*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Direct-Path Signal Cross-Correlation Estimation for Sound Source Localization in Reverberation

*Wei Xue,Ying Tong,Guohong Ding,Chao Zhang,Tao Ma,Xiaodong He,Bowen Zhou*

JD AI Research

{xuewei27,tongying,dingguohong,chao.zhang,tao.ma,xiaodong.he,bowen.zhou}@jd.com

## Abstract

Sound source localization (SSL) is challenging in presence of reverberation since the cross-correlation between the direct-path signals in different microphones, which indicates the spatial information of the sound source, is interfered by the reverberation signal components. A novel algorithm is proposed in this paper to estimate the cross-correlation of the *direct-path* speech signals, such that the robustness of SSL to reverberation can be improved. The proposed method follows a similar scheme to the multichannel linear prediction (MCLP), which is commonly used for speech dereverberation, while avoids the explicit estimation of the direct-path signal of each channel. This is achieved by revealing the relationship between the direct-path signal cross-correlation (DPCC) and the MCLP coefficient vector, and finally deriving the DPCC by using only the multichannel reverberant signals. It is also shown that the pre-whitening operation, which is widely used for SSL, can be inherently integrated into the estimated DPCC. An adaptive method is further derived to facilitate online frame-level SSL. The proposed method can be easily applied to conventional cross-correlation based SSL methods by using the DPCC rather than the full cross-correlation. Experiments conducted in various reverberant conditions demonstrate the effectiveness of the proposed method.

**Index Terms**: sound source localization, reverberation, microphone arrays

## 1. Introduction

Sound source localization (SSL) is of importance in modern speech processing related systems such as teleconferencing, robotics, and hands-free devices *etc*. For instance, given the spatial information of the sound source, the head of a robot can steer towards the active speaker, and speech processing algorithms ( *e.g.* beamforming [1, 2] and speech dereverberation [3, 4]) can be designed to extract the target speech signal. A microphone array is commonly deployed to perform a spatial sampling of the sound field, and SSL can be achieved by analysing the cross-correlations between the microphone array signals.

Conventional methods for SSL can be categorized into high-resolution spectral [5, 6], steered beamformer response power (SRP) [7], and time delay estimation (TDE) [8–13] based methods, among which the generalized cross-correlation (GCC) [8, 14] based TDE methods have become a mainstream. However, these methods generally adopt a free-field signal model [15] that only considers the direct-path propagation from the source to microphone. Therefore, their performances often degrade dramatically in highly-reverberant environments, in which the reflections from directions other than the direct path also present, and contaminate the direct-path signal cross-correlation (DPCC) that is required for SSL of the target source.

To overcome this problem, several methods [16–19] are proposed to exploit the time-frequency (TF) units with less reverberation for SSL, and these TF units are selected either by the coherence test [16] or the direct-path dominance test [17, 18]. Methods based on a more realistic signal model are also developed. In [20, 21], the room impulse responses (RIRs) from the source to different microphones, which characterize the full sound propagation process in the reverberant environment, are estimated by blind system identification (BSI). Time delays among microphone signals are then computed from the direct paths in the estimated RIRs to give an SSL result. An under-modelled BSI based method is proposed in [22], which estimates the early RIRs instead of the whole impulse responses. In addition, relative transfer function (RTF) based methods [23–26] are developed by generalising the anechoic steering vector into the RTF, and the RTF can be estimated by solving a set of linear equations to obtain the least square solutions.

Since the DPCC is essential for reverberation-robust SSL, in this paper, we aim to estimate the DPCC directly from the multichannel reverberant observations. An advantage of the proposed method is that it can be easily applied to all cross-correlation based SSL methods by using the DPCC rather than the cross-correlation of the overall signal. The proposed method is developed by exploiting multichannel linear prediction (MCLP) [27], which is used for speech dereverberation, while avoiding the estimation of the direct-path signal of each channel from explicit speech dereverberation. The relationship between the DPCC and the MCLP vector is derived. An optimization problem to compute the MCLP vector for SSL is formulated, and inherently integrates the pre-whitening operation which is widely used for SSL to remove the effect of speech autocorrelation. The DPCC is finally computed only based on the multichannel reverberant signals. An adaptive method is further derived for online SSL in each frame. Experiments conducted in real reverberant environments with different source-array distances demonstrate the effectiveness of the proposed method.

