INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Multi-Channel Linear Prediction Based on Binaural Coherence for Speech Dereverberation

Hong Liu, Xiuling Wang, Miao Sun, Cheng Pang

Key Laboratory of Machine Perception (Ministry of Education),
Shenzhen Graduate School, Peking University, China
hongliu@pku.edu.cn, {wangxiuling, sunmiao, chengpang}@sz.pku.edu.cn

Abstract

It has been shown that the multi-channel linear prediction (MCLP) can achieve blind speech dereverberation effectively. However, it always degrades the binaural cues which are exploited for human sound localization, i.e., interaural time differences (ITD) and interaural level differences (ILD). To overcome this problem, the multiple input-single output structure of conventional MCLP is modified to a binaural input-output structure for suppressing reverberation and preserving binaural cues simultaneously. First, by employing a binaural coherence model with head shadowing effects, the variance of desired signal can be estimated the same to both ears, which can ensure no modification of ILD. Then, the variance is utilized to calculate the prediction coefficients in a maximum-likelihood (ML) sense. Finally, the desired signals can be obtained as the prediction errors in MCLP. And since the algorithm does not disturb the phase of input signal, the ITD cue is kept. Evaluations with measured binaural room impulse responses (BRIRs) show that the proposed method yields a good performance on both speech dereverberation and binaural cues preservation.

Index Terms: binaural dereverbaration, binaural cues, coherence, head shadowing, multi-channel linear prediction

1. Introduction

In an enclosed space, reverberation always decreases the speech quality and intelligibility, because of the reflections from the walls, floors, ceilings or furniture, especially in applications of hands-free devices, hearing aids, teleconferencing, sound localization, automatic speech recognition (ASR), etc [1]. Many methods have been proposed for single- and multi- microphone dereverberation, which can be broadly classified into three categories [2], inverse filtering [3], spectral enhancement [4] and probabilistic model-based approach [5,6]. The significant drawback is that most of these dereverberation methods are single/multiple input-single output techniques or independent bilateral signal processing, which means performing monaural enhancement on each side of the device without taking the spatial information into account [7]. Hence, the binaural cues, particularly interaural time differences (ITD) and interaural level differences (ILD), are severely degraded, which are exploited by human auditory system for the ability of sound localization [8,9]. It is well known that the spatial perception can be used to increase the speech intelligibility [10, 11]. The fact motivates the dereverberation with preserving binaural cues. Recently, the dereverberation methods with binaural cues preservation are suggested in [7, 12, 13], which are based on the multichannel wiener filter (MWF) [7, 12] and Kalman-EM scheme [13].

An efficient blind dereverberation method based on multi-channel linear prediction (MCLP) in the short-time Fourier transform (STFT) domain was proposed in [5]. It is assumed that the late reverberation can be predicted from the previous frames of the reverberant signal, unknown parameters of the prediction filter and time-varying Gaussian (TVG) model can be estimated by using the maximum-likelihood (ML) rule [14]. Several works have been recently presented based on MCLP for improving the dereverberation performance [2, 15–17], such as adding sparse priors to the desired signal [2, 15]. But there is a problem that these methods are focusing on the amount of dereverberation only, without considering about the effects on binaural cues and hence, the spatial information is disturbed. Therefore, we propose a method based on MCLP aiming at enabling a tradeoff between the dereverbaration performance and the preservation of the binaural cues. As the early reverberation is often beneficial to the spatial awareness [1], we only focus on the suppression of the late reverberation here.

In this paper, a binaural dereverberation method based on MCLP is proposed. First, a modified binaural input-output structure is employed to make a data-link between two ears and take the original spatial information into account. Then, a binaural coherence model with head shadowing effects is used to calculate the auto-power spectral density (APSD) of the estimated desired signal in each iteration, which is then applied to estimate its variance. In this way, the prediction coefficients on each ear can be calculated with the same value of the desired signal variance, which can keep the ILD greatly. And the dereverberation algorithm we used does not disturb the phase of the input signal with the overlap-add, hence the ITD cue can be kept. By doing so, the effects of reverberation can be reduced without affecting the impression on the sound scene for human.

