Correlation Maximization-Based Sampling Rate Offset Estimation for Distributed Microphone Arrays

Lin Wang and Simon Doclo, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate the sampling rate mismatch problem in distributed microphone arrays and propose a correlation maximization algorithm to blindly estimate the sampling rate offset between two asynchronously sampled microphone signals. We approximate the sampling rate offset with a linear-phase drift model in the short-time Fourier transform (STFT) domain and show that the correlation coefficient between two microphone signals tends to present the highest value when the sampling of the two microphone signals is synchronized. Based on this finding we propose the correlation maximization algorithm, which performs sampling rate compensation on two microphone signals with different possible offset values and calculates their correlation coefficient after compensation. The offset value that leads to the largest correlation coefficient is chosen as the optimal estimate. Since the precision of the STFT linear-phase drift model used in the algorithm degrades as the sampling rate offset or the signal length is increased, we further propose a two-stage exhaustive search scheme to detect the optimal sampling rate offset. This scheme is able to minimize the influence of the linear-phase drift model error in order to improve the sampling rate offset estimation accuracy. Both simulated as well as real-world experiments confirm the effectiveness of the proposed algorithm.**

***Index Terms*—Correlation coefficient, distributed microphone array, sampling rate offset.**

I. Introduction

MICROPHONE array processing techniques, such as beamforming and blind source separation, are widely used for noise reduction due to their spatial filtering capability to suppress interfering signals that arrive from undesired directions [1]–[7]. Despite obvious advantages over single-microphone techniques, traditional microphone arrays have their limitations because they usually capture the sound field locally, typically at a relatively large distance from the sound sources. Furthermore, due to space and energy constraints, especially in portable devices the array is often limited in physical size and the number of microphones. In recent years, distributed microphone arrays or so called wireless acoustic sensor networks (WASN) have attracted increased attention from the research community due to their ability to overcome these limitations [8], [9]. A WASN generally consists of a set of wireless nodes that are spatially distributed over the environment, usually in an ad-hoc fashion. Each node contains at least one microphone, an analog/digital converter (ADC), and a processing and communication unit [10]. Via wireless communication, the microphone nodes can physically cover a much larger area, which vastly increases the amount of spatial information and overcomes the array-size limitation. All the microphones inside the network work collaboratively, either in a centralized or a distributed way, towards a common goal such as noise reduction or source localization. However, algorithm design for WASN is quite challenging due to several issues, such as communication bandwidth constraints [11], [12], sampling rate mismatch [13], [14], unknown microphone positions [17]–[19], and distributed signal processing [20]–[22]. This paper addresses the sampling rate mismatch problem between different nodes.

In a WASN, the nodes are individually connected to their own ADC and clock source, and capture the acoustic scene asynchronously. Since the oscillators or crystals, which are used for generating the clock signals, always have a certain tolerance in their nominal frequencies, a sampling rate mismatch between two independent nodes is inevitable. Depending on the used devices, the mismatch may range from just a few ppm to many hundreds of ppm (parts per million, 10^{-6}) [13], [14]. With mismatched sampling frequencies, the unit lengths of digital samples at two nodes become different. Consequently, the temporal information (e.g., time difference) between the two nodes drifts with time. This will significantly degrade the performance of most microphone array signal processing algorithms which assume unique time differences of arrival (TDOA) of the sound sources [13]–[16]. Thus, sampling rate offset correction is crucial for a distributed microphone network. Generally, two processing stages are involved: sampling rate offset (SRO) estimation and compensation.

Once the SRO between two microphones is known, the mismatch can be easily compensated. A straightforward way is to resample the digital audio stream in the time domain. Besides the traditional resampling approach using cascaded interpolators and decimators [23], sample-wise interpolation filters, such as the Lagrange polynomial interpolation method [10] and the "sinc" interpolation method [14], are commonly used in practical applications. Sampling rate compensation can also be performed in the short-time Fourier transform (STFT) domain [24].

Manuscript received July 02, 2015; revised November 21, 2015; accepted December 24, 2015. Date of publication January 12, 2016; date of current version February 23, 2016. The work was supported in part by the Postdoctoral Researcher Fellowship by the Alexander von Humboldt Foundation and in part by the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard Hendriks.

L. Wang is with the Signal Processing Group, Department of Medical Physics and Acoustics, University of Oldenburg, 26111 Oldenburg, Germany, and also with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: wanglin_2k@yahoo.com; lin.wang@uni-oldenburg.de).

S. Doclo is with the Signal Processing Group, Department of Medical Physics and Acoustics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: simon.doclo@uni-oldenburg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2517326

Estimating the SRO remains a challenging task, for which several approaches have been proposed, either in a referenced way or in a blind way.

- One example of the referenced approach is to broadcast known calibration signals among all the recording devices [13], [25]. To generate the calibration signal, frequency-modulated radio transmitters and receivers are required so that interference with recorded audio signals can be avoided. Another approach is to exchange network packet time-stamps between the nodes, based on which the clock skew between two nodes can be estimated [26], [27]. Network communication protocols are however required to support the exchange of time stamps.
- The SRO can also be blindly estimated from microphone signals without using any reference information. In [10] the SRO is estimated from spatially coherent but stationary environment noise, by analyzing the phase drift of the spatial coherence function of the microphone signals. In [24] a maximum likelihood method is proposed to estimate the SRO from the recording of physically static sound sources, by assuming a zero-mean multivariate Gaussian probability distribution for the STFT coefficients of the microphone signals. In [28] an SRO estimation algorithm is proposed in combination with blind source separation, where the microphone signals are first compensated with all possible offset values and then separated. The offset value which leads to the best separation result is regarded as the optimal estimate.

In this paper we investigate approaches for blind SRO estimation from the microphone signals. The contribution of the paper is summarized as below. First, by approximating the SRO with a linear-phase drift model in the STFT domain, we theoretically prove that the correlation between two independent microphones recording the same acoustic event is closely related to SRO, and that the correlation coefficient tends to show the largest value only when the sampling of the two microphones is synchronized. Second, we propose a correlation maximization based SRO estimation algorithm, which estimates the offset value as the one maximizing the correlation between two microphones after sampling rate compensation. We theoretically show that the algorithm performs well as long as the acoustic signal show a certain coherence in the microphones. Third, we propose a two-stage exhaustive search scheme to improve the robustness of the proposed algorithm to the STFT linear-phase modeling error, which becomes pronounced when the SRO is large or the considered signal is long. The effectiveness of the proposed algorithm is confirmed using both simulated as well as real-world experiments.

The paper is organized as follows. Section II introduces some background about the sampling rate mismatch problem, including linear-phase drift modeling and sampling rate compensation. The correlation maximization algorithm as well as the two-stage exhaustive search scheme are proposed in Section III. The relationship between the proposed correlation maximization algorithm and existing algorithms is discussed in Section IV. Experimental results are presented in Section V.

II. Preliminaries

A. Sampling Rate Offset Formulation in the STFT Domain

In [10], [24] it has been shown that the sampling rate mismatch problem can be approximated as a linear-phase drift model in the STFT domain. Since this model is the foundation of the proposed algorithm, we present a detailed derivation below.