The rest of the paper is organised as follows. Section 2 reviews the basics on signal model. The proposed method for DPCC estimation and SSL is presented in Section 3. The experimental setup and results are given in Sections 4, which is followed by conclusions.

## 2. Signal Model

Considering a reverberant environment which consists of a speech source and an $M$-element microphone array, in the time domain, the signal captured by the $m$-th microphone at time $t$, $y_m(t)$, is given by

$$y_m(t) = s(t) \star h_m + v_m(t), \tag{1}$$

where $s(t)$ is the time-domain speech source signal, $h_m = [h_{m,0}, h_{m,1}, ..., h_{m,L-1}]^T$ is the $L$-tap time-invariant RIR from the source to the $m$-th microphone, and $\star$ denotes the convolution operation. The additive environmental noise is represented

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1488

by $v_m(t)$, and is assumed to be uncorrelated with the source signal and the noise signals from other microphones.

The time-domain signal model can be transformed into the short-time Fourier transform (STFT) domain as long as the frame length is larger than the RIR length $L$. In the reverberant environment when the frame length might be shorter than $L$, the convolutive transfer function approximation [28] can be adopted, by which the time-domain RIR is divided into several overlapping segments. The STFT-domain reverberant signal is expressed as the narrowband convolution between the Fourier coefficients of the segmented RIR and the source signal. In the vector-form, the $m$-th microphone signal $Y_m(k,f)$ in the TF unit $(k,f)$ can be expressed as

$$Y_m(t,f) = \mathbf{h}_m^H(f)\mathbf{s}(t,f) + V_m(t,f), \tag{2}$$

where $\mathbf{h}_m(f) = [H_{m,0}(f), H_{m,1}(f), ..., H_{m,P-1}(f)]^T$ is a $P \times 1$ vector with $H_{m,p}(f)$ as the Fourier coefficient of the $p$-th segment of the RIR $h_m$, and $(\cdot)^H$ denotes Hermitian transpose. The $\mathbf{s}(t,f) = [S(t,f), S(t-1,f), ..., S(t-P+1,f)]^T$ is the source signal vector, and $V_m(t,f)$ is the STFT of the noise signal $v_m(t)$.

If the frame length of the STFT is short, the direct-path propagation is mainly characterized by $H_{m,0}(f)$, which represents the spatial information of the source. The direct-path signal of the $m$-th microphone, $x_{m,\mathrm{d}}(t,f)$, is written as

$$x_{m,\mathrm{d}}(t,f) = H_{m,0}^*(f)S(t,f), \tag{3}$$

in which the effect of reverberation is eliminated, and $(\cdot)^*$ denotes conjugation operation. The goal of this paper is to estimate the cross-correlation between the direct-path signals $x_{m,\mathrm{d}}(t,f)$ for $m = 1, 2, ..., M$ such that the spatial information of the source can be extracted from the reverberant observations.

# 3. Proposed Method

In this section, a new algorithm is developed to estimate the DPCC by following a scheme similar to speech dereverberation, and avoids the time-consuming explicit estimation of the direct-path signal. The relationship between the DPCC and the MCLP coefficient vector, which is used for speech dereverberation, is derived, which enables the cross-correlation being finally estimated solely based on the multichannel reverberant signal observations. The SSL is achieved by using the direct-path signal cross-correlations for the conventional steered-response power phase transform (SRP-PHAT) method [29]. This section will also show the choice of the optimization criterion to obtain the cross-correlation of the pre-whitened direct-path signals using the proposed method. Furthermore, an adaptive algorithm is derived for online frame-level updating.

