_q)}, ..., e^{-j\omega\Delta t_m(\theta_q)}, ..., e^{-j\omega\Delta t_{M-1}(\theta_q)}] \in \mathbb{C}^{M\times 1}$, where the first microphone is taken as the reference, $\Delta t_m(\theta_q)$ is the time delay from the m^{th} microphone to the reference microphone.

2.2. MCLP-based methods

In MCLP-based methods, the dereverberated signals $\mathbf{d}(n,\omega)$ can be given as:

$$\mathbf{d}(n,\omega) = \mathbf{y}(n,\omega) - \mathbf{G}_t^H(\omega)\tilde{\mathbf{y}}_t(n,\omega). \quad (3)$$

where $\mathbf{G}_t(\omega) = [\mathbf{g}_{t_1}(\omega), ..., \mathbf{g}_{t_m}(\omega), ..., \mathbf{g}_{t_M}(\omega)] \in \mathbb{C}^{K_t M\times M}$ is a temporal filter matrix at frequency bin ω , $\mathbf{g}_{t_m}(\omega)$ is the temporal filter vector for the m^{th} microphone. $\tilde{\mathbf{y}}_t(n,\omega) = \left[\mathbf{y}^T(n-\Delta_t-1,\omega), ..., \mathbf{y}^T(n-\Delta_t-K_t,\omega)\right]^T$ is the stacked observation signal matrix, Δ_t denotes the delay tap index of the beginning frame of the late reverberation component, K_t denotes the prediction order. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and complex conjugate transpose operator, respectively. The MCLP-based methods estimate the filter matrix $\mathbf{G}_t(\omega)$ to obtain dereverberated speech signals.

3. Proposed Algorithm

The proposed algorithm consists of two parts: dual-path MCLP filters for dereverberation and multi-norm beamforming for denoising.

3.1. Dual-path MCLP filter branch

3.1.1. Dereverberated signals with dual-path filters

Let us define $\mathbf{G}_f(n) = [\mathbf{g}_{f_1}(n), ..., \mathbf{g}_{f_m}(n), ..., \mathbf{g}_{f_M}(n)] \in \mathbb{C}^{((2K_f+1)M)\times M}$ as the frequential filter matrix at time frame n, where $\mathbf{g}_{f_m}(n)$ is the filter for the m^{th} microphone at time frame n, K_f denotes the prediction order in frequential dimension. With both temporal filter matrix in (3) and the proposed frequential filter matrix, the early component in TF domain microphone array signals can be given as:

$$\mathbf{x}(n,\omega) = \mathbf{y}(n,\omega) - \mathbf{G}_t^H(\omega)\tilde{\mathbf{y}}_t(n,\omega) - \mathbf{G}_f^H(n)\tilde{\mathbf{y}}_f(n,\omega), \quad (4)$$

where $\tilde{\mathbf{y}}_f(n,\omega) = [\mathbf{y}^T(n,\omega-K_f), ..., \mathbf{y}^T(n,\omega), ..., \mathbf{y}^T(n,\omega+K_f)]^T$ is the wide band observation signal matrix at time frame n. The dual-path filters in both temporal and frequential dimensions yield more comprehensive modeling of the late reverberation.

3.1.2. Dual-path MCLP filter

In the dual-path MCLP filter for dereverberation, we minimize the l_2 norm as well as the l_1 norm of the dereverberated signals, as the information bearing signals such as speech signals exhibit sparsity in their STFT coefficients:

$$\left\{\hat{\mathbf{G}}_t(\omega), \hat{\mathbf{G}}_f(n)\right\} = \arg\min_{\mathbf{G}_t,\mathbf{G}_f}\sum_{n=1}^{N}\sum_{\omega=1}^{\Omega}(\|\mathbf{x}(n,\omega)\|_2^2 + \lambda_{\mathbf{z}}\|\mathbf{x}(n,\omega)\|_1), \quad (5)$$