Consider a continuous-time signal $\tilde{z}[t]$ which is sampled as two discrete-time signals $z_1(n)$ and $z_2(n)$, with sampling rates $f_{s1} = f_s$ and $f_{s2} = f_s + \varepsilon$, respectively, where $\varepsilon \ll f_s$ is the SRO, and t and n denote continuous time and discrete time indices, respectively. The relationship between $z_1(n)$, $z_2(n)$, and $\tilde{z}[t]$ can be expressed as

$$z_1(n) = \tilde{z}\left[\frac{n}{f_s}\right], \quad z_2(n) = \tilde{z}\left[\frac{n}{f_s+\varepsilon}\right]. \tag{1}$$

Using an N-point window $w(n)$, the STFT of $z_1(n)$ with shift N_s is given by

$$\begin{aligned} Z_1(k,l) &= \sum_{n=0}^{N-1} w(n) z_1(lN_s+n) \exp\left(-j\frac{2\pi kn}{N}\right) \\ &= \sum_{n=0}^{N-1} w(n)\tilde{z}\left[\frac{lN_s}{f_s}+\frac{n}{f_s}\right]\exp\left(-j\frac{2\pi kn}{N}\right), \end{aligned} \tag{2}$$

where k and l are the frequency and frame indices, respectively. Similarly, the STFT of $z_2(n)$ is expressed as

$$\begin{aligned} Z_2(k,l) &= \sum_{n=0}^{N-1} w(n)\tilde{z}\left[\frac{lN_s}{f_s+\varepsilon}+\frac{n}{f_s+\varepsilon}\right]\exp\left(-j\frac{2\pi kn}{N}\right) \\ &\approx \sum_{n=0}^{N-1} w(n)\tilde{z}\left[\frac{lN_s}{f_s}+\frac{n}{f_s}-\frac{lN_s\varepsilon}{f_s^2}\right]\exp\left(-j\frac{2\pi kn}{N}\right), \end{aligned} \tag{3}$$

where the approximation in (3) holds when $\varepsilon \ll f_s$.

Let us denote the l-th windowed frames of $z_1(n)$ and $z_2(n)$, respectively, as

$$z_{1l}^w(n) = w(n)\tilde{z}\left[\frac{lN_s}{f_s}+\frac{n}{f_s}\right], \tag{4}$$

and

$$z_{2l}^w(n) = w(n)\tilde{z}\left[\frac{lN_s}{f_s}+\frac{n}{f_s}-\frac{lN_s\varepsilon}{f_s^2}\right]. \tag{5}$$

If the contents in these two frames are similar, $z_{2l}^w(n)$ can be seen as a shifted version of $z_{1l}^w(n)$, with a non-integer shift $-\frac{lN_s\varepsilon}{f_s}$. The relationship between $Z_1(k,l)$ and $Z_2(k,l)$ can then be expressed as

$$\boxed{Z_2(k,l) \approx Z_1(k,l)\cdot\exp\left(-j\frac{2\pi k\left(\frac{lN_s\varepsilon}{f_s}\right)}{N}\right)} \tag{6}$$

Based on (6), the SRO can be approximated as a linear-phase drift model in the STFT domain: the magnitude remains unity while the phase varies linearly with respect to time and frequency. To satisfy the assumption of similar contents of $z_{1l}^w(n)$

and $z_{2l}^{w}(n)$, the shift size between the two frames should be small enough, i.e.,

$$\boxed{\frac{lN_s\varepsilon}{f_s} \ll N} \tag{7}$$

Due to this condition, the precision of the linear-phase drift model degrades for large SROs or long signals.

B. Resampling Techniques

The aim of resampling is to convert a digital sequence $z(n)$, with original sampling rate f_o, to a sequence $\hat{z}(n)$, with target sampling rate $f_d = f_o + \varepsilon$. Resampling can be performed either in the STFT domain or in the time domain.

1) STFT-domain Resampling: Based on the linear-phase drift model in (6), the relationship between $z(n)$ and $\hat{z}(n)$ can be expressed in the STFT domain as

$$\hat{Z}(k,l) = Z(k,l) \cdot \exp\left(-j\frac{2\pi k\left(\frac{lN_s\varepsilon}{f_o}\right)}{N}\right), \tag{8}$$

where $Z(k,l)$ and $\hat{Z}(k,l)$ are the STFTs of $z(n)$ and $\hat{z}(n)$, respectively. Resampling can hence be performed straightforwardly by modifying the phase of $Z(k,l)$. However, due to the limited precision of the linear-phase drift model (cf. (7)), the STFT-domain resampling approach can only roughly adjust the sampling rate.

2) Time-domain Resampling: In [10], [29] a time-domain fourth-order Lagrange interpolation algorithm has been presented for resampling. First, the original signal $z(n)$ is interpolated by a factor of 4, obtaining the interpolated signal $\tilde{z}(\tilde{n})$. Define $\dot{n} = \lfloor 4n\frac{f_o}{f_d} \rfloor$ as the closest index in $\tilde{z}(\tilde{n})$ from left to the time n/f_d, where the operation "$\lfloor\cdot\rfloor$" represents the integer part of the argument. The resampled signal can be calculated by using

$$\hat{z}(n) = \beta_1\tilde{z}(\dot{n}-1) + \beta_2\tilde{z}(\dot{n}) + \beta_3\tilde{z}(\dot{n}+1) + \beta_4\tilde{z}(\dot{n}+2), \tag{9}$$

where the definitions of the four interpolation coefficients $\beta_1 - \beta_4$ can be found in [10].

Compared with STFT-domain resampling, the time-domain approach is not affected by the SRO value nor the signal length, and can obtain a higher resampling precision. In this paper, both resampling techniques are used: STFT-domain resampling is used when estimating the SRO in the STFT domain, whereas time-domain resampling is used to adjust the sampling rate accurately.

III. Correlation Maximization Based Sampling Rate Offset Estimation

Assume an unknown reverberant and noisy acoustic environment. The signals received at two asynchronously sampled microphones are $x_1(n)$ and $x_2(n)$, with sampling rates f_s and $f_s + \varepsilon_o$, respectively, where ε_o is the SRO. The task is to blindly estimate ε_o from the microphone signals $x_1(n)$ and $x_2(n)$. After introducing the underlying principle in Section III-A, we propose the correlation maximization based SRO estimation algorithm in Section III-B and discuss factors that influence the estimation accuracy in Section III-C.

A. Correlation Coefficient Versus Sampling Rate Offset

Suppose $\bar{x}_2(n)$, which in practice is unknown, is the signal of the second microphone sampled at f_s. The STFTs of $x_1(n)$, $x_2(n)$ and $\bar{x}_2(n)$ are denoted as $X_1(k,l)$, $X_2(k,l)$ and $\mathfrak{X}_2(k,l)$, respectively. Based on the linear-phase drift model in (6), the relationship between $X_2(k,l)$ and $\mathfrak{X}_2(k,l)$ can be expressed as

$$\frac{X_2(k,l)}{\mathfrak{X}_2(k,l)} = \exp\left(-j\frac{2\pi k\left(\frac{lN_s\varepsilon_o}{f_s}\right)}{N}\right) = \mathrm{e}^{-j\alpha(k,l)}, \tag{10}$$

where

$$\alpha(k,l) = \frac{2\pi k\left(\frac{lN_s\varepsilon_o}{f_s}\right)}{N}, \tag{11}$$

denotes the phase drift due to sampling rate offset.

The correlation coefficient between $X_1(k,l)$ and $\mathfrak{X}_2(k,l)$ is defined as:

$$\rho_{ko}(X_1,\mathfrak{X}_2) = \frac{\sum_l X_1(k,l)\mathfrak{X}_2^*(k,l)}{\sqrt{\sum_l |X_1(k,l)|^2}\sqrt{\sum_l |\mathfrak{X}_2(k,l)|^2}}. \tag{12}$$

Using (10), the correlation coefficient between $X_1(k,l)$ and $X_2(k,l)$ is defined as:

$$\begin{aligned}\rho_k(X_1,X_2) &= \frac{\sum_l X_1(k,l)X_2^*(k,l)}{\sqrt{\sum_l |X_1(k,l)|^2}\sqrt{\sum_l |X_2(k,l)|^2}} \\ &= \frac{\sum_l X_1(k,l)\mathfrak{X}_2^*(k,l)\mathrm{e}^{j\alpha(k,l)}}{\sqrt{\sum_l |X_1(k,l)|^2}\sqrt{\sum_l |\mathfrak{X}_2(k,l)|^2}}.\end{aligned} \tag{13}$$

We refer to ρ_{ko} as the *synchronous correlation coefficient*, and ρ_k the *asynchronous correlation coefficient*.