## 3.1. MCLP Coefficient Vector

For the ease of discussion, the additive noise is temporarily ignored in this subsection, and the noise issue will be discussed later in Section 3.2.3. The multichannel speech signal have temporal and spatial correlations [27, 30–32], therefore, by using MCLP [27], the reverberation signal component of the $m$-th microphone, $\hat{x}_{m,\mathrm{r}}(t,f)$, can be estimated as

$$\hat{x}_{m,\mathrm{r}}(t,f) = \mathbf{g}_m^H(t,f)\mathbf{y}(t-1,f), \tag{4}$$

where $\mathbf{g}_m(t,f)$ is the MCLP coefficient vector for the $m$-th microphone. $\mathbf{y}(t-1,f)$ is a $MP \times 1$ vector consisting of the multichannel reverberant observations in the past $P$ frames, and is given by

$$\mathbf{y}(t-1,f) = [\mathbf{y}_1(t-1,f)^T, ..., \mathbf{y}_M(t-1,f)^T]^T,$$

$\mathbf{y}_m(t-1,f)$ is defined for the microphone $m$ as $\mathbf{y}_m(t-1,f) = [Y_m(t-1,f), Y_m(t-2,f), ..., Y_m(t-P,f)]^T$.

According to (2), $\mathbf{y}(t-1,f)$ can be rewritten as

$$\mathbf{y}(t-1,f) = \mathbf{H}^H(f)\tilde{\mathbf{s}}(t-1,f). \tag{5}$$

Here $\mathbf{H}(f) = [\mathbf{H}_1(f), \mathbf{H}_2(f), ..., \mathbf{H}_M(f)]$ is a $(2P-1)\times MP$ matrix, $\mathbf{H}_m(f)$ is of $(2P-1)\times P$ and is defined as

$$\mathbf{H}_m(f) = \begin{bmatrix} H_{m,0} & 0 & \dots & 0 \\ H_{m,1} & H_{m,0} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ H_{m,P-1} & H_{m,P-2} & \dots & H_{m,0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & H_{m,P-1} \end{bmatrix}, \tag{6}$$

and $\tilde{\mathbf{s}}(t-1,f) = [S(t-1,f), S(t-2,f), ..., S(t-2P+1,f)]^T$ is a $(2P-1)\times 1$ vector.

## 3.2. DPCC Estimation

### *3.2.1. Relationship with the MCLP Coefficient Vector*

The direct-path signal of the $m$-th microphone is estimated by subtracting the reverberation signal component from the observed signal:

$$\begin{aligned}\hat{x}_{m,\mathrm{d}}(t,f) &= Y_m(t,f) - \hat{x}_{m,\mathrm{r}}(t,f) \\ &= Y_m(t,f) - \mathbf{g}_m^H(t,f)\mathbf{y}(t-1,f).\end{aligned} \tag{7}$$

Given (7), the DPCC between the $i$-th and $j$-th microphones is calculated as

$$\begin{aligned}r_{i,j}^{\mathrm{d}}(t,f) &= \mathbb{E}\{\hat{x}_{i,\mathrm{d}}(t,f)\hat{x}_{j,\mathrm{d}}^*(t,f)\} \\ &= \mathbb{E}\{[Y_i(t,f) - \mathbf{g}_i^H(t,f)\mathbf{y}(t-1,f)]\times \\ &\quad [Y_j^*(t,f) - \mathbf{y}^H(t-1,f)\mathbf{g}_j(t,f)]\} \\ &= r_{i,j}(t,f) - \mathbf{g}_i^H(t,f)\mathbf{r}_j(t,f) - \mathbf{r}_i^H(t,f)\mathbf{g}_j(t,f) \\ &\quad + \mathbf{g}_i^H(t,f)\mathbf{R}(t-1,f)\mathbf{g}_j(t,f),\end{aligned} \tag{8}$$

thus, it is expressed as a function of the MCLP coefficient vector $\mathbf{g}_m(t,f)$ for $m = i, j$, as well as and a few cross-correlation related terms $r_{i,j}(t,f)$, $\mathbf{r}_i(t,f)$, $\mathbf{r}_j(t,f)$, and $\mathbf{R}(t,f)$.