2. MCLP-based dereverberation

Suppose a scenario in an enclosure where there is a single speech source captured by M microphones, $M/2$ microphones on each side of a head. Let $s(n,k)$ be the source signal in the short-time Fourier transform (STFT) domain with frequency bin index $k \in \{1,\dots,K\}$ and time index $n \in \{1,\dots,N\}$. The STFT coefficients of the observed signal at the $m\text{-}th$ microphone on the left or right ear can be represented as [14]:

$$x_{j,m}(n,k) = \sum_{L=0}^{L_h-1} h_{j,m}(L,k)s(n-L,k) + e_{j,m}(n,k), \quad (1)$$

where $j \in \{l,r\}$ represents the signals corresponding to the left or right ear, $h_{j,m}(n,k)$ is the room impulse response (RIR) of length L_h between the speech source and the $m\text{-}th$ microphone, $e_{j,m}(n,k)$ denotes the additive noise and modeling errors of the RIR convolution. As in [16], by assuming $e_m(n,k) = 0$, the

http://dx.doi.org/10.21437/Interspeech.2016-729

observed signal at a chosen reference microphone (e.g., $m = 1$) can be expressed in the multi-channel linear prediction form as:

$$x_{j,1}(n,k) = \sum_{m=1}^{M/2} \sum_{L=0}^{L_c-1} \Big(c_{l,m}(L,k) x'_{l,m}(n-\tau-L,k) + c_{r,m}(L,k) x'_{r,m}(n-\tau-L,k) \Big) + d_j(n,k),\ j \in \{l,r\}, \tag{2}$$

where $c_{j,m}$ are the coefficients of the prediction filter with the length of L_c, $x'_{j,m}$ denotes the time-aligned signal performed by means of the generalized cross-correlation with phase transform (GCC-PHAT) [18]. The first term in Eq. (2) is the late reverberation, which is predicted from the past observed signals with $c_{j,m}$. The second term $d_j(n,k) = \sum_{L=0}^{\tau-1} h_{j,1}(L,k) s(n-L,k)$ is the desired signal for each ear including the direct speech signal and early reflections with the prediction delay τ. The MCLP in Eq. (2) can be written in vector form as:

$$\boldsymbol{x}_{j,1}(k) = \boldsymbol{X}^{\tau}(k)\boldsymbol{C}_j(k) + \boldsymbol{d}_j(k),\ j \in \{l,r\}, \tag{3}$$

with

$$\begin{aligned}
\boldsymbol{X}^{\tau}(k) &= [\boldsymbol{X}^{\tau}_{l,1}(k), \ldots, \boldsymbol{X}^{\tau}_{l,M/2}(k), \boldsymbol{X}^{\tau}_{r,1}(k), \ldots, \boldsymbol{X}^{\tau}_{r,M/2}(k)], \\
\boldsymbol{C}_j(k) &= [\boldsymbol{C}^T_{l,1}(k), \ldots, \boldsymbol{C}^T_{l,M/2}(k), \boldsymbol{C}^T_{r,1}(k), \ldots, \boldsymbol{C}^T_{r,M/2}(k)]^T, \\
\boldsymbol{d}_j(k) &= [\boldsymbol{d}_j(1,k), \ldots, \boldsymbol{d}_j(N,k)]^T, \\
\boldsymbol{X}_{j,m}(k) &= [\bar{\boldsymbol{X}}_{j,m}(1,k), \ldots, \bar{\boldsymbol{X}}_{j,m}(N,k)]^T, \\
\boldsymbol{C}_{j,m}(k) &= [c_{j,m}(0,k), \ldots, c_{j,m}(L_c-1,k)]^T, \\
\bar{\boldsymbol{X}}_{j,m}(n,k) &= [x'_{j,m}(n,k), \ldots, x'_{j,m}(n-L_c+1,k)]^T,
\end{aligned}$$