where $\mathbf{x}(n,\omega)$ is given in (4), $\lambda_{\mathbf{z}}$ is sparsity penalization parameter, $\hat{\mathbf{G}}_t(\omega)$, $\hat{\mathbf{G}}_f(n)$ are the estimated $\mathbf{G}_t$ and $\mathbf{G}_f$, respectively. For the sake of simplicity, ω and n in $\mathbf{G}_t(\omega)$, $\mathbf{G}_f(n)$, $\tilde{\mathbf{y}}_t(n,\omega)$, and $\tilde{\mathbf{y}}_f(n,\omega)$ are omitted in the following of this paper. The problem (5) can be solved via Proximal Alternating Linearized Minimization (PALM) [21] method. By introducing an extra variable $\mathbf{z}(n,\omega) = \mathbf{x}(n,\omega)$, the augmented Lagrangian of (5) can be given as:

$$\mathcal{L}(\mathbf{G}_t,\mathbf{G}_f,\mathbf{z}(n,\omega),\boldsymbol{\eta}) = \sum_{n=1}^{N}\sum_{\omega=1}^{\Omega}(\|\mathbf{x}(n,\omega)\|_2^2 + \lambda_{\mathbf{z}}\|\mathbf{z}(n,\omega)\|_1 + \mathcal{R}e\{\boldsymbol{\eta}^H(\mathbf{x}(n,\omega)-\mathbf{z}(n,\omega))\} + \frac{1}{2\rho_{\mathbf{G}}}\|\mathbf{x}(n,\omega)-\mathbf{z}(n,\omega)\|_2^2), \quad (6)$$

where $\mathcal{R}e\{\cdot\}$ is the real part operator, $\rho_{\mathbf{G}}$ is the penalization parameter of the convex term and $\boldsymbol{\eta}$ is the Lagrange multiplier vector. Utilising the method in [22], the solutions for $\mathbf{G}_t$, $\mathbf{G}_f$, $\mathbf{z}(n,\omega)$, and $\boldsymbol{\eta}$ in the $(l+1)^{th}$ iteration are given as:

$$\mathbf{G}_t^{(l+1)} = \left(\sum_{n=1}^{N}\left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t^H\right)^{-1} \cdot \sum_{n=1}^{N}\left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_t\mathbf{y}^H(n,\omega) + \frac{1}{2}\tilde{\mathbf{y}}_t\boldsymbol{\eta}^{(l)H} - \left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_f^H\mathbf{G}_f^{(l)} - \frac{1}{2\rho_{\mathbf{G}}}\tilde{\mathbf{y}}_t\mathbf{z}^{(l)H}(n,\omega), \quad (7)$$

and:

$$\mathbf{G}_f^{(l+1)} = \left(\sum_{\omega=1}^{\Omega}\left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_f\tilde{\mathbf{y}}_f^H\right)^{-1} \cdot \sum_{\omega=1}^{\Omega}\left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_f\mathbf{y}^H(n,\omega) + \frac{1}{2}\tilde{\mathbf{y}}_f\boldsymbol{\eta}^{(l)H} - \left(1+\frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{y}}_f\tilde{\mathbf{y}}_t^H\mathbf{G}_t^{(l+1)} - \frac{1}{2\rho_{\mathbf{G}}}\tilde{\mathbf{y}}_f\mathbf{z}^{(l)H}(n,\omega), \quad (8)$$

and:

$$\mathbf{z}^{(l+1)}(n,\omega) = \mathcal{S}_{\lambda_{\mathbf{z}}/\mu_{\mathbf{z}}}\left(\mathbf{z}^{(l)}(n,\omega) - \frac{1}{\mu_{\mathbf{z}}}\nabla_{\mathbf{z}}V(\mathbf{G}_t^{(l+1)},\mathbf{G}_f^{(l+1)},\mathbf{z}^{(l)}(n,\omega),\boldsymbol{\eta}^{(l)})\right), \quad (9)$$