Considering that in (13) the term $X_1(k,l)\mathfrak{X}_2^*(k,l)$ depends on unknown acoustic events in the environment while the term $\alpha(k,l)$ in (11) depends on the SRO, it is reasonable to assume that these two terms are statistically independent of each other, i.e.,

$$\begin{aligned}&\mathrm{E}\left(X_1(k,l)\mathfrak{X}_2^*(k,l)\mathrm{e}^{j\alpha(k,l)}\right) \\ &\quad = \mathrm{E}(X_1(k,l)\mathfrak{X}_2^*(k,l)) \cdot \mathrm{E}(\mathrm{e}^{j\alpha(k,l)}),\end{aligned} \tag{14}$$

where $\mathrm{E}(\cdot)$ denotes mathematical expectation. Assuming ergodic processes where the expectation operator over realizations can be replaced by time-domain averaging over frames, in each frequency k it follows that,

$$\begin{aligned}&\frac{1}{L}\sum_l \left(X_1(k,l)\mathfrak{X}_2^*(k,l)\mathrm{e}^{j\alpha(k,l)}\right) \\ &\approx \frac{1}{L}\sum_l (X_1(k,l)\mathfrak{X}_2^*(k,l)) \cdot \frac{1}{L}\sum_l \mathrm{e}^{j\alpha(k,l)},\end{aligned} \tag{15}$$

where L is the total number of considered time frames.

Using (12), (13) and (15), the relationship between ρ_k and ρ_{ko} can now be expressed as

$$\rho_k(X_1, X_2) = \rho_{ko}(X_1, \mathfrak{X}_2)\frac{1}{L}\sum_l \mathrm{e}^{j\alpha(k,l)}. \tag{16}$$

Taking the absolute value of both sides of (16), it follows that,

$$\boxed{|\rho_k| = |\rho_{ko}| \cdot |\tfrac{1}{L}\sum_l \mathrm{e}^{j\alpha(k,l)}| = |\rho_{ko}| \cdot \gamma_{kL}(\varepsilon_o)} \tag{17}$$

where

$$\gamma_{kL}(\varepsilon) = \left|\frac{1}{L}\sum_l \mathrm{e}^{j\alpha(k,l)}\right| = \left|\frac{1}{L}\sum_l \exp\left(-j\frac{2\pi k\left(\frac{lN_s\varepsilon}{f_s}\right)}{N}\right)\right|, \tag{18}$$

is defined as the *attenuation factor*.

It can be easily verified that $\gamma_{kL}(\varepsilon)$ lies in the interval [0,1] and is equal to 1 only when $\varepsilon = 0$. Accordingly, $|\rho_k|$ will be equal to $|\rho_{ko}|$ only when $\varepsilon = 0$, and be smaller than $|\rho_{ko}|$ otherwise. Next, we will exploit this property to estimate the SRO.

B. Proposed Algorithm

1) Correlation Maximization (CM) Algorithm: Since the maximum correlation coefficient is achieved only when the two microphone signals are sampled synchronously, we propose to compensate the sampling rate of the microphone signals with different possible offsets and estimate the optimal value as the one that maximizes the correlation coefficient between the compensated microphone signals. This is expressed as

$$\tilde{\varepsilon}_{\mathrm{CM}}(k) = \arg\max_{\varepsilon}\{|\rho_{k\varepsilon}|\}, \tag{19}$$

where $\rho_{k\varepsilon}$ is the correlation coefficient between X_1 and $X_{2\varepsilon}$, where $X_{2\varepsilon}$ is obtained by compensating X_2 with an SRO ε. For convenience of computation, the compensation is performed in the STFT domain directly (cf. Section II-B). After sampling rate compensation, the correlation coefficient is expressed as

$$\begin{aligned}\rho_{k\varepsilon} &= \rho_k\left(X_1(k,l), X_{2\varepsilon}(k,l)\right)\\ &= \rho_k\left(X_1(k,l), X_2(k,l)\exp\left(j\frac{2\pi k\left(\frac{lN_s\varepsilon}{f_s}\right)}{N}\right)\right)\\ &= \rho_{ko}\cdot\frac{1}{L}\sum_l \exp\left(j\frac{2\pi k\left(\frac{lN_s(\varepsilon-\varepsilon_o)}{f_s}\right)}{N}\right).\end{aligned} \tag{20}$$

Similarly to (17), it hence follows that

$$|\rho_{k\varepsilon}| = |\rho_{ko}| \cdot \gamma_{kL}(\varepsilon_o - \varepsilon). \tag{21}$$

Obviously, $|\rho_{k\varepsilon}|$ is maximized when $\varepsilon = \varepsilon_o$. Considering the whole frequency band by defining

$$\bar{\rho}_\varepsilon = \sum_k |\rho_{k\varepsilon}|, \tag{22}$$

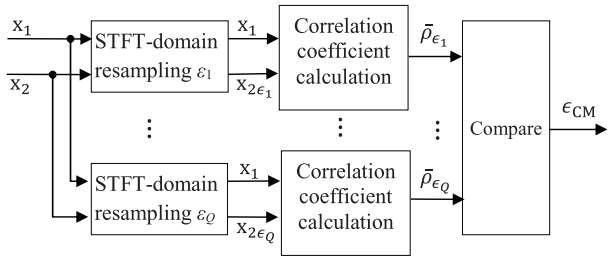


Fig. 1. Block diagram of the exhaustive search scheme in the STFT domain.

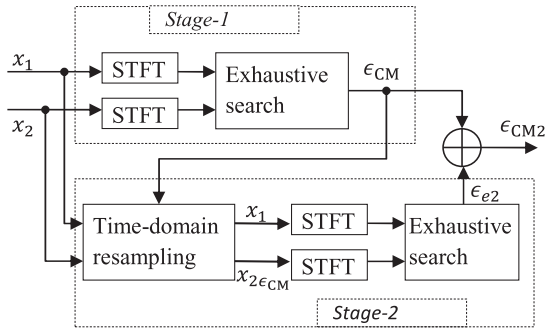


Fig. 2. Block diagram of the two-stage processing for sampling rate offset estimation.

the offset value can be estimated as

$$\boxed{\varepsilon_{\mathrm{CM}} = \arg\max_{\varepsilon}\bar{\rho}_\varepsilon} \tag{23}$$

The solution to (23) is difficult to be expressed in an analytical form. However, since there is only one parameter to be optimized, the optimal value can be straightforwardly determined using an exhaustive search scheme as shown in Fig. 1, i.e., the microphone signals are compensated with all possible Q offset values and the offset which leads to the maximum correlation coefficient is selected as the optimal estimate.

2) Two-stage Search Scheme (CM-2): The correlation maximization algorithm is derived based on the linear-phase drift model in the STFT domain. However, as stated in Section II-A, the precision of this model degrades when the SRO is large or when the considered signal is long. To overcome this drawback, an improved search scheme is proposed which involves two stages as shown in Fig. 2.

In the first stage, after transforming the time-domain signals $x_1(n)$ and $x_2(n)$ into the STFT domain, an exhaustive search (as depicted in Fig. 1) is applied in the STFT domain to estimate the offset $\varepsilon_{\mathrm{CM}}$. This estimate usually deviates from the true offset value due to the linear-phase drift model error.

In the second stage, a time-domain resampling (9) is applied to $x_2(n)$ using $\varepsilon_{\mathrm{CM}}$. Next, the offset between $x_1(n)$ and the compensated signal $x_{2\varepsilon_{\mathrm{CM}}}(n)$ is estimated as ε_{e2}, using the same exhaustive search scheme in the STFT domain. As mentioned in Section II-B, the precision of time-domain resampling is independent of SRO and signal length. After time-domain resampling, the residual offset between $x_1(n)$ and $x_{2\varepsilon_{\mathrm{CM}}}(n)$, which is equal to $\varepsilon_o - \varepsilon_{\mathrm{CM}}$, is much smaller than the original offset ε_o between $x_1(n)$ and $x_2(n)$. As the precision of the linear-phase drift model improves for smaller offset values, the

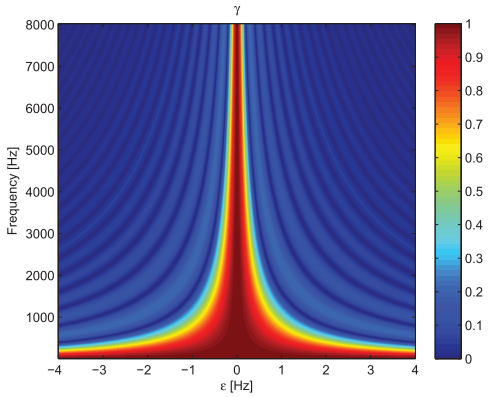


Fig. 3. Variation of γ_{kL} with respect to frequency and SRO ε (signal length 10 s, $f_s = 16$ kHz, $N = 1024$, $N_s = N/2$).

