ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096992

DEREVERBERATION IN ACOUSTIC SENSOR NETWORKS USING WEIGHTED PREDICTION ERROR WITH MICROPHONE-DEPENDENT PREDICTION DELAYS

Anselm Lohmann[1]*, Toon van Waterschoot*[2]*, Joerg Bitzer*[3]*, Simon Doclo*[1]

[1]University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
[2]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Leuven, Belgium
[3]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
anselm.lohmann@uni-oldenburg.de

ABSTRACT

In the last decades several multi-microphone speech dereverberation algorithms have been proposed, among which the weighted prediction error (WPE) algorithm. In the WPE algorithm, a prediction delay is required to reduce the correlation between the prediction signals and the direct component in the reference microphone signal. In compact arrays with closely-spaced microphones, the prediction delay is often chosen microphone-independent. In acoustic sensor networks with spatially distributed microphones, large time-differences-of-arrival (TDOAs) of the speech source between the reference microphone and other microphones typically occur. Hence, when using a microphone-independent prediction delay the reference and prediction signals may still be significantly correlated, leading to distortion in the dereverberated output signal. In order to decorrelate the signals, in this paper we propose to apply TDOA compensation with respect to the reference microphone, resulting in microphone-dependent prediction delays for the WPE algorithm. We consider both optimal TDOA compensation using crossband filtering in the short-time Fourier transform domain as well as band-to-band and integer delay approximations. Simulation results for different reverberation times using oracle as well as estimated TDOAs clearly show the benefit of using microphone-dependent prediction delays.

***Index Terms*—** Dereverberation, weighted prediction error, acoustic sensor networks, prediction delay

1. INTRODUCTION

When recording a speech source using microphones inside a room, reverberation due to acoustic reflections may degrade the quality and intelligibility of the recorded speech. While early reflections may be beneficial, late reverberation typically reduces both speech intelligibility as well as automatic speech recognition performance [1, 2]. Hence, effective dereverberation is required for many speech communication applications, such as voice-controlled systems, hearing aids and hands-free telephony [3–14]. A popular blind dereverberation algorithm is the weighted prediction error (WPE) algorithm [10–14], which is based on multi-channel linear prediction (MCLP). WPE performs dereverberation by first estimating the late reverberant component in a reference microphone from the delayed reverberant microphone signals and then subtracting the estimate from the reference microphone signal. Several variants of the WPE algorithm have been proposed, e.g., aiming at controlling sparsity of the dereverberated output signal in the time-frequency domain [11, 13].

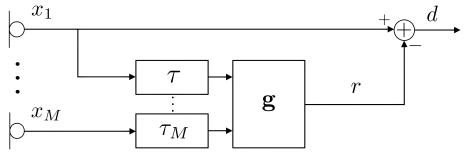


Fig. 1: WPE with microphone-dependent prediction delays

The delay, called prediction delay, plays an important role in WPE and is introduced to reduce the correlation between the prediction signals and the direct component in the reference microphone signal, hence aiming to preserve the direct component and early reflections [10, 11]. The prediction delay determines the trade-off between the amount of residual reverberation and the distortion in the desired component. It is typically chosen according to the autocorrelation of speech and is therefore often chosen microphone-independent. Whereas this may be a good choice for compact arrays with closely-spaced microphones, in acoustic sensor networks with spatially distributed microphones large inter-microphone distances may exist, leading to large and diverse time-differences-of-arrival (TDOAs) of the speech source between microphones. If uncompensated for, the effective prediction delay in each microphone may be sub-optimal, leading to distortion or excess reverberation in the desired component.

Aiming at performing TDOA compensation with respect to the reference microphone in the short-time Fourier transform (STFT)-domain, in this paper we propose to use microphone-dependent prediction delays in the WPE algorithm. We perform either optimal TDOA compensation with non-integer delays or coarse TDOA compensation with integer delays. In the STFT-domain, optimal TDOA compensation needs to be implemented using crossband filtering. However, since crossband filters introduce additional computational complexity and using more crossband filters does not necessarily imply lower reconstruction error in subbands [15], we also propose using a band-to-band approximation. Simulation results for an acoustic sensor network with spatially distributed microphones using different reverberation times as well as estimated and oracle TDOAs show the proposed microphone-dependent prediction delays outperform the microphone-independent prediction delay in the WPE algorithm, where the best dereverberation performance is achieved using non-integer delays with crossband filtering closely followed by the computationally less complex band-to-band approximation.