where $j \in \{l,r\}$ and $(\cdot)^T$ denotes non-conjugate transposition. Here, $\boldsymbol{X}^{\tau}_{j,m}(k) \in \mathbb{R}^{N \times L_c}$ is constructed with $\boldsymbol{X}_{j,m}(k)$ delayed for τ frames. With the estimated prediction coefficients, the desired speech signal can be estimated as follows:

$$\widehat{\boldsymbol{d}}_j(k) = \boldsymbol{x}_{j,1}(k) - \boldsymbol{X}^{\tau}(k)\widehat{\boldsymbol{C}}_j(k),\ j \in \{l,r\}, \tag{4}$$

where $\widehat{(\cdot)}$ denotes an estimated variable. Therefore, the desired signal can be interpreted as the prediction error in MCLP [14].

In the speech dereverberation based on MCLP, the desired signal is always modeled as a time-varying Gaussian (TVG) model, which means that it is an independent zero-mean random variable following a circular complex Gaussian distribution in each time-frequency bin [2,14–17]. The probability density function (PDF) of the desired signal is defined as:

$$P(d_j(n,k)) = N_c(d_j(n,k); 0, \lambda_j(n,k)),\ j \in \{l,r\}, \tag{5}$$

where the variance $\lambda_j(n,k)$ is an unknown and time-varying parameter to be estimated.

Let $\boldsymbol{\lambda}_j(k) = [\lambda_j(1,k), \ldots, \lambda_j(N,k)]^T$, the likelihood function for the k-th frequency bin can be derived as:

$$L(\boldsymbol{C}_j(k), \boldsymbol{\lambda}_j(k)) = P(\boldsymbol{d}_j(k)) = \prod_{n=1}^{N} P(d_j(n,k)),\ j \in \{l,r\}. \tag{6}$$

The unknown parameters $\boldsymbol{C}_j(k)$ and $\boldsymbol{\lambda}_j(k)$ can be estimated by maximizing Eq. (6), i.e., by minimizing the following negative log likelihood function with $\boldsymbol{D}_{\boldsymbol{\lambda}_j(k)} = diag(\boldsymbol{\lambda}_j(k))$ as:

$$\min_{\boldsymbol{C}_j(k), \boldsymbol{\lambda}_j(k)} \boldsymbol{d}_j^H(k) \boldsymbol{D}^{-1}_{\boldsymbol{\lambda}_j(k)} \boldsymbol{d}_j(k) + \sum_{n=1}^{N} log \pi \lambda_j(n,k), j \in \{l,r\}. \tag{7}$$

Then, an alternating optimization procedure is used to recover the desired signal as in [15]. And the solution of the vector $\boldsymbol{\lambda}_j(k)$ is that $\widehat{\boldsymbol{\lambda}}_j(k) = |\widehat{\boldsymbol{d}}_j(k)|^2$.

3. Dereverberation with binaural coherence

In this section, the proposed binaural cues-preserved dereverberation method is introduced. To preserve ILD, a binaural coherence model with head shadowing effects is used to estimate the variance of the desired signal, which is different from the conventional MCLP. Since the algorithm is processed after time-alignment, the dereverberation performance can be robust over the entire azimuth range. And as the signal is reconstructed by using overlap-add with the phase information of the original signal, the ITD can be unaffected. The schematic diagram of our method is depicted in Fig. 1.

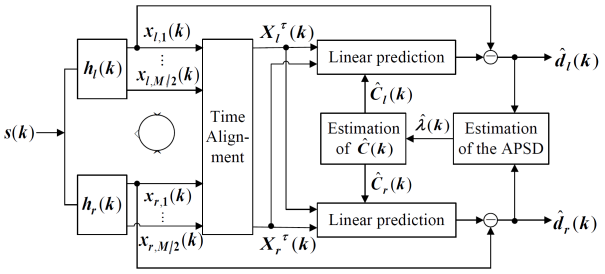

Figure 1: Block diagram of the proposed method

3.1. Estimation of $\lambda_j(n,k)$

In the conventional MCLP, the variances of the desired signal on each ear are estimated independently in a bilateral configuration as $\widehat{\boldsymbol{\lambda}}_j(k) = |\widehat{\boldsymbol{d}}_j(k)|^2$, without considering about the effects on the binaural cues. In our approach, a binaural coherence model is applied to estimate $\lambda_j(n,k)$, which ensures the same variance of the desired signal on both ears, and hence, the ILD cue can be kept. The coherence function between two signals $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ can be defined as [19]:

$$\boldsymbol{\Gamma}_{\boldsymbol{z}_1\boldsymbol{z}_2} = \frac{\boldsymbol{\Phi}_{\boldsymbol{z}_1\boldsymbol{z}_2}}{\sqrt{\boldsymbol{\Phi}_{\boldsymbol{z}_1\boldsymbol{z}_1} \cdot \boldsymbol{\Phi}_{\boldsymbol{z}_2\boldsymbol{z}_2}}}, \tag{8}$$

where $\boldsymbol{\Phi}_{\boldsymbol{z}_1\boldsymbol{z}_1}$ and $\boldsymbol{\Phi}_{\boldsymbol{z}_2\boldsymbol{z}_2}$ represent the APSDs of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ respectively, and $\boldsymbol{\Phi}_{\boldsymbol{z}_1\boldsymbol{z}_2}$ is the cross-power spectral density (CPSD) between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Here, the reverberant room is approximated by a 3D diffuse noise field [20]. And because of the object in the line-of-sight, the coherence with head shadowing effects on the input signals to both ears is approximated as [7]:

$$\widehat{\boldsymbol{\Gamma}}^{head}_{\boldsymbol{z}_1\boldsymbol{z}_2}(f) = \sum_{p=1}^{P} a_p \cdot \exp(-\frac{f-b_p}{c_p})^2, \tag{9}$$

where f is the frequency. The model order P is set to 3, and the constants a_p, b_p and c_p are set the same as in [7]. Then, the variance is estimated by using the APSD [19] of the estimated desired signal (obtained from last iteration). Here, we let $\widehat{\boldsymbol{\Phi}}(n,k) = \widehat{\Phi}_{\widehat{d}_l^{(i)}\widehat{d}_l^{(i)}}(n,k) + \widehat{\Phi}_{\widehat{d}_r^{(i)}\widehat{d}_r^{(i)}}(n,k)$, the variance of the desired signal can be defined in the following as:

$$\widehat{\lambda}^{(i)}(n,k) = \widehat{\Phi}_{\widehat{d}\widehat{d}}(n,k) = \frac{Re\{\widehat{\Phi}_{\widehat{d}_l^{(i)}\widehat{d}_r^{(i)}}(n,k)\} - \frac{1}{2} Re\{\widehat{\boldsymbol{\Gamma}}^{head}_{\widehat{\boldsymbol{d}}_l\widehat{\boldsymbol{d}}_r}(f)\} \widehat{\boldsymbol{\Phi}}(n,k)}{1 - Re\{\widehat{\boldsymbol{\Gamma}}^{head}_{\widehat{\boldsymbol{d}}_l\widehat{\boldsymbol{d}}_r}(f)\}}, \tag{10}$$

where the function $Re\{\cdot\}$ returns the real part of its argument and $(\cdot)^{(i)}$ denotes the iterating value at the i-th iteration. Since

Algorithm 1: *Outline of the proposed algorithm.* $\| x \|_\infty$ *denotes the maximum absolute value of the elements in* x

Parameters: L_c and τ in (2)
Input: $\boldsymbol{x}_{j,m}(k), \forall j, m, k$
Output: $\boldsymbol{d}_j(k), \forall j, k$
Initialization: $\widehat{\boldsymbol{d}}_j^{(0)}(k) = \boldsymbol{x}_{j,1}(k)/A_j$ with $A_j = \| \boldsymbol{x}_{j,1}(k) \|_\infty$

```
time-aligned with GCC
for k = 1, ..., K do
    calculate X^τ(k)
    for i = 0, ..., i_max do
        λ̂_j^(i)(n,k) ← max{Φ̂_{d̂d̂}(n,k), ε},
        calculate Φ̂_{d̂d̂}(n,k) as in (10)
        Ĉ_j^(i+1)(k) ← calculate as in (13)
        d̂_j^(i+1)(k) ← x_{j,1}(k) − X^τ(k) Ĉ_j^(i+1)(k)
    end for
    d̂_j(k) = A_j d̂_j(k)
end for
return d_j(k)
```