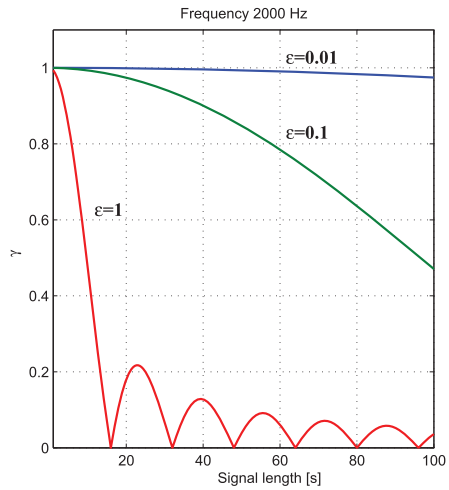


Fig. 4. Variation of γ_{kL} with respect to signal length for a fixed ε, at frequency 2000 Hz.

residual offset can be estimated much more accurately. The final offset is computed as the sum of the two estimates, i.e.,

$$\boxed{\varepsilon_{\mathrm{CM2}} = \varepsilon_{\mathrm{CM}} + \varepsilon_{e2}} \tag{24}$$

To further refine the SRO estimate, it is possible to employ even more processing stages. However, in practice, we have found that two stages are typically sufficient to obtain an accurate SRO estimate.

C. Discussion

Based on (21), the estimation accuracy of the proposed CM algorithm mainly depends on two factors: the attenuation factor γ_{kL} and the synchronous correlation coefficient ρ_{ko}.

The attenuation factor γ_{kL} plays an important role when locating the maximum correlation coefficient. The larger the sensitivity of $\gamma_{kL}(\varepsilon)$ to ε, the easier it is to detect the correct offset value. As defined in (18), γ_{kL} is determined by three parameters: the frequency k, the signal length L, and the SRO ε. For $f_s = 16$ kHz, $N = 1024$, $N_s = N/2$, $L = 312$ (corresponding to a signal length of 10 s), and ε varying between -4 Hz and 4 Hz, Fig. 3 shows, as an example, how γ_{kL} varies with frequency and SRO. From Fig. 3 it can be verified that γ_{kL} lies in the interval [0,1] and is only equal to 1 when $\varepsilon = 0$. In addition, it can be observed that γ_{kL} decreases for increasing ε, and that the decrease is more pronounced at higher frequencies than at lower frequencies. This demonstrates that it is easier to achieve an accurate estimation at high frequencies. For frequency 2000 Hz, Fig. 4 depicts how γ_{kL} varies with signal length (from 0 to 100 s) when ε is fixed at 0.01 Hz, 0.1 Hz, and 1 Hz, respectively. It can be observed that for all three offsets γ_{kL} decreases with increasing signal length. The decrease is evident when the offset is large (e.g., 0.1 Hz and 1 Hz). For the small offset (0.01 Hz), the decrease of γ_{kL} can be observed only when the signal is long enough. This demonstrates that using a long signal is important for improving the estimation accuracy. It should be noted that the precision of the linear-phase drift model degrades with increasing signal length and this issue can be addressed by the two-stage search scheme in the CM-2 algorithm.

The estimation accuracy also depends on the synchronous correlation coefficient ρ_{ko}. The larger ρ_{ko}, the easier it is to detect the variation of $\gamma_{KL}(\varepsilon)$ and hence the correct offset value. In the extreme case when $\rho_{ko} = 0$, $\rho_{k\varepsilon}$ in (21) does not vary with $\gamma_{KL}(\varepsilon)$, making it impossible to detect the correct SRO, even when ε is large. The synchronous correlation coefficient ρ_{ko} mainly depends on the acoustic environment.

- Static environment. For a single static coherent source, the synchronous correlation coefficient between two microphones is large for all frequencies. For a static diffuse-like environment, e.g., a mixture of multiple coherent sources, the synchronous correlation coefficient between two microphones is small at high frequencies and large at low frequencies, depending on the inter-microphone distance. When the distance between two microphones is too large, the microphone signals become uncorrelated.
- Dynamic environment, e.g., one or multiple moving sources. Depending on the trajectory and the speed, a moving source may create time-varying delays between two microphone signals (similar to sampling rate mismatch), such that the peak location of the correlation coefficient function may deviate from the true SRO value. Since the movement of the source is typically unknown, the deviation of the peak location is unpredictable, especially when the signal is short. In general, the proposed algorithm can only recover the SRO up to this unknown deviation and hence will not work for moving sources. However, there exist special cases in which the peak location deviation induced by the source movement is close to 0. For instance, when the source is moving around the microphones (i.e., there is enough spatial diversity) and the signal is long enough, the moving source can be considered as a combination of static sources distributed around the microphones, exhibiting diffuse-like characteristics. In this case, the synchronous correlation coefficient in (12) tends to be small at high frequencies and large at low frequencies (depending on the inter-microphone distance), and the deviation of the peak location tends to be 0, such that the proposed algorithm may reach the correct SRO estimate. This will be experimentally validated in Section V-D.

IV. Relationship with Existing Algorithms

In this section the relationship between the proposed algorithm and two other algorithms which exploit the STFT-domain linear-phase drift model to estimate the SRO [10], [24] are explored.

A. Linear-phase Drift (LPD) Estimation

In [10] the phase shift of the correlation coefficient is exploited to estimate the SRO. Given $X_1(k,l)$, $X_2(k,l)$ and $\mathfrak{X}_2(k,l)$, the synchronous and asynchronous instantaneous correlation coefficients are defined as

$$\varrho_o(k,l) = \frac{X_1(k,l)\mathfrak{X}_2^*(k,l)}{\sqrt{|X_1(k,l)|^2|\mathfrak{X}_2(k,l)|^2}}, \tag{25}$$

and

$$\varrho(k,l) = \frac{X_1(k,l)X_2^*(k,l)}{\sqrt{|X_1(k,l)|^2|X_2(k,l)|^2}}, \tag{26}$$

where 'instantaneous' means that the correlation coefficient is calculated per time-frequency bin.

Assuming that the sound sources in the acoustic environment are stationary, i.e., $\varrho_o(k,l) = \varrho_o(k)$, it can be easily shown that

$$\varrho(k,l) = \varrho_o(k)\exp\left(j\frac{2\pi N_s k l \varepsilon_o}{N f_s}\right) = \varrho_o(k)\exp\left(j\alpha(k,l)\right). \tag{27}$$

The SRO can be estimated from the variation (with respect to time) of the phase term in (27), i.e.,

$$\boxed{\varepsilon_{\mathrm{LPD}}(k) = \frac{\angle(\varrho(k,l_2)/\varrho(k,l_1))}{\frac{2\pi N_s k(l_2-l_1)}{N f_s}}} \tag{28}$$

where $\angle(\cdot)$ denotes the phase of the argument, and l_1 and l_2 denote the indices of two neighbouring time frames. The estimate $\varepsilon_{\mathrm{LPD}}(k)$ is equal to ε_o at all frequencies.

In [10], it has been proposed to further improve the estimation accuracy of the LPD algorithm by averaging the estimates across multiple time segments and multiple valid frequency bins (where no frequency aliasing occurs). More specifically, by dividing the complete signal into I segments (each containing P frames) and averaging the estimates at the I time segments and also the $\bar{K}$ valid frequency bins leads to

$$\bar{\varepsilon}_{\mathrm{LPD}} = \frac{1}{\bar{K}}\sum_k \frac{N f_s}{2\pi P N_s k}\angle\left\{\frac{1}{I-1}\left(\sum_{i=1}^{I-1}\frac{\varrho(k,iP)}{\varrho(k,(i-1)P)}\right)\right\}. \tag{29}$$

The averaging in theory will not change the estimate, but in practice can help to reduce the influence of outlier estimation using one time segment or one frequency bin alone.