2. SIGNAL MODEL

We consider a single speech source captured by a set of M microphones in a reverberant room. Similarly as in [10, 11, 13], we don't consider

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956369 and the ERC Consolidator Grant: SONORA (no. 773268), and from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2177/1 - Project ID 390895286.

additive noise in this paper. In the STFT-domain, let $s(k,n)$ denote the clean speech signal with $k \in \{1,...,K\}$ the subband index and $n \in \{1,...,N\}$ the time frame index. The reverberant signal at the m-th microphone $x_m(k,n)$ can be written as

$$x_m(k,n) = \sum_{l=0}^{L_h-1} h_m(k,l)s(k,n-l) + e_m(k,n), \tag{1}$$

where $h_m(k,n)$ denotes the subband convolutive transfer function with length L_h between the speech source and the m-th microphone, and $e_m(k,n)$ denotes the subband modelling error. Without loss of generality, we define the first microphone as the reference microphone. Assuming the error term $e_m(k,n)$ in (1) can be disregarded, the dereverberation problem as is depicted in Fig. 1 can be formulated as

$$d(k,n) = x_1(k,n) - r(k,n). \tag{2}$$

The desired component $d(k,n) = \sum_{l=0}^{L_d-1} h_1(k,l)s(k,n-l)$ consists of the direct path and early reflections in the reference microphone signal $x_1(k,n)$, where L_d denotes the temporal cut-off between early and late reflections. The undesired component $r(k,n) = \sum_{l=L_d}^{L_h-1} h_1(k,l)s(k,n-l)$, which we aim to estimate, is the late reverberant component in the reference microphone signal $x_1(k,n)$. Using the MCLP model [10], the undesired late reverberant component $r(k,n)$ can be written as the sum of delayed filtered versions of all reverberant microphone signals, i.e.

$$r(k,n) = \sum_{m=1}^{M} \sum_{l=0}^{\tilde{L}_g-1} \tilde{g}_m(k,l)x_m(k,n-\tau-l), \tag{3}$$

where $\tilde{g}_m(k,n)$ denotes the m-th prediction filter of length $\tilde{L}_g$ and $\tau = L_d$ denotes the (microphone-independent) prediction delay.
The prediction delay aims at reducing the correlation between the prediction signal $x_m(k,n-\tau)$ and the desired component $d(k,n)$. It is typically chosen according to the autocorrelation of speech and therefore often microphone-independent. In this paper we propose to generalise the MCLP model in (3) to allow for a microphone-dependent prediction delay τ_m (see Fig. 1), i.e.

$$\boxed{r(k,n) = \sum_{m=1}^{M} \sum_{l=0}^{L_g-1} g_m(k,l)x_m(k,n-\tau_m-l),} \tag{4}$$

where $g_m(k,l)$ denotes the m-th prediction filter of length L_g. Using (4), the signal model in (2) can be rewritten in vector notation as

$$\mathbf{d}(k) = \mathbf{x}_1(k) - \mathbf{X}_{\boldsymbol{\tau}}(k)\mathbf{g}(k), \tag{5}$$

with

$$\mathbf{d}(k) = \begin{bmatrix} d(k,1) & \cdots & d(k,N) \end{bmatrix}^T \in \mathbb{C}^N, \tag{6}$$

$$\mathbf{x}_1(k) = \begin{bmatrix} x_1(k,1) & \cdots & x_1(k,N) \end{bmatrix}^T \in \mathbb{C}^N, \tag{7}$$

where N denotes the number of time frames. The multi-channel delayed convolution matrix $\mathbf{X}_{\boldsymbol{\tau}}(k)$ in (5) is defined as