where $V(\mathbf{G}_t^{(l+1)},\mathbf{G}_f^{(l+1)},\mathbf{z}^{(l)}(n,\omega),\boldsymbol{\eta}^{(l)}) = \mathcal{R}e\{\boldsymbol{\eta}^{(l)H}(\mathbf{y}(n,\omega)-\mathbf{G}_t^{(l+1)H}\tilde{\mathbf{y}}_t-\mathbf{G}_f^{(l+1)H}\tilde{\mathbf{y}}_f-\mathbf{z}^{(l)}(n,\omega))\} + \frac{1}{2\rho_{\mathbf{G}}}\|\mathbf{y}(n,\omega)-\mathbf{G}_t^{(l+1)H}\tilde{\mathbf{y}}_t-\mathbf{G}_f^{(l+1)H}\tilde{\mathbf{y}}_f-\mathbf{z}^{(l)}(n,\omega)\|_2^2$, $\mu_{\mathbf{z}}$ is

thresholding parameter and $\mathcal{S}_{\lambda_{\mathbf{z}}/\mu_{\mathbf{z}}}(\mathbf{v})$ is the soft thresholding operator [23] of the vector $\mathbf{v}$ such that:

$$S_{\lambda_{\mathbf{z}}/\mu_{\mathbf{z}}}(\mathbf{v}) = \begin{cases} \mathbf{v} - \lambda_{\mathbf{z}}/\mu_{\mathbf{z}}, & \text{if } \mathbf{v} \geq \lambda_{\mathbf{z}}/\mu_{\mathbf{z}}, \\ \mathbf{v} + \lambda_{\mathbf{z}}/\mu_{\mathbf{z}}, & \text{if } \mathbf{v} \leq -\lambda_{\mathbf{z}}/\mu_{\mathbf{z}}, \\ 0, & \text{otherwise} \end{cases}$$

and:

$$\boldsymbol{\eta}^{(l+1)} = \boldsymbol{\eta}^{(l)} + \gamma(\mathbf{y}(n,\omega) - \mathbf{G}_t^{(l+1)H}\tilde{\mathbf{y}}_t - \mathbf{G}_f^{(l+1)H}\tilde{\mathbf{y}}_f - \mathbf{z}^{(l+1)}(n,\omega)), \tag{10}$$

where γ is the step size parameter.

With $\hat{\mathbf{G}}_t$ and $\hat{\mathbf{G}}_f$ estimated in (5), the estimated early component of microphone signals can be given as:

$$\hat{\mathbf{x}}(n,\omega) = \mathbf{y}(n,\omega) - \hat{\mathbf{G}}_t^H\tilde{\mathbf{y}}_t - \hat{\mathbf{G}}_f^H\tilde{\mathbf{y}}_f. \tag{11}$$

3.2. Multi-norm beamforming branch

The multi-norm beamforming is applied to the signals dereverberated by the dual-path MCLP filters, namely $\hat{\mathbf{x}}(n,\omega)$ in (11). The power as well as the l_1 norm of the denoised signals are minimized while the signals from the target direction of arrival (DOA) are preserved:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \left(\|\mathbf{w}^H\hat{\mathbf{x}}(n,\omega)\|_2^2 + \lambda_{\mathbf{w}}\|\mathbf{w}^H\hat{\mathbf{x}}(n,\omega)\|_1 \right), \quad \text{s.t.} \quad \mathbf{w}^H\mathbf{a}(\theta_s) = 1, \tag{12}$$

where $\mathbf{w}$ is the beamforming filter vector, $\hat{\mathbf{w}}$ is the estimated $\mathbf{w}$ and $\lambda_{\mathbf{w}}$ is the sparsity penalization parameter.