The above-mentioned cross-correlation related terms are defined as

$$r_{i,j}(t,f) = \mathbb{E}\{Y_i(t,f)Y_j^*(t,f)\}, \tag{9}$$

$$\mathbf{r}_m(t,f) = \mathbb{E}\{Y_m^*(t,f)\mathbf{y}(t-1,f)\}, \tag{10}$$

$$\mathbf{R}(t-1,f) = \mathbb{E}\{\mathbf{y}(t-1,f)\mathbf{y}^H(t-1,f)\}, \tag{11}$$

and represent the inter-channel cross-correlation, the correlation vector between the $m$-th microphone and the multichannel signal vector $\mathbf{y}(t-1,f)$, and the signal covariance matrix, respectively. $\mathbb{E}[\cdot]$ stands for the mathematical expectation. We note that all of these cross-correlation related terms can be directly computed from the multichannel reverberant signals.

#### 3.2.2. MCLP Coefficient Vector for SSL

Next, the MCLP coefficient vector $\mathbf{g}_m(t,f)$ for $m = i, j$ in (8), is computed from the multichannel observations. In speech dereverberation, several criteria (*e.g.* weighted prediction error (WPE) and generalized WPE (GWPE) [33]) are available to formulate an optimization problem for $\mathbf{g}_m(t,f)$. Since it is well-known that the autocorrelation of the non-white speech signal imposes a negative effect on the cross-correlation based SSL, a pre-whitening operation is usually applied to the speech signals. In the rest of this subsection, an optimization criterion specialized for the SSL task is designed.

By modelling the speech source as a $(2P-1)$-order auto-regression (AR) process [31] stimulated by a white Gaussian signal and ignoring the noise, (2) becomes

$$Y_m(t,f) = \mathbf{h}_m^H(f)[\mathbf{C}^H(f)\tilde{\mathbf{s}}(t-1,f) + \mathbf{e}(t,f)], \quad (12)$$

where

$$\mathbf{C}(f) = \begin{bmatrix} c_1 & 1 & 0 & \dots & 0 \\ c_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_{2P-2} & 0 & 0 & \dots & 1 \\ c_{2P-1} & 0 & 0 & \dots & 0 \end{bmatrix} \quad (13)$$

is the $(2P-1) \times P$ dimensional AR coefficient matrix, and $\mathbf{e}(t,f) = [e(t,f), 0, ..., 0]^T$ is a $P \times 1$ vector with $e(t,f)$ following the complex white Gaussian distribution.

Based on (5) and (12), (7) can be rewritten as

$$\begin{aligned}\hat{x}_{m,\mathrm{d}}(t,f) = &[\mathbf{C}(f)\mathbf{h}_m(f) - \mathbf{H}(f)\mathbf{g}_m(t,f)]^H \tilde{\mathbf{s}}(t-1,f) \\ &+ \mathbf{h}_m^H(f)\mathbf{e}(t,f). \end{aligned} \quad (14)$$

Note that $\mathbf{h}_m^H(f)\mathbf{e}(t,f) = H_{m,0}^*(f)e(t,f)$ is a white Gaussian signal multiplied by a scalar, $H_{m,0}^*(f)$, which represents the direct-path propagation of the RIR.

To remove the effect of speech autocorrelation, the first term of the right-hand side of (14) need to be eliminated. Using the fact that $\mathbf{h}_m^H(f)\mathbf{e}(t,f)$ is complex white Gaussian, it leads to the minimum mean square error optimization problem on (7):

$$\hat{\mathbf{g}}_m(t,f) = \arg\min_{\mathbf{g}_m(t,f)} \mathbb{E}[\hat{x}_{m,\mathrm{d}}(t,f)\hat{x}_{m,\mathrm{d}}^*(t,f)]. \quad (15)$$

The optimal solution of (15) is given by

$$\hat{\mathbf{g}}_m(t,f) = [\mathbf{R}(t-1,f)]^{-1}\mathbf{r}_m(t,f). \quad (16)$$

Substituting (16) into (8) and after some simplifications, the DPCC between the microphone $i$ and $j$ can be finally estimated only based on the multichannel reverberant observations:

$$\begin{aligned} r_{i,j}^{\mathrm{d}}(t,f) = &r_{i,j}(t,f) \\ &- \mathbf{r}_i^H(t,f)[\mathbf{R}(t-1,f)]^{-1}\mathbf{r}_j(t,f). \end{aligned} \quad (17)$$

#### 3.2.3. An Adaptive Algorithm

The DPCC in (17) can be computed in the batch mode, in which the whole utterance is used to estimate the cross-correlation related terms defined from (9) to (11). However, for the online case, the matrix inversion in (17) for each TF bin is overly time-consuming. An adaptive algorithm is further derived in this subsection to facilitate the online frame-level estimation of the DPCC.