B. Maximum Likelihood (ML) Estimation

In [24] the spatial stationarity of the acoustic environment is exploited to estimate the SRO. Assuming the STFT coefficients of the synchronously sampled microphone signals can be modelled using a zero-mean multivariate normal distribution, the SRO can be estimated by maximizing a likelihood function, which evaluates the fit with a zero-mean multivariate normal distribution. The likelihood function is defined as

$$J(\boldsymbol{V}_\varepsilon,\varepsilon) = \sum_{k,l}\left\{-\log\pi^2 - \log(\det(\boldsymbol{V}_\varepsilon(k)) - \boldsymbol{X}_\varepsilon(k,l)^{\mathrm{H}}\boldsymbol{V}_\varepsilon(k)^{-1}\boldsymbol{X}_\varepsilon(k,l)\right\}, \tag{30}$$

where $\det(\cdot)$ denotes the determinant of a matrix, $\boldsymbol{X}_\varepsilon(k,l) = [X_1(k,l), X_{2\varepsilon}(k,l)]^{\mathrm{T}}$, and $\boldsymbol{V}_\varepsilon(k) = \frac{1}{L}\sum_l \boldsymbol{X}_\varepsilon(k,l)\boldsymbol{X}_\varepsilon^{\mathrm{H}}(k,l)$ is the covariance matrix. After derivation, the SRO can be estimated as

$$\boxed{\varepsilon_{\mathrm{ML}} = \arg\max_\varepsilon\left\{-\sum_k \log\det\left(\sum_l \boldsymbol{X}_\varepsilon(k,l)\boldsymbol{X}_\varepsilon^{\mathrm{H}}(k,l)\right)\right\}} \tag{31}$$

The optimization problem in (31) can not be solved analytically, and thus an exhaustive search scheme can be employed to find the optimal solution. In [24], a golden section search scheme is proposed to accelerate the search speed.

C. Relationship Between the Algorithms

The LPD algorithm and the proposed CM algorithm both exploit the concept of linear-phase drift to estimate the SRO. Using instantaneous correlation coefficients at different time frames, the LPD algorithm analytically estimates the slope of the linear-phase drift, from which the SRO can be easily calculated. The LPD algorithm is straightforward; however, it requires the acoustic environment to be stationary such that the phase of the synchronous correlation coefficient in (25) does not vary with time. This assumption is not always met in practice. In contrast, the CM algorithm does not need this assumption. By exploiting the independence between the phase drift and the acoustical signals, it estimates the SRO by maximizing the (non-instantaneous) correlation coefficient between compensated microphone signals.

Although at first sight the ML and the proposed CM algorithm are derived from totally different perspectives, i.e., the ML algorithm maximizes the zero-mean multivariate normal distribution of the compensated signals and the CM algorithm maximizes the correlation coefficient between the compensated signals, it can be shown that they are closely related to each other. Using (20), the objective function of the CM algorithm in (22) can be rewritten as

$$J_{\mathrm{CM}} = \sum_k\left|\frac{\sum_l X_1 X_{2\varepsilon}^*}{\sqrt{\sum_l |X_1|^2\sum_l |X_2|^2}}\right|, \tag{32}$$

whereas, based on (31), the objective function of the ML algorithm can be rewritten as

$$\begin{aligned} J_{\mathrm{ML}} &= -\sum_k \log\det\left(\begin{bmatrix}\sum_l |X_1|^2 & \sum_l X_1 X_{2\varepsilon}^* \\ \sum_l X_{2\varepsilon} X_1^* & \sum_l |X_{2\varepsilon}|^2\end{bmatrix}\right) \\ &= -\sum_k \log\left(\sum_l |X_1|^2\sum_l |X_2|^2 - |\sum_l X_1 X_{2\varepsilon}^*|^2\right), \end{aligned} \tag{33}$$

where for conciseness the indices (k, l) have been neglected in (32) and (33). Both objective functions are quite similar since they both aim to maximize the term $|\sum_l X_1 X_{2\varepsilon}^*|$. The difference is that (32) employs a normalization processing, whereas (33) employs a logarithm operation. We nevertheless expect the SRO estimation performance of both the CM and the ML algorithms to be similar.

V. Experimental Results

The experiment section is divided into four parts. The first part shows examples of linear-phase drift modeling of sampling rate offset (Section V-A). The second part compares the performance of the proposed SRO estimation algorithm with existing algorithms for three acoustic scenarios: a stationary environment with diffuse noise (Section V-B), a stationary environment with static speakers (Section V-C), and a dynamic environment with moving speakers or time-varying acoustic events (Section V-D). The third part applies the considered SRO estimation and resampling algorithms to blind source separation with simulated as well as real-world data (Section V-E and V-F). Finally, we comment on the computational complexity of the considered algorithms (Section V-G).

The performance of the SRO estimation algorithms is evaluated using the SRO estimation error, which is defined as the difference between the estimated offset ε_e and the true offset value ε_o, i.e.,

$$E = |\varepsilon_e - \varepsilon_o|. \tag{34}$$

Four SRO estimation algorithms are considered:

- The linear-phase drift (LPD) algorithm [10].
- The maximum likelihood (ML) algorithm [24] with an exhaustive search scheme.
- The proposed correlation maximization (CM) algorithm with an exhaustive search scheme.
- The proposed correlation maximization algorithm with an improved two-stage search scheme (CM-2) depicted in Fig. 2.

For all experiments the nominal sampling rate is $f_s = 16$ kHz, and the STFT frame length is equal to 1024 with 50% overlap. ML and CM use the same exhaustive search scheme, where the exhaustive search area is equal to $[-10, 10]$ Hz with a precision of 0.0005 Hz. Similarly to [10], the LPD algorithm averages the estimates across multiple time segments and frequency bins, with the length of each time segment being 0.5 s.

A. Linear-phase Drift Modeling

For this experiment a white Gaussian noise signal $z_1(n)$ with a length of 30 s has been used. The sampling rate of the test signal $z_2(n)$ is adjusted to $f_s + \varepsilon$, using the time-domain resampling method presented in Section II-B. The original and the resampled signals are transformed into the STFT domain and denoted as $Z_1(k, l)$ and $Z_2(k, l)$, respectively. According to (6), the phase of $\frac{Z_2(k,l)}{Z_1(k,l)}$ should in theory vary linearly with respect to time while the magnitude should always be 1.

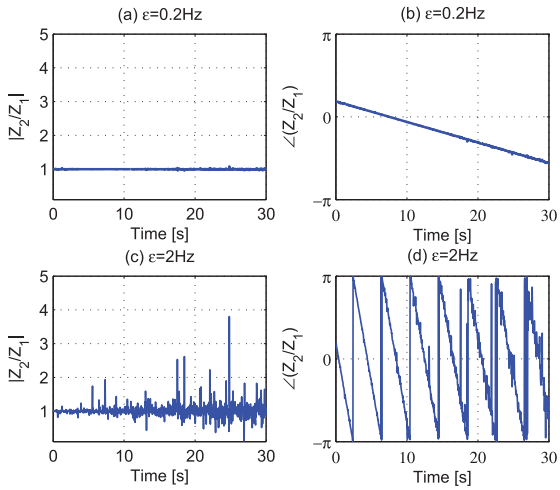


Fig. 5. Magnitude and phase of the linear-phase drift model in the STFT domain, at frequency 2000 Hz for the SRO (a) (b) $\varepsilon = 0.2$ Hz and (c) (d) $\varepsilon = 2$ Hz.