The problem (12) can be solved via Alternating Direction Method of Multipliers (ADMM) method [24]. Introducing an extra variable $z_{\mathbf{w}}(n) = \mathbf{w}^H\hat{\mathbf{x}}(n,\omega)$, the augmented Lagrangian of (12) can be given as:

$$\mathcal{L}(\mathbf{w}, z_{\mathbf{w}}(n), \eta_{\mathbf{w}}) = \sum_{n=1}^{N} \Big(\|\mathbf{w}^H\hat{\mathbf{x}}(n,\omega)\|_2^2 + \lambda_{\mathbf{w}} \|z_{\mathbf{w}}(n)\|_1 + \mathcal{R}e\{\eta_{\mathbf{w}}^H(\mathbf{w}^H\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}(n))\} + \frac{1}{2\rho_{\mathbf{w}}}\|\mathbf{w}^H\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}(n)\|_2^2 \Big), \quad \text{s.t. } \mathbf{w}^H\mathbf{a}(\theta_s) = 1. \tag{13}$$

where $\eta_{\mathbf{w}}$ is Lagrange multiplier. The problem (13) can be solved via some iterative steps similar to problem (6). In the $(l+1)^{th}$ iteration, the details for solving $\mathbf{w}^{(l+1)}$ will be given in Appendix. $z_{\mathbf{w}}^{(l+1)}(n)$ can be given as:

$$z_{\mathbf{w}}^{(l+1)}(n) = \mathcal{S}_{\lambda_{\mathbf{w}}/\mu_{\mathbf{w}}} \left(z_{\mathbf{w}}^{(l)}(n) - \frac{1}{\mu_{\mathbf{w}}}\nabla_{\mathbf{z}_{\mathbf{w}}} V(\mathbf{w}^{(l+1)}, z_{\mathbf{w}}^{(l)}(n), \eta_{\mathbf{w}}^{(l)}) \right), \tag{14}$$

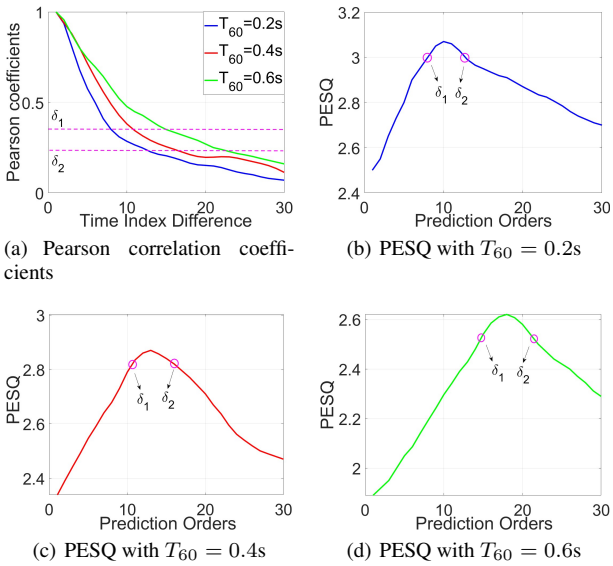
where $V(\mathbf{w}^{(l+1)}, z_{\mathbf{w}}^{(l)}(n), \eta_{\mathbf{w}}^{(l)}) = \mathcal{R}e\{\eta_{\mathbf{w}}^{(l)H}(\mathbf{w}^{(l+1)H}\hat{\mathbf{x}}(n,\omega) z_{\mathbf{w}}^{(l)}(n))\} + \frac{1}{2\rho_{\mathbf{w}}}\|\mathbf{w}^{(l+1)H}\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}^{(l)}(n)\|_2^2$, $\mu_{\mathbf{w}}$ is thresholding parameter and $\mathcal{S}_{\lambda_{\mathbf{w}}/\mu_{\mathbf{w}}}(\cdot)$ is the soft thresholding operator [23], $\eta_{\mathbf{w}}^{(l+1)}$ can be given as:

$$\eta_{\mathbf{w}}^{(l+1)} = \eta_{\mathbf{w}}^{(l)} + \gamma_{\mathbf{w}}(\mathbf{w}^{(l+1)H}\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}^{(l+1)}(n)). \tag{15}$$

where $\gamma_{\mathbf{w}}$ is the step size parameter.