the APSD of the signal may not be negative or singular, the maximum threshold $\boldsymbol{\Gamma}_{\max}$ for the coherence function to ensure that $1 - Re\{\widehat{\boldsymbol{\Gamma}}^{head}_{\hat{\boldsymbol{a}}_l \hat{\boldsymbol{a}}_r}(f)\} > 0$ is set to 0.99. The estimation of $\Phi_{d_j d_j}(n,k)$ and $\Phi_{d_l d_r}(n,k)$ are performed by a recursive periodogram approach with smoothing factor $0 \leq \alpha \leq 1$ as:

$$\begin{aligned}\widehat{\Phi}_{\hat{d}_j^{(i)}\hat{d}_j^{(i)}}(n,k) =& \alpha \widehat{\Phi}_{\hat{d}_j^{(i)}\hat{d}_j^{(i)}}(n-1,k) + \\ & (1-\alpha)|\widehat{d}_j^{(i)}(n,k)|^2,\ j \in \{l,r\},\end{aligned} \tag{11}$$

$$\begin{aligned}\widehat{\Phi}_{\hat{d}_l^{(i)}\hat{d}_r^{(i)}}(n,k) =& \alpha \widehat{\Phi}_{\hat{d}_l^{(i)}\hat{d}_r^{(i)}}(n-1,k) + \\ & (1-\alpha)\widehat{d}_l^{(i)}(n,k) \cdot \widehat{d}_r^{(i)*}(n,k),\end{aligned} \tag{12}$$

where $*$ denotes the complex conjugate.

3.2. Binaural output

With the same variance of the desired signal for each ear in Eq. (10), the important ILD cue can be preserved in the binaural dereverberation structure. By using the ML rule and alternating optimization procedure, which has been mentioned in Section 2, the prediction coefficients and desired signals can be calculated. In the alternating optimization procedure, firstly, assuming the variances of the desired signal $\widehat{\boldsymbol{\lambda}}(k)$ are fixed to the values from the i-th iteration, the optimization problem of prediction vector $\widehat{\boldsymbol{C}}_j^{(i+1)}(k)$ can be obtained by minimizing Eq. (7). Secondly, with the estimated value of $\widehat{\boldsymbol{C}}_j^{(i+1)}(k)$ from the first step and according to Eq. (4), the desired signal $\widehat{\boldsymbol{d}}_j^{(i+1)}(k)$ can be obtained. Then, update $\widehat{\boldsymbol{\lambda}}(k)$ with Eq. (10). After the i-th iteration, the element-wise solution can be given as:

$$\begin{aligned}\widehat{\boldsymbol{C}}_j^{(i+1)}(k) =& \Big((\boldsymbol{X}^\tau(k))^H \boldsymbol{D}^{-1}_{\widehat{\boldsymbol{\lambda}}^{(i)}(k)} \boldsymbol{X}^\tau(k) \Big)^{-1} \\ & (\boldsymbol{X}^\tau(k))^H \boldsymbol{D}^{-1}_{\widehat{\boldsymbol{\lambda}}^{(i)}(k)} \boldsymbol{x}_{j,1}(k),\ j \in \{l,r\},\end{aligned} \tag{13}$$

$$\widehat{\boldsymbol{d}}_j^{(i+1)}(k) = \boldsymbol{x}_{j,1}(k) - \boldsymbol{X}^\tau(k) \widehat{\boldsymbol{C}}_j^{(i+1)}(k),\ j \in \{l,r\}, \tag{14}$$

with the complex conjugate transposition $(\cdot)^H$. The iterating procedure is repeated at a maximum number or until covergence with the initialization as $\widehat{d}_j^{(0)}(n,k) = x_{j,1}(n,k)$.