Fig. 5 depicts the magnitude and the phase of $\frac{Z_2(k,l)}{Z_1(k,l)}$ for SRO $\varepsilon = 0.2$ Hz and $\varepsilon = 2$ Hz, respectively, at frequency 2000 Hz. The upper panels illustrate that the linear-phase drift model holds well for a small offset $\varepsilon = 0.2$ Hz. However, the precision of the model degrades when the offset is increased, as illustrated by the lower panels for $\varepsilon = 2$ Hz. It is additionally observed that the model remains precise in the beginning of the signal, but worsens as time evolves. This demonstrates that the precision of the linear-phase drift model also depends on the signal length. In summary, the observations made in Fig. 5 are consistent with condition (7).

B. Diffuse Noise

In this experiment the performance of the SRO estimation algorithms is evaluated in spherically diffuse white noise, generated using the algorithm in [31], for different inter-microphone distances and signal lengths. The considered inter-microphone distances are $\{2, 10, 20, 50, \infty\}$ cm, where in principle the correlation of the microphone signals decreases with increasing inter-microphone distance and ∞ denotes uncorrelated noise. The considered signal lengths are $\{10, 20, 40, 60\}$ s. The second microphone signal is resampled using the time-domain resampling method presented in Section II-B, where the SRO is randomly chosen between $[-10, 10]$ Hz with a precision of 0.01 Hz. We implement 100 realizations, where for each realization the considered algorithms are applied to estimate the SRO. The average estimation error is calculated from the 100 realizations.

As an example, Fig. 6 depicts the two-stage search results of CM-2, for a signal length of 10 s, inter-microphone distance 10 cm, and true SRO 8.33 Hz. The y-axis denotes the correlation measure computed by (22). As shown in Fig. 6, for both stages a peak value can be clearly detected. Due to the limited precision of the linear-phase drift model, the estimated SRO in the first stage is 8.2845 Hz, which deviates from the true value. After sampling rate compensation, the residual offset is

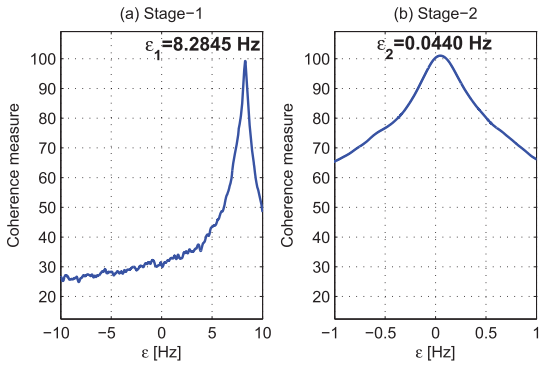


Fig. 6. Example of the two-stage search by CM-2. The true SRO is 8.33 Hz. The estimation results in (a) Stage-1 and (b) Stage-2 are 8.2845 Hz and 0.0440 Hz, respectively. The final SRO estimate is 8.3285 Hz.

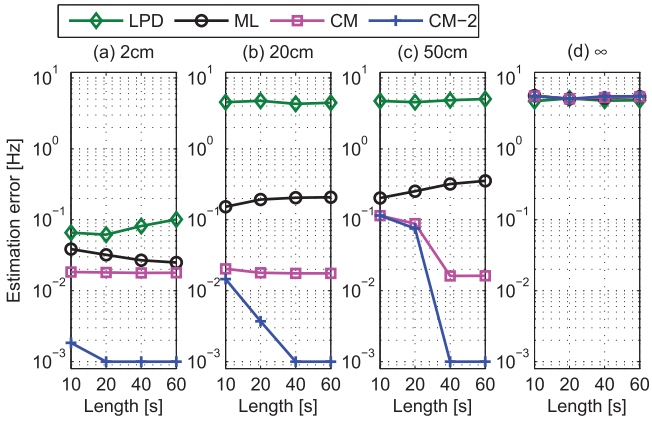


Fig. 7. Average SRO estimation error for the considered algorithms in diffuse noise with an inter-microphone distance of (a) 2 cm, (b) 20 cm, (c) 50 cm, and (d) ∞ (corresponding to uncorrelated noise).

estimated as 0.0440 Hz in the second stage. The final SRO is estimated as 8.3285 Hz, which is very close to the true value.

The average SRO estimation errors in diffuse noise for the four considered algorithms are shown in Fig. 7 (in log-scale), with each panel denoting a different inter-microphone distance. For convenience the estimation error is limited to a lower value of 0.001 Hz. The following observations can be made.

1) LPD yields the worst performance among all considered algorithms, especially when the microphones are far apart ($\geq$ 20 cm). Since the correlation between two microphones decreases as the inter-microphone distance increases, the assumption of a time-invariant instantaneous correlation coefficient does not hold any more, especially at higher frequencies.

2) For ML, CM, and CM-2, the estimation performance degrades when the inter-microphone distance increases. As shown in the fourth panel, all algorithms fail when the microphone distance is ∞, i.e., the microphone signals are completely uncorrelated.

3) For both CM and CM-2, the performance improves with the signal length. CM-2 significantly outperforms CM for the first three inter-microphone distances when the signal is long enough ($\geq$ 40 s).

4) CM outperforms ML for all considered scenarios, although they have similar objective functions (cf. (32) and (33)). A possible explanation for the better performance of CM is that the normalization operation in (32) reduces the dynamic range of the data, making it more robust to linear-phase drift modeling errors.

5) For ML, it can be observed that at 2 cm inter-microphone distance the performance improves with signal length while at 20 cm and 50 cm inter-microphone distances the performance degrades with signal length. For ML and CM, it should be realized that there are both advantages and disadvantages when using long signals for SRO estimation. On the one hand, increasing the signal length increases the sensitivity of the attenuation factor to the SRO (cf. Section III-C), thus improving the estimation accuracy. On the other hand, increasing signal length increases linear-phase drift modelling errors, thus degrading the estimation performance especially when the signal correlation is low (e.g., at high frequencies and large inter-microphone distances for diffuse noise). Thus the SRO estimation performance also degrades with increasing microphone distance. For ML, the advantages seem to exceed the disadvantages for a small inter-microphone distance (2 cm), leading to an improved estimation performance with increasing signal length. However, for a large inter-microphone distance (20 cm and 50 cm), the disadvantages seem to exceed the advantages, leading to a degraded estimation performance with increasing signal length. In contrast, CM seems to be more robust against linear-phase drift modelling errors, hence showing an improved estimation performance with increasing signal length for all considered inter-microphone distances (2 cm, 20 cm, 50 cm).

C. Static Environments

In this experiment one or three static speakers are talking in a simulated enclosure of size $4\text{ m} \times 5\text{ m} \times 2.5\text{ m}$ with a reverberation time 400 ms. The distance between two microphones is 20 cm. The speakers are randomly placed at a distance of 2 m from the microphones. The room impulse responses between the speakers and the microphones have been generated using the image-source method [30]. The considered signal lengths are $\{10, 20, 40, 60\}$ s. The second microphone signal is resampled with an SRO randomly chosen between $[-10, 10]$ Hz with a precision of 0.01 Hz. We implement 100 realizations and calculate the average estimation error from all realizations.

The average SRO estimation errors for the four considered algorithms are shown in Fig. 8. LPD yields the worst performance since the instantaneous correlation coefficient varies with time, due to non-stationarity of speech signals. ML and CM perform similarly for all considered scenarios. CM-2 yields the best performance and can achieve very accurate SRO estimation when the signal is long enough.

D. Dynamic Environments

In this experiment the performance for two dynamic acoustic environments are evaluated. The first scenario considers switched acoustic events, i.e., the first half of the microphone

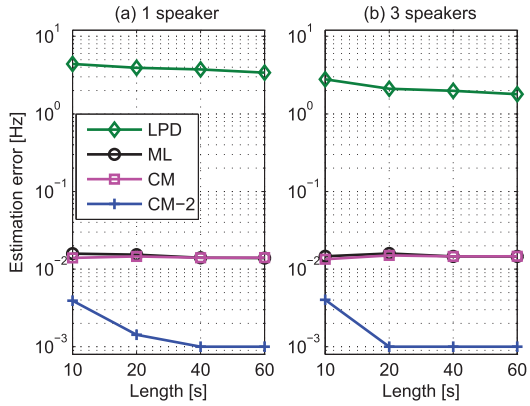


Fig. 8. Average SRO estimation error for the considered algorithms in a static acoustic environment with (a) 1 speaker and (b) 3 speakers.