With the estimation result $\hat{\mathbf{w}}$ obtained by (12), the enhanced speech signals can be given as:

$$\hat{s}(n,\omega) = \hat{\mathbf{w}}^H\hat{\mathbf{x}}(n,\omega). \tag{16}$$



(a) Pearson correlation coefficients (b) PESQ with $T_{60} = 0.2$s (c) PESQ with $T_{60} = 0.4$s (d) PESQ with $T_{60} = 0.6$s

Figure 1: *Pearson correlation coefficients and PESQ with different T_{60} values.*

4. Proposed Prediction Order Selection Method

In this section, the prediction order selection method is present. Firstly, let us define $y_{1_t}^i(t)$, $t = 0, ..., T$ as the i^{th} individual sample of the temporal signals at the reference microphone at time index t in the Monte Carlo experiments. The Pearson correlation coefficients between $y_{1_t}^i(0)$ and $y_{1_t}^i(t)$ can be given as:

$$\varrho_{y_{1_t}}(0,t) = \frac{\sum_{i=1}^{I}(y_{1_t}^i(0) - \bar{y}_{1_t}(0))(y_{1_t}^i(t) - \bar{y}_{1_t}(t))}{\sqrt{\sum_{i=1}^{I}(y_{1_t}^i(0) - \bar{y}_{1_t}(0))^2 \sum_{i=1}^{I}(y_{1_t}^i(t) - \bar{y}_{1_t}(t))^2}}, \tag{17}$$

where I denotes the total sample number, $\bar{(\cdot)}$ denotes the mean value operator. In Fig. 1(a) the Pearson correlation coefficients $\varrho_{y_{1_t}}(0,t)$ in (17) are presented as a function of time lag t under different T_{60} with $I = 50$, namely 50 Monte Carlo experiments. It can be observed that, due to the appearance of the late reverberation, $\varrho_{y_{1_t}}(0,t)$ decreases gently as the time lag increases.

In our experiments, we have found that, for different T_{60} configurations, a single threshold δ can be selected within the range $\delta_1 \leq \delta \leq \delta_2$, such that the corresponding time lag is also an approximate optimal prediction order of the temporal filters in our proposed MCLP based method. This can be illustrated in Fig.1 (b)-(d), where the Perceptual Evaluation of Speech Quality (PESQ) of our proposed method is plotted as a function of the temporal filter prediction order K_t under different T_{60} values, while the frequential filter prediction order K_f is fixed. The prediction orders corresponding to the rose-red circles of PESQ values in Fig. 1(b)-(d) are equal to the time lags corresponding to δ_1 and δ_2 in Fig. 1(a). Hence, in our following work, the optimal prediction order in the time filters is selected as:

$$K_t = 1/2(K_{\delta_1} + K_{\delta_2}), \tag{18}$$

where K_{δ_1} and K_{δ_2} are the time lags corresponding to δ_1 and δ_2 in Fig. 1(a), respectively. In this manner, an approximate optimal prediction order for different T_{60} values can be determined

by calculating the Pearson correlation coefficients between the temporal signals at different time lags, which is efficient for real-world applications. This optimal prediction order selection approach can be applied to other MCLP-based method as well.

The optimal prediction order in the frequential filters in our proposed method can be determined in a similar way.

5. Simulation Experiments

In this section, numerical simulations are presented to illustrate the validity of the proposed method. Source signals are speeches from the TIMIT database [25] sampled at 16 kHz. A uniform linear microphone array (ULA) composed of 8 microphones with the inter-element space equal to 0.03 m is utilised. The RIRs are generated using the image method [26] in a room with size of 6 m $\times$ 6 m $\times$ 3 m and the additive noises are Gaussian white. The Hanning windows with the length of 512 samples are used for signal analysis and synthesis with hop size of half-length. 100 Monte Carlo simulations are conducted for each configuration.