The dereverberation scheme based on MCLP with binaural cues preservation is summarized in Algorithm 1. Note that for each frequency bin k, the matrix $\boldsymbol{X}_j^\tau(k), j \in \{l,r\}$ is normalized with the maximum magnitude of the STFT coefficients of the reference microphone signal $\boldsymbol{x}_{j,1}$. Also, our method can be applied to MCLP with sparse priors in [2], i.e., first, modifying the MCLP model to the binaural input-output structure. Then, the variance of the desired signal is estimated as $\left(\widehat{\lambda}^{(i)}(n,k)\right)^{2-\delta}$ with the shape parameter δ ($0 \leq \delta \leq 2$) of the complex generalized Gaussian (CGG) prior, and $\widehat{\lambda}^{(i)}(n,k)$ is calculated by using our approach in Eq. (10). And this method will be compared with other three approaches in our experiments in Section 4.

4. Experiments and analysis

To test the performance of our approach, experiments are carried out with four different binaural room impulse responses (BRIRs) from the Aachen Impulse Response (AIR) database [21]. The selected BRIRs are measured with a dummy head (one microphone on each ear, i.e., $M = 2$) in different reverberant environments at a microphone distance 0.17 m. Reverberation time RT_{60}, louder speaker-microphone distance $Dist.$, azimuth angle θ between head and loudspeaker of the BRIRs are shown in Table 1. The evaluation is performed by using 80 utterances of 10 male and 10 female (4 utterances for each speaker) from the TIMIT database [22], which are convolved with BRIRs to generate the reverberant speech signals. The average length of an utterance is approximately 3.8 *sec*.

Table 1: Properties of the different rooms

Room	$Dist.$	RT_{60}	θ
Office Room	$1m$	$0.45s$	90° (frontal)
Lecture Room	$7.1m$	$0.85s$	90°
Stairway Hall	$2m$	$0.83s$	$0^\circ, 15^\circ, \ldots, 90^\circ$
Aula Carolina	$5m$	$5.16s$	90°

Four different methods are compared here. The conventional bilateral MCLP (labeled as MCLP) [5], in which the prediction coefficients and the variances of the desired signals on two ears are calculated independently without any data-link, the variances $\widehat{\boldsymbol{\lambda}}_j(k)$ are estimated as $\widehat{\boldsymbol{\lambda}}_j(k) = |\widehat{\boldsymbol{d}}_j(k)|^2$, and the $\boldsymbol{X}^\tau(k)$ in (3) (4) (13) (14) is represented by $\boldsymbol{X}_j^\tau(k), j \in \{l,r\}$. The conventional MCLP with CGG sparse prior (labeled as MCLP-SP) [2]. Our method, i.e., calculating the variance of the desired signal using Eq. (10) in a binaural structure (labeled as MCLP-COH). And our method with CGG, which has been mentioned in Section 3.2 (labeled as MCLP-SPCOH). In the experiments, the sampling frequency is 16kHz, the STFT is calculated using a Hann window with the frame length of 256 and 50% overlap. The length of the prediction filter is set to $L_c = 10$, the prediction delay $\tau = 2$, and $\varepsilon = 10^{-8}$. And the shape parameter in MCLP-SP and MCLP-SPCOH is set to $\delta = 0.5$. Note that the maximum number of the iteration i_{max} is set to 1 for MCLP, MCLP-COH and 10 for MCLP-SP, MCLP-SPCOH. Because in MCLP or MCLP-COH, the variance is updated with no spectral priors, the results indicate that the quality often degrades in the following iterations, the same as is mentioned in [15].