**Algorithm 1:** An adaptive DPCC estimation algorithm

$T$: the number of frames, $F$: the number of frequency bins.
**for** $t = P+1$ *to* $T$ **do**
  **for** $f = 1$ *to* $F$ **do**
    1. Recursive smoothing update of $r_{i,j}(t,f)$ and $\mathbf{r}_m(t,f)$ using the smoothing factor in (20);
    2. Update $\mathbf{R}^{-1}(t-1,f)$ using (19);
    3. Update the DPCC using (17);
  **end**
**end**

For online application, the cross-correlation related terms are usually updated by recursive smoothing. By taking the update of $\mathbf{R}(t,f)$ in (11) for example, the update scheme is expressed as

$$\mathbf{R}(t,f) = \alpha\mathbf{R}(t-1,f) + (1-\alpha)\mathbf{y}(t,f)\mathbf{y}^H(t,f), \quad (18)$$

where $\alpha \in (0,1]$ is a smoothing factor.

Consider the update at frame $t+1$. Given the observations of a new frame, the $r_{i,j}(t,f)$ and $\mathbf{r}_m(t,f)$ in (9) and (10) can be updated similarly as (18), and can be substituted into (17). Further, to avoid the matrix inversion in each TF bin, the Sherman-Morrison-Woodbury formula [34, 35] is adopted to compute $\mathbf{R}^{-1}(t,f)$ for the frame $t+1$. From (18), it yields

$$\begin{aligned} &\mathbf{R}^{-1}(t) \\ &= \alpha^{-1}\mathbf{R}^{-1}(t-1) - \frac{\frac{1-\alpha}{\alpha^2}\mathbf{R}^{-1}(t-1)\mathbf{y}(t)\mathbf{y}^H(t)\mathbf{R}^{-1}(t-1)}{1+\frac{1-\alpha}{\alpha}\mathbf{y}^H(t)\mathbf{R}^{-1}(t-1)\mathbf{y}(t)} \\ &= \alpha^{-1}\mathbf{R}^{-1}(t-1) - \frac{(1-\alpha)\mathbf{u}(t)\mathbf{u}^H(t)}{\alpha^2+\alpha(1-\alpha)\mathbf{y}^H(t)\mathbf{u}(t)}, \end{aligned} \quad (19)$$

where the frequency index "$f$" is omitted for simplicity, and $\mathbf{u}(t) = \mathbf{R}^{-1}(t-1)\mathbf{y}(t)$ is an $MP \times 1$ vector.

The adaptive algorithm for estimating the DPCC is summarized in Algorithm 1. In practice, to improve the robustness to noise, similar to [36], the smoothing factor for the recursive updating is adjusted by the speech presence probability (SPP) as

$$\alpha(t,f) = \alpha_d + (1-\alpha_d)[1-p(t,f)], \quad (20)$$

where $\alpha_d \in (0,1)$ is a constant, $p(t,f)$ is the SPP which can be estimated by [37]. In this way, a fast update rate can be obtained for large SPP, and the cross-correlation terms do not update in noise-dominated TF bins when the SPP $p(t,f) = 0$.

### 3.3. SSL

The estimated DPCC can be directly used by the conventional cross-correlation based methods. In this paper, the SRP-PHAT method [29] is used to obtain a final SSL estimation. Details of using the DPCC for SRP-PHAT is omitted here.

## 4. Evaluation

In this section, the proposed method is compared with the conventional SRP-PHAT algorithm, and the direct-path RTF (DP-RTF) method [26] in different real reverberant conditions.