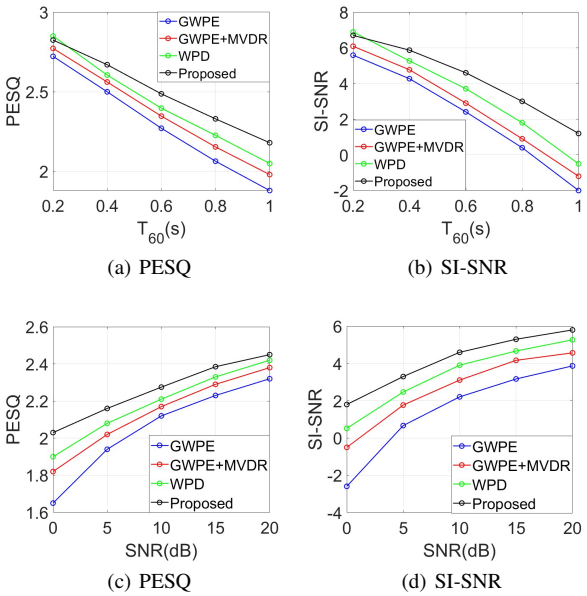
In all the experiments, our proposed method is compared with the baseline methods including the GWPE method [9] (legend: "GWPE") , the cascade of GWPE and MVDR method [12] (legend: "GWPE+MVDR") and the Weighted Power minimization Distortionless response (WPD) beamforming [14] (legend: "WPD") method. The PESQ and Scale-Invariant Signal-to-Noise Ratio (SI-SNR) are used as the evaluation metrics.

In Fig. 2(a) and 2(b), PESQ and SI-SNR obtained by the baseline methods and the proposed method are plotted as a function of T_{60}, with SNR set to 25 dB. $K_t = \{10, 14, 18, 22, 24\}$ and $K_f = \{2, 4, 6, 8, 10\}$ for $T_{60} = \{0.2s, 0.4s, 0.6s, 0.8s, 1s\}$. It can be seen that, with the dual-path MCLP filters which can more comprehensively removing the late reverberation, the advantage of the proposed method becomes more significant in high reverberation cases. When T_{60} is small enough (200 ms in this experiment), the performance of the proposed method is slightly worse than that of the baseline method WPD. One plausible explanation is that temporal filters alone are sufficient for removing the late reverberation when T_{60} is sufficiently small. A trade-off between dual-path MCLP filters or temporal-only MCLP filters should be considered based on varying T_{60} in our proposed method.

Fig. 2(c) and 2(d) show PESQ and SI-SNR as a function of SNR, with T_{60} set to $0.3s$. $K_t = 12$ and $K_f = 3$. It can be observed that, as a joint optimization method, WPD outperforms the cascaded GWPE and MVDR method ("GWPE+MVDR"). However, due to the l_1 norm constraint on the denoised signals, our proposed method exhibits a more powerful denoising capability and outperforms all the baseline methods across all the SNRs.

6. Conclusion

In this paper, we propose a speech enhancement method using dual-path MCLP filters and multi-norm beamforming. The proposed method demonstrates superior performance in both dereverberation and denoising compared to the baseline methods, particularly in high reverberation scenarios. In addition, we have proposed an efficient method for selecting optimal prediction orders in both temporal and frequential filters for the proposed method, which can also be applied to other MCLP-based methods. Simulation results validate the effectiveness of our proposed method.



(a) PESQ (b) SI-SNR

(c) PESQ (d) SI-SNR

Figure 2: *Performance metrics under different configurations.*

7. Appendix

To solve $\mathbf{w}^{(l+1)}$, a new augmented Lagrangian can be derived as:

$$\mathcal{L}(\mathbf{w}^{(l+1)}, \eta_1) = \sum_{n=1}^{N}(\|\mathbf{w}^{(l+1)H}\hat{\mathbf{x}}(n,\omega)\|_2^2 + \mathcal{R}e\{\eta_{\mathbf{w}}^{(l)H}(\mathbf{w}^{(l+1)H}$$
$$\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}^{(l)}(n,\omega))\} + \frac{1}{2\rho_{\mathbf{w}}}\|\mathbf{w}^{(l+1)H}\hat{\mathbf{x}}(n,\omega) - z_{\mathbf{w}}^{(l)}(n,\omega)\|_2^2) +$$
$$\mathcal{R}e\{\eta_1^H(\mathbf{w}^{(l+1)H}\mathbf{a}(\theta_s) - 1)\} + \frac{1}{2\rho_1}\|\mathbf{w}^{(l+1)H}\mathbf{a}(\theta_s) - 1\|_2^2, \tag{19}$$