$$\mathbf{X}_{\boldsymbol{\tau}}(k) = \begin{bmatrix} \mathbf{X}_{\tau_1}(k) & \cdots & \mathbf{X}_{\tau_M}(k) \end{bmatrix} \in \mathbb{C}^{N \times ML_g}, \tag{8}$$

where $\mathbf{X}_{\tau_m}(k) \in \mathbb{C}^{N \times L_g}$ is the convolution matrix of $\mathbf{x}_m(k)$ delayed by τ_m frames with $\tau_1 = \tau$ denoting the prediction delay in the reference microphone and $\mathbf{g}(k) \in \mathbb{C}^{ML_g}$ is the stacked vector of all prediction filter coefficients $g_m(k,n)$. The problem of speech dereverberation, i.e. estimation of the desired component $\mathbf{d}(k)$, is now reduced to estimating the filter $\mathbf{g}(k)$ predicting the undesired late reverberation.

3. WPE ALGORITHM

To estimate the prediction filter $\mathbf{g}(k)$, it has been proposed in [10] to model the desired component $d(k,n)$ using a time-varying Gaussian (TVG) model. This is equivalent to modelling the desired component in each time-frequency bin by means of a zero-mean circular Gaussian distribution with time-varying variance $\lambda(k,n)$, i.e.

$$\mathcal{N}_{\mathbb{C}}(d(k,n);0,\lambda(k,n)) = \frac{1}{\pi\lambda(k,n)} e^{-\frac{|d(k,n)|^2}{\lambda(k,n)}}, \tag{9}$$

where the variance $\lambda(k,n)$ is an unknown parameter which needs to be estimated. Since the TVG model does not assume any dependency across frequencies or time frames, the likelihood function is given by

$$\mathcal{L}(\mathbf{g}(k),\boldsymbol{\lambda}(k)) = \prod_{n=1}^{N} \mathcal{N}_{\mathbb{C}}(d(k,n);0,\lambda(k,n)), \tag{10}$$

with

$$\boldsymbol{\lambda}(k) = \begin{bmatrix} \lambda(k,1) & \cdots & \lambda(k,N) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^N. \tag{11}$$

The prediction filter $\mathbf{g}(k)$ and the variances $\boldsymbol{\lambda}(k)$ can then be estimated by maximising the log-likelihood function, which is equivalent to solving the following optimisation problem [10, 11]

$$\min_{\boldsymbol{\lambda}(k),\mathbf{g}(k)} \sum_{n=1}^{N} \left(\frac{|d(k,n)|^2}{\lambda(k,n)} + \log\pi\lambda(k,n) \right). \tag{12}$$

Since no closed-form solution exists for the joint optimisation problem in (12), it has been proposed in [10] to use an alternating optimisation procedure, where two simpler sub-problems are solved in an iterative manner. More in particular, in each iteration the cost function in (12) is first minimized with respect to the prediction vector $\mathbf{g}(k)$, assuming that the variances $\boldsymbol{\lambda}(k)$ are fixed to the values from the previous iteration. Using these values for the prediction vector $\mathbf{g}(k)$, the cost function in (12) is then minimized with respect to the variances $\boldsymbol{\lambda}(k)$. At a given iteration i, with index k omitted, the estimated prediction filter and the estimated variances are given by

$$\hat{\mathbf{g}}^{(i+1)} = \left(\mathbf{X}_{\boldsymbol{\tau}}^H \mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}}^{-1} \mathbf{X}_{\boldsymbol{\tau}} \right)^{-1} \mathbf{X}_{\boldsymbol{\tau}}^H \mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}}^{-1} \mathbf{x}_1, \tag{13}$$

$$\hat{\boldsymbol{\lambda}}^{(i+1)} = |\hat{\mathbf{d}}^{(i+1)}|^2 = |\mathbf{x}_1 - \mathbf{X}_{\boldsymbol{\tau}}\mathbf{g}^{(i+1)}|^2, \tag{14}$$

where $\mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}} = \mathrm{diag}(\hat{\boldsymbol{\lambda}}^{(i)})$ and $\hat{\mathbf{d}}^{(i+1)}$ is the estimated desired component calculated using (5) with the $|.|$ and $(.)^2$ operators applied element-wise.