### 4.1. Data Description and Experimental Setup

Three six-element uniform circular microphone arrays with identical frequency responses are used to simultaneously collect the multichannel reverberant speech signals in a real room. The radius of the circular array is $4.1\,\mathrm{cm}$, and the dimension of the room is $3\,\mathrm{m} \times 4.2\,\mathrm{m} \times 3\,\mathrm{m}$. The target speaker is seated at $(1.5, 0.5, 1.4)\,\mathrm{m}$, and the microphone arrays are positioned on the horizontal plane with centre at $(1.5, 1.5, 1.4)\,\mathrm{m}$, $(1.5, 2.5, 1.4)\,\mathrm{m}$ and $(1.5, 3.5, 1.4)\,\mathrm{m}$, which correspond to the source-array distance as $1\,\mathrm{m}$, $2\,\mathrm{m}$ and $3\,\mathrm{m}$, respectively. The reverberation time of the room is approximately $400\,\mathrm{ms}$. It should be noted that although the reverberation time is the same for the three distances, the direct-to-reverberant ratios (DRRs) [3] are different, which influence the difficulty of SSL.

Twenty volunteers were asked to read 50 randomly selected sentences, and each sentence is a combination of a wakeup word and a request phrase, with the duration ranging from $3\,\mathrm{s}$ to $12\,\mathrm{s}$. Therefore, in total 1000 sentences were collected for each source-array distance. The microphone arrays captured the signals at the $16\,\mathrm{kHz}$ sampling rate and $32\,\mathrm{bit}$ resolution. The ambient environmental noise was also present, and the signal-to-noise ratios (SNRs) were approximately 25 dB, 22 dB and 20.5 dB for the $1\,\mathrm{m}$, $2\,\mathrm{m}$ and $3\,\mathrm{m}$ distances.

For all algorithms, the analysis window for STFT is 512 samples Hamming window with 50% overlap. The other parameters of the proposed method are chosen as $P = 20$, and $\alpha_d = 0.92$. The SSL is conducted in each frame on the $5^\circ$ spatial resolution, and only the azimuth is estimated. The frame-level accuracy, utterance-level accuracy and frame-level mean absolute error are used to evaluate the SSL performance. For each frame, the estimation is supposed to be correct if the absolute error is less than $10^\circ$. The utterance-level estimation is defined as successful if the frame-level accuracy of the utterance is higher than 0.5, since the true azimuth could be determined by majority voting with $10^\circ$ error tolerance.

### 4.2. Results

The frame-level accuracy and utterance-level accuracy for different source-array distances are depicted in Fig. 1. It is observed that the proposed method yields the best performance for all cases tested. In Fig. 1 (a), it is shown that when the source-array distance is $1\,\mathrm{m}$, the proposed method can achieve a frame-level accuracy of 92.24%, which leads to a 100.00% utterance-level accuracy for the tested utterances as shown in Fig. 1 (b). When increasing the source-array distance, all methods suffer from performance degradation as a result of the decreased DRR and SNR. However, the proposed method can still have a 81.35% frame-level accuracy and 94.50% utterance-level accuracy when the source-array distance is $3\,\mathrm{m}$, which outperforms the conventional SRP-PHAT method that also exploits signal pre-whitening, and the DP-RTF method that generalises the anechoic steering vector into RTF for SSL.

Fig. 2 displays the error bars of the frame-level absolute errors computed from 66,802 frames for each source-array distance. With the increase of the source-array distance, generally the comparison methods have a larger variation in the SSL estimations over frames, however, the proposed method can always attain an accurate and stable estimation in different conditions, which indicates a better robustness against reverberation.

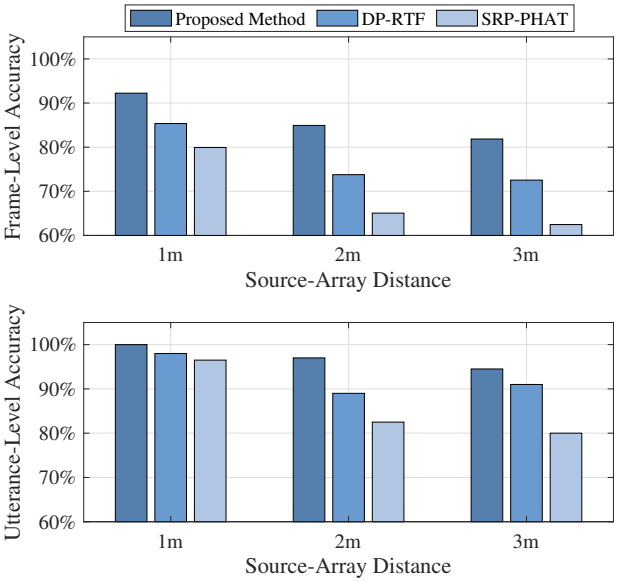


Figure 1: *Frame-level (a) and utterance-level (b) estimation accuracy of different algorithms for different source-array distances. The error tolerance is* $10^\circ$.