The problem (19) can be solved via several iterative steps. In the $(j+1)^{th}$ iteration, $\mathbf{w}$ can be given as:

$$\mathbf{w}^{(l+1,j+1)} = \left(\sum_{n=1}^{N}(1 + \frac{1}{2\rho_{\mathbf{w}}})\hat{\mathbf{x}}(n,\omega)\hat{\mathbf{x}}^H(n,\omega) + \frac{\mathbf{a}(\theta_s)\mathbf{a}^H(\theta_s)}{2\rho_1}\right)^{-1}$$
$$\left(\sum_{n=1}^{N}\hat{\mathbf{x}}(n,\omega)(\frac{1}{2\rho_{\mathbf{w}}}z_{\mathbf{w}}^{(l)H}(n,\omega) - \frac{1}{2}\eta_{\mathbf{w}}^{(l)H}) + \mathbf{a}(\theta_s)(\frac{1}{2\rho_1} - \frac{1}{2}\eta_1^{(j)H})\right), \tag{20}$$

where $\eta_1^{(j+1)}$ can be given as:

$$\eta_1^{(j+1)} = \eta_1^{(j)} + \gamma_1(\mathbf{w}^{(l+1,j+1)H}\mathbf{a}(\theta_s) - 1). \tag{21}$$

8. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62101013).

9. References

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[2] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 75–95, Aug. 1998.

[3] R. Maas, E. A. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, p. 297–300.

[4] Y. Ephraim and I. Cohen, *Recent advancements in speech enhancement*. The Electronic Handbook, 2006.

[5] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, p. 2162–2176, Dec. 2015.

[6] J. Jensen and R. Heusdens, "Improved subspace-based single-channel speech enhancement using generalized super-gaussian priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 862–872, Mar. 2007.

[7] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, p. 430–440, Feb. 2007.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, p. 1717–1731, Sep. 2010.

[9] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse respons shortening," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, p. 2707–2720, Dec. 2012.

[10] J. Benesty and J. Chen, *Microphone Array Signal Processing*. Germany: Springer-Verlag, 2008.

[11] M. Delcroix, "Strategies for distant speech recognition in reverberant environments," no. 1, p. 60, Aug. 2015.

[12] W. Yang and G. Huang, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," in *International Workshop on Acoustic Signal Enhancement(IWAENC)*, 2018.

[13] M. Togami, "Multichannel online speech dereverberation under noisy environments," pp. 1078–1082, Aug. 2015.

[14] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process*, p. 903–907, Jun. 2019.

[15] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Proc, Oct. 2020, p. 46–50.

[16] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet:dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Mar. 2022, p. 6857–6861.

[17] Z. Wang, S. Cornell, and S. Choi, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," *arXiv preprint arXiv*, p. 2209.03952, 2022.

[18] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.

[19] S. Wang, X. Kong, X. Peng, H. Movassagh, V. Prakash, and Y. Lu, "Dasformer: Deep alternating spectrogram transformer for multi/single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Island, Jun. 2023, pp. 1–5.

[20] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, p. 546–555, May. 2009.

[21] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *J. Acoust. Soc. Amer*, vol. 146, no. 1/2, p. 459–494, 2014.

[22] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic, 1982.

[23] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn*, no. 1, pp. 1–122, 2011.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM*, 1993.

[26] E. A. P. Habets, "Room impulse response generator," in *The Netherlands, Tech. Rep. 2.4, 2006. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator*, 2006.