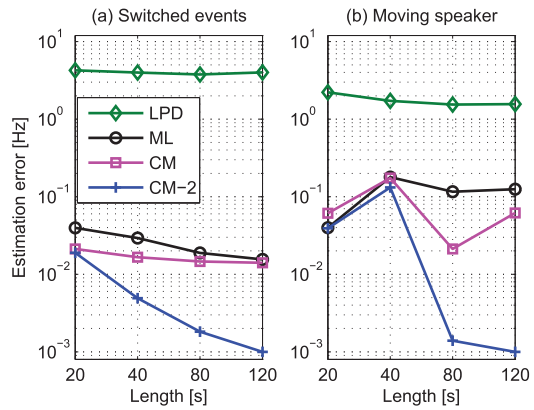


Fig. 9. Average SRO estimation error for the considered algorithms in a dynamic acoustic environment with (a) switched events and (b) a moving speaker.

signal contains a static speaker while the second half contains pure diffuse noise. In the second scenario, we consider the special case (cf. Section III-C) where a speaker is moving around the microphones inside the same simulated enclosure in Section V-C. The speaker is moving back and forth between $0°$ and $180°$ along a circle with a radius of 2 m (at a speed of $3.3°/s$). For both scenarios the inter-microphone distance is 20 cm and the considered signal lengths are $\{20, 40, 80, 120\}$ s. The second microphone signal is resampled with an SRO randomly chosen between $[-10, 10]$ Hz with a precision of 0.01 Hz. We implement 100 realizations and calculate the average estimation error from all realizations.

In comparison to static environments, estimating the SRO in dynamic environments is a challenging task. The average SRO estimation errors by the four considered algorithms are shown in Fig. 9. LPD again yields the largest estimation error for both dynamic scenarios. For the first scenario with switched acoustic events, the performance of CM, CM-2, and ML improves with increasing signal length. CM performs slightly better than ML. CM-2 yields the best performance, almost achieving perfect estimation when the signal is longer than 80 s. In the second scenario with a moving speaker, CM performs slightly better than ML although both algorithms yield relatively large estimation errors (> 0.02 Hz), even for long signals. CM-2 is still able

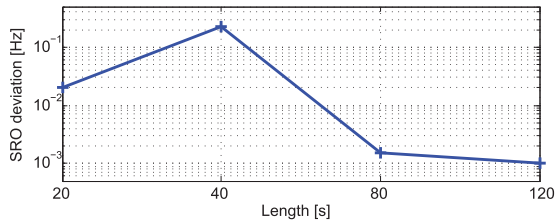


Fig. 10. SRO deviation for a moving speaker with different signal lengths.

to achieve very accurate estimation provided that the signal is long enough (> 80 s).

For the second scenario, the large estimation error of CM-2 for short signals is mainly due to speaker movement, such that the peak location of the correlation coefficient function deviates from the true SRO value (cf. discussion in Section III-C). To support this argument, we have computed the deviation of SRO estimation due to speaker movement by applying CM-2 to two synchronized microphone signals (i.e., SRO being 0) for the scenario with a moving speaker. Fig. 10 depicts the obtained SRO estimation deviation for different signal lengths. As can be observed in Fig. 10, the SRO estimation deviation is relatively large (and unpredictable) for short signals. However, when the signal is long enough, the SRO estimation deviation approaches 0. The SRO estimation deviation curve in Fig. 10 is consistent with the SRO estimation error curve for CM-2 in Fig. 9(b), which is obtained with SRO varying between $[-10, 10]$ Hz. This demonstrates that CM-2 can recover the SRO up to the deviation due to speaker movement.

E. Application to Blind Source Separation

In this section we evaluate the influence of sampling rate offset on the performance of a frequency-domain blind source separation (BSS) algorithm [6] and we evaluate the performance improvement when applying the considered SRO estimation algorithms. We simulate a similar scenario as in Section V-C with two static speakers and two microphones with an inter-microphone distance of 20 cm. The two speakers are randomly placed at a distance of 2 m from the microphones. The considered signal lengths are $\{10, 20, 40, 60, 100\}$ s. The second microphone signal is resampled with an SRO chosen from the set $\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}, \cdots, 5, 10, 15, 20\}$ Hz. The BSS performance is evaluated using the output signal-to-interference ratio (SIR), which is obtained by averaging the SIRs of two sources at the BSS output. The SIR is calculated with a toolbox in [32]. For all testing scenarios, the input SIR of the microphone signals is around 0 dB.

Fig. 11 depicts the BSS performance for various SROs and signal lengths without sampling rate compensation. When the SRO $\varepsilon \leq 0.01$ Hz, the BSS performance for the considered signal lengths is hardly affected by sampling rate offset. When $\varepsilon > 0.01$ Hz, it can be clearly observed that the BSS performance degrades as ε increases. The influence of sampling rate offset on the BSS performance highly depends on the signal length: long signals are more sensitive to sampling rate offset than short signals. When $\varepsilon > 1$ Hz, BSS fails in most cases.

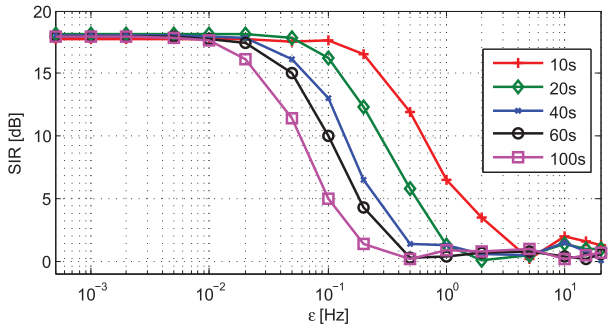


Fig. 11. BSS performance in terms of output SIR for various SROs and signal lengths without sampling rate compensation.

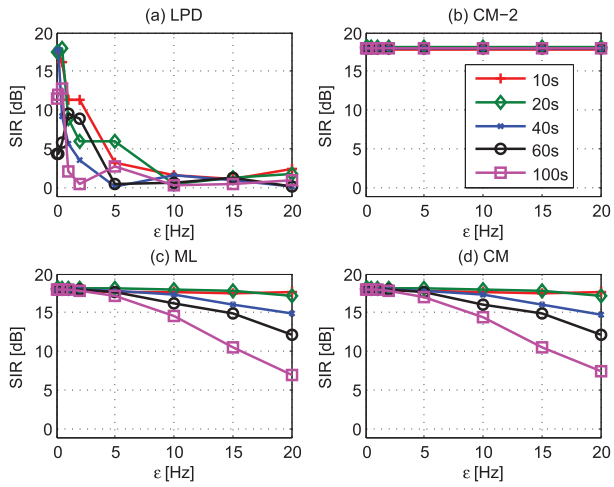


Fig. 12. BSS performance in terms of output SIR for various SROs and signal lengths after sampling rate compensation with (a) LPD, (b) CM-2, (c) ML, and (d) CM.

Fig. 12 depicts the BSS performance after sampling rate compensation with the considered algorithms. In general, LPD yields the worst performance; the performance of CM is similar to the performance of ML, and CM-2 yields the best performance, which is consistent with the observations made in the previous sections. The output SIR obtained by CM-2 remains constant at 18 dB for the considered SROs and signal lengths. ML and CM yield a decreased SIR when the SRO or the signal length is increased. More specifically, for signals shorter than 20 s, the output SIRs of ML and CM remain almost constant at 18 dB for the considered SROs. For signals longer than 20 s, their output SIRs start to drop with increasing ε, especially when $\varepsilon > 5$ Hz.