The performance is evaluated in terms of a non-intrusive measurement speech to reverberant modulation energy ratio (SRMR) [23], an intrusive measurement cepstral distance (CD) [24], ITD and ILD. In the following, the results are the improve-

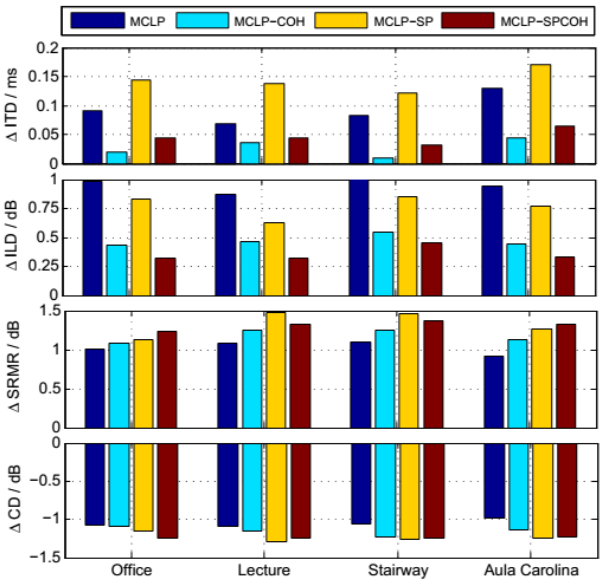

Figure 2: Results for different reverberant environments. ΔSRMR and ΔCD are the averaged results between the left and right signal. Note that for MCLP and MCLP-SP, ITD and ILD are calculated in a bilateral structure.

ments of the measurements, i.e., $\Delta I = I_{derev} - I_{rev}$, $\Delta Q = |Q_{derev} - Q_{rev}|$, with $I \in \{\text{SRMR, CD}\}$, $Q \in \{\text{ITD, ILD}\}$, $(\cdot)_{derev}$ and $(\cdot)_{rev}$ are the evaluations on dereverberated signals and original speech signals, respectively. Here, ITD is evaluated by GCC-PHAT [18], and ILD is simply calculated by the energy ratio as $10\log_{10}(\frac{E_l}{E_r})$, where E_l and E_r are the energy of the left and right signal respectively. The results (averaged over all utterances) are shown in Fig. 2.

It can be observed in Fig. 2 that for all experiments, the reverberation is suppressed in terms of ΔSRMR. Comparing ΔITD and ΔILD of MCLP with MCLP-COH, MCLP-SP with MCLP-SPCOH, the binaural cues can be preserved distinctly by using our method. Since the variance of the desired signal is estimated to be the same on each ear in our binaural structure, the ILD cue is not modified. There is an exception that it seems that the ITD is preserved slightly better with MCLP-COH than MCLP-SPCOH. But in theory, they should be the same as the phases of original signals are kept. It is due to the signal distortion and the remaining reverberation in the signals, which will be improved in our future work.

Comparing the ΔSRMR in four different reverberant environments in Fig. 2, it can be seen that the SRMR values are improved more in the lecture room and the stairway hall than in the office room. It is because the selected BRIRs of the lecture room and the stairway hall are with longer RT_{60} than the office room, the more reverberant the room, the more amount of dereverberation can be obtained. The results also show a good enhancement performance on the Aula-Carolina BRIR, which is an extreme case here with $RT_{60} = 5.16s$. It indicates the robustness of the proposed method in highly reverberant environment. However, by preserving the binaural cues, the dereverberation performance degrades slightly as expected in some cases in terms of ΔSRMR, such as in the lecture room and the stairway hall. It is caused by the accumulated errors of the approximation of the binaural coherence model and the estimation of the APSDs for calculating the variance of the desired signal. Besides, the CDs can be reduced effectively. And the ΔCDs of MCLP and MCLP-COH (or MCLP-SP and MCLP-SPCOH) are similar, which means our method does not cause more distortions compared to the conventional method. Note that the dereverberation and binaural cues preservation performance will be better with $M \geq 2$, i.e., more than one microphone on each ear. Because with the increase of the microphone number, more original speech and spatial information can be captured.