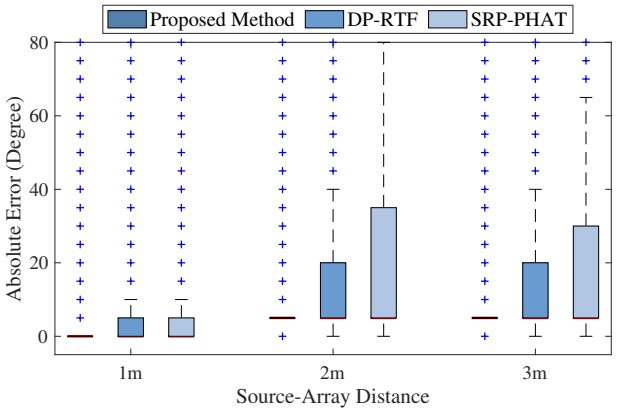


Figure 2: *Error bars of the absolute SSL errors as a function of source-array distance. Whiskers are 1.5 times the interquartile range.*

## 5. Conclusions

A novel algorithm for the DPCC estimation is proposed to improve the robustness of SSL in reverberation. The proposed algorithm is based on the MCLP, and requires no explicit estimation of the direct-path signals from different microphones. The relationship between the direct-path cross-correlation and the MCLP vector is revealed, and an optimal criterion to estimate the DPCC is formulated. In this framework, the pre-whitening operation is inherently integrated. Moreover, an adaptive algorithm is further derived for frame-level online SSL. The experiments on multichannel recordings in real reverberant environments demonstrate the superiority of the proposed method in terms of both accuracy and robustness.

# 6. References


[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[2] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 158–170, Jan. 2010.

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

[4] H. A. Javed, A. H. Moore, and P. A. Naylor, "Spherical harmonic rake receivers for dereverberation," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[5] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[6] J. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 18–21.

[7] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[9] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 157–160, Mar. 2007.

[10] W. Xue and W. Liu, "Direction of arrival estimation based on subband weighting for noisy conditions," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 142–145.

[11] W. Xue, W. Liu, and S. Liang, "Noise robust direction of arrival estimation for speech source with weighted bispectrum spatial correlation matrix," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 837–851, 2015.

[12] W. Xue, S. Liang, and W. Liu, "Interference robust DOA estimation of human speech by exploiting historical information and temporal correlation." in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 2895–2899.

[13] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.

[14] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, May 2001.

[15] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[16] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Am.*, vol. 123, pp. 2136–2147, 2008.

[17] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[18] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[19] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, nov 2004.

[20] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*. Springer Berlin Heidelberg, 2003, pp. 227–247.

[21] K. Kowalczyk, E. A. P. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Process. Lett.*, vol. 20, pp. 653–656, 2013.

[22] W. Xue, M. Brookes, and P. A. Naylor, "Under-modelled blind system identification for time delay estimation in reverberant environments," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016.

[23] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[24] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[25] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[26] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environment," *IEEE J. Sel. Topics Signal Process.*, 2019.

[27] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise de-reverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[28] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[29] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[30] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation of speech signals based on linear prediction," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, vol. 2, Jeju Island, Korea, Oct. 2004, pp. 877–881.

[31] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Multichannel Kalman filtering for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018.

[32] ——, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1833–1847, 2018.

[33] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[34] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[35] Y. Zeng and R. C. Hendriks, "Distributed estimation of the inverse of the correlation matrix for privacy preserving beamforming," *Signal Processing*, vol. 107, pp. 109–122, feb 2015.

[36] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[37] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.