F. Real-World Experiment

In this experiment we evaluate the BSS performance using recorded data with real devices. The recording is made in a room of size 7 m $\times$ 7 m $\times$ 3 m with a reverberation time of about 400 ms. We use two independent devices, a Samsung S3 smartphone and a Gopro Hero3 camera, which are placed 10 cm apart in the center of the room. The speech signals are played through two Genelec 8010 loudspeakers, which are randomly placed at a distance of 1 m from the recording devices. The signal length is 100 s. The recordings contain slight noise from an air conditioner. The sampling rate of the devices are 44.1 kHz (Samsung) and 48 kHz (Gopro), respectively. The recorded signals from both devices are downsampled to 16 kHz before processing. The SRO between both devices is about 5.5 Hz at a sampling rate of 16 kHz. To obtain the clean reference signals in the microphones, which are required to calculate the SIR [32], we first played speech signals separately through each loudspeaker and then simultaneously through both loudspeakers. For all testing scenarios, the input SIR of the microphone signals is around 3 dB.

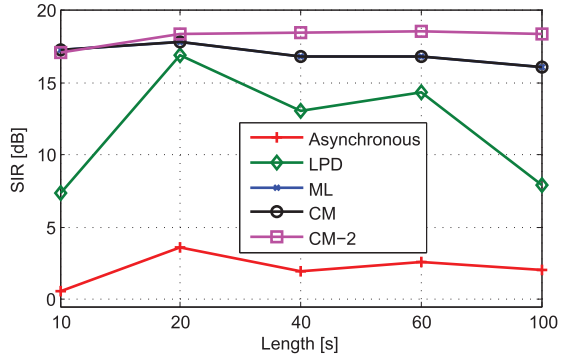


Fig. 13. BSS performance in terms of output SIR for real recorded data before (asynchronous) and after sampling rate compensation by the considered algorithms. The SRO between the devices is about 5.5 Hz at a sampling rate of 16 kHz. Note that the SIR curves of CM and ML are overlapping.

Fig. 13 depicts the BSS performance after sampling rate compensation with the considered algorithms for different signal lengths varying from 10 s to 100 s. All algorithms improve the output SIR with respect to the asynchronously sampled signals. LPD yields the worst performance, and CM and ML achieve a very similar SRO estimation accuracy, hence having practically overlapping SIR curves. CM-2 yields a similar performance as CM and ML for a signal length of 10 s but outperforms CM and ML for longer signals. It can be expected that the advantage of CM-2 becomes even more evident when the SRO between two devices is larger.

G. Computational Complexity

The computational complexity of the ML, CM and CM-2 algorithms is dominated by the exhaustive search procedure, whose computational complexity is proportional to KLQ, where Q is the number of SRO candidates, which depends on the search region and the search step. For the same search region and search step, the computational complexity of ML and CM is similar, whereas the computational complexity of CM-2 is twice of CM. In contrast, LPD computes the SRO analytically and does not require an exhaustive search procedure. The computational complexity of LPD is proportional to KL, which is much smaller than the complexity of the other algorithms.

As an example, we have compared the computation time of the Matlab implementations of all considered algorithms on an Intel CPU i5@3.33 GHz with 4 GB RAM, using a 20 s long signal at a sampling rate of 16 kHz. The computation time of LPD is only 0.32 s. For the ML, CM and CM-2 algorithms,

we have used a search region $[-10, 10]$ Hz and a search step 0.0005 Hz, resulting in $Q = 4 \times 10^4$. The computation time is 1345 s (ML), 1316 s (CM) and 2633 s (CM-2), i.e., the computation time for each SRO candidate is around 0.03 s. In practice, the computational complexity of the exhaustive search procedure can be significantly reduced (to tens of seconds) by using a coarse-to-fine search scheme, or by using an intelligent search scheme (e.g., the golden section search scheme [24] and the iterative quadric function approximation scheme [28]), or by reducing the search region based on prior knowledge of the recording devices. In addition, since the SRO usually does not vary a lot over time, the SRO value does not need to be updated frequently once it has been estimated.

VI. Conclusions

In this paper, we have proposed a correlation maximization based algorithm to blindly estimate the SRO between two asynchronously sampled microphone signals[1]. By approximating the SRO with a linear-phase drift model in the STFT domain, the proposed algorithm estimates the SRO as the one maximizing the correlation coefficient between two resynchronized signals. Furthermore, since the precision of the linear-phase drift model degrades for large SROs and for large signal lengths, a two-stage search scheme is proposed to minimize the influence of the model errors.

Simulated and real-world experiments validate the performance of the proposed correlation maximization algorithm. The algorithm works well for static acoustic environments as long as the correlation between the microphone signals is large enough. The algorithm also shows promising results for dynamic acoustic environments for the special case when the source is moving around the microphones and when the signal is long enough.

References

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

[2] S. Makino, T. W. Lee, and H. Sawada, Eds. *Blind Speech Separation*, Berlin, Germany: Springer-Verlag, 2007.

[3] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[4] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[5] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, pp. 1–13, 2010, Article ID 797962.

[6] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 549–557, Mar. 2011.

[7] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[8] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," *Proc. IEEE Symp. Commun. Veh, Technol. Benelux (SCVT)*, Ghent, Belgium, 2011, pp. 1–6.

[9] S. Haykin and K. J. R. Liu, Eds. *Handbook on Array Processing and Sensor Networks*, Hoboken, NJ, USA: Wiley, 2010.

[10] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Aachen, Germany, 2012, pp. 1–4.

[11] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 38–51, Jan. 2009.

[12] T. C. Lawin-Ore and S. Doclo, "Analysis of rate constraints for MWF-based noise reduction in acoustic sensor networks," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 269–272.

[13] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, China: 2003, pp. 840–843.

[14] E. Robledo-Arnuncio, T. S. Wada, and B. H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New York, NY, USA, 2007, pp. 34–37.

[15] H. Ding and D. I. Havelock, "Drift-compensated adaptive filtering for improving speech intelligibility in cases with asynchronous inputs," *EURASIP J. Adv. Signal Process.*, pp. 1–12, 2010, Article ID 621064.

[16] M. Pawig, G. Enzner, and P. Vary, "Adaptive sampling rate correction for acoustic echo control in voice-over-IP," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 189–199, Jan. 2010.

[17] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 70–83, Jan. 2005.

[18] T. K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1623–1636, Oct. 2015.

[19] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of Ad-hoc arrays using time difference of arrivals," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 1018–1033, Feb. 2016.

[20] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Aachen, Germany, 2012, pp. 1–4.

[21] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks - part I: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, Oct. 2010.

[22] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 233–246, Jan. 2012.

[23] L. R. Rabiner, *Multirate Digital Signal Processing*, Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[24] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185–196, Feb. 2015.

[25] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed adhoc audio networks and its use for blind source separation," *Proc. IEEE Int. Symp. Multimedia Software Eng.*, Miami, FL, USA, 2004, pp. 18–25.

[26] J. Schmalenstroeer and R. Haeb-Umbach, "Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model," *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013, pp. 1–5.

[27] Q. M. Chaudhari, "A simple and robust clock synchronization scheme," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 328–332, Feb. 2012.

[28] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Seattle, WA, USA, 2008, pp. 1–4.

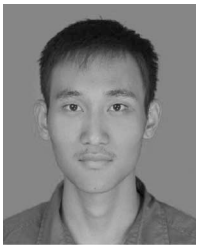
[29] L. Erup, F. M. Gardner, and R. A. Harris, "Interpolation in digital modems, II- implementation and performance," *IEEE Trans. Commun.*, vol. 41, no. 6, pp. 998–1008, Jun. 1993.

[30] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[31] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[32] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[1] Matlab code available at https://sites.google.com/site/linwangsig/sro.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003, and the Ph.D. degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow in the University of Oldenburg, Germany. Since 2014, he has been a Postdoctoral Researcher with Queen Mary University of London, UK. His research interests include video and audio compression, microphone array, blind source separation, and 3D audio processing. (https://sites.google.com/site/linwangsig).

Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation - Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Adaptive Systems Laboratory (McMaster University, Canada). From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009 he is a Full Professor at the University of Oldenburg, Germany, and scientific advisor for the project group Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997 (with Erik De Clippel), the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen) and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). He was member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2008-2013) and Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. Prof. Doclo has served as guest editor for several special issues (*IEEE Signal Processing Magazine*, *Elsevier Signal Processing*) and is associate editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*.