Table 2: Comparison of ITD for different θ (azimuth angle) in the stairway hall environment

θ	90°	60°	45°	30°	0°
MCLP	0.0461	0.2763	0.3908	0.4739	0.5666
MCLP-COH	**0.0492**	**0.3349**	**0.4414**	**0.5562**	**0.6641**
MCLP-SP	-0.0258	0.2423	0.3772	0.4494	0.5294
MCLP-SPCOH	**0.0375**	**0.3215**	**0.4232**	**0.5045**	**0.6646**
original	0.0625	0.375	0.50	0.5625	0.675

Table 3: Comparison of ILD for different θ (azimuth angle) in the stairway hall environment

θ	90°	60°	45°	30°	0°
MCLP	-0.87	3.91	6.31	7.73	8.69
MCLP-COH	**-0.80**	**3.53**	**5.83**	**7.14**	**8.01**
MCLP-SP	-0.57	3.87	6.24	7.41	8.29
MCLP-SPCOH	**-0.70**	**3.56**	**5.76**	**6.98**	**7.69**
original	-0.71	3.3	5.29	6.29	6.87

The results of ITD and ILD for different azimuths in the stairway hall environment can be found in Table 2 and Table 3 respectively. The results of other azimuths are not presented here for the space limitation. The more closer the value is to the original signal, the better performance it can be achieved in preserving the binaural cues. Comparing the ITD or ILD of MCLP with MCLP-COH, MCLP-SP with MCLP-SPCOH, the binaural cues can be kept efficiently with our method. Also, it can be seen that the lowest influence of the algorithms is at $\theta = 90°$. Since the ITD or ILD is small in the frontal direction.

5. Conclusions

This work introduced a binaural dereverberation method based on MCLP. By employing a binaural coherence model which has taken the head shadowing effects into account, the parameters of MCLP can be estimated without disturbing the ILD cue. And by using the algorithm without affecting the phase of the original signal in a binaural input-output structure, the ITD cue can be kept. Experimental results have shown that the proposed method can preserve the binaural cues while has little effect on the dereverberation performance. As the experiments are carried out focusing on the reverberation only, without adding noise, it can be extended to speech enhancement in both reverberant and noisy environment in our future work.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046, 60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011), Natural Science Foundation of Guangdong Province(No. 2015A030311034), Science and Technology Innovation Commission of Shenzhen Municipality (No. JCYJ20130331144631730, No. JCYJ20130331144716089).

7. References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2010.

[2] A. Jukic, N. Mohammadiha, T. Van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. on Audio Speech & Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics Speech & Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[4] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770 – 773, 2009.

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

[6] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and kalman smoothing," *Signal Processing Conference (EUSIPCO)*, pp. 1 – 5, 2013.

[7] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. on Audio Speech & Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[8] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Trans. on Signal Processing (TSP)*, vol. 63, no. 18, pp. 4771–4783, 2015.

[9] H. Liu and X. Li, "Time delay estimation for speech signal based on foc-spectrum," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1732–1735, 2012.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, rev. ed. Mit Press, 1996.

[11] C. Pang, J. Zhang, and H. Liu, "Direction of arrival estimation based on reverberantion weighting and noise error estimation," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3436–3440, 2015.

[12] S. Braun, M. Torcoli, D. Marquardt, and E. A. P. Habets, "Multichannel dereverberation for hearing aids with interaural coherence preservation," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 124–128, 2014.

[13] G. Schning and O. Glemser, "An online dereverberation algorithm for hearing aids with binaural cues preservation," *Chemische Berichte*, vol. 110, no. 9, pp. 3231–3234, 2015.

[14] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. on Audio Speech & Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[15] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on itakura-saito distance minimization," *IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 245–248, 2012.

[16] A. Jukic, N. Mohammadiha, T. Van Waterschoot, and T. Gerkmann, "Multichannel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," *IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 96–100, 2015.

[17] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. on Audio Speech & Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[18] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans.on Acoustics Speech & Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] I. A. Mccowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. on Speech & Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[20] H. Kuttruff, *Room Acoustics*. London: Taylor & Francis, 2000.

[21] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Proceeding of Digital Signal Processing (DSP)*, pp. 1 – 5, 2009.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[23] T. H. Falk and W. Y. Chan, "A non-intrusive quality measure of dereverberated speech," *IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[24] S. Furui, *Digital Speech Processing: Synthesis, and Recognition, Second Edition, Revised and Expanded*. New York: Marcel Dekker, 2001.