In [11], an extension to the TVG model was proposed by including a hyperprior with a sparsity-promoting parameter p, aimed at better describing the sparse nature of speech in the STFT-domain. It was shown in [11] that the solution for the variances in (14) simply changed to

$$\hat{\boldsymbol{\lambda}}^{(i+1)} = |\hat{\mathbf{d}}^{(i+1)}|^{2-p}. \tag{15}$$

4. MICROPHONE-DEPENDENT PREDICTION DELAYS

In acoustic sensor networks with spatially distributed microphones, large and diverse TDOAs of the speech source between the microphones may exist. If uncompensated for, the effective prediction delays in the microphones may be suboptimal, leading to distortion or excess reverberation in the WPE output signal (see evaluation in Section 5). In this section we propose to perform TDOA compensation using the microphone-dependent prediction delays in (4). We perform either optimal TDOA

compensation with non-integer microphone-dependent prediction delays or coarse TDOA compensation with integer microphone-dependent prediction delays. In Section 4.1 we first define the problem in the time-domain. It can be shown that the optimal TDOA compensation filter in the time-domain can be written as a combination of integer frame-delays and crossband filters in the STFT-domain (see Section 4.2). To reduce computational complexity, we also propose using a band-to-band approximation (Section 4.3) or using integer frame-delays (Section 4.4).

4.1. TDOA compensation in time-domain

Let TDOA_m be the TDOA between microphone m and the reference microphone. This TDOA can be decomposed as

$$\text{TDOA}_m = \delta_m^{\text{frame}} \times L_{\text{shift}} + \delta_m^{\text{samp}} + \delta_m^{\text{frac}}, \tag{16}$$

with the integer frame-delay δ_m^{frame}, the integer delay δ_m^{samp} and the fractional delay δ_m^{frac} defined as

$$\delta_m^{\text{frame}} = \lfloor \frac{\text{TDOA}_m}{L_{\text{shift}}} \rceil, \tag{17}$$

$$\delta_m^{\text{samp}} = \lfloor \text{TDOA}_m - \delta_m^{\text{frame}} \times L_{\text{shift}} \rfloor, \tag{18}$$

$$\delta_m^{\text{frac}} = \text{TDOA}_m - \delta_m^{\text{frame}} \times L_{\text{shift}} - \delta_m^{\text{samp}}, \tag{19}$$

where L_{shift} denotes the STFT frame shift, $\lfloor . \rceil$ denotes the operator for nearest integer rounding and $\lfloor . \rfloor$ denotes the operator for lower integer rounding. The optimal time-domain TDOA compensation filter $\underline{u}'_m[t]$ can be written as [16]

$$\underline{u}'_m[t] = \delta[t - \delta_m^{\text{frame}} \times L_{\text{shift}}] * \delta[t - \delta_m^{\text{samp}}] * \text{sinc}[t - \delta_m^{\text{frac}}], \tag{20}$$

where $\delta[t]$ denotes a unit impulse with discrete-time index t and $*$ denotes convolution. It should be noted that $\delta[t - \delta_m^{\text{frame}} \times L_{\text{shift}}]$ can be implemented using integer frame-delays in the STFT-domain, whereas $\underline{u}_m[t] = \delta[t - \delta_m^{\text{samp}}] * \text{sinc}[t - \delta_m^{\text{frac}}]$ needs to be implemented using crossband filters in the STFT-domain.

4.2. Non-integer prediction delays using crossband filtering

It was shown in [15] that any time-domain filter, e.g. $\underline{u}_m[t]$, can be implemented in the STFT-domain using crossband filters

$$u_m(k, k', n) = \{\underline{u}_m[t] * \phi_{k,k'}[t]\}|_{t=nL_{\text{shift}}}, \tag{21}$$

where $u_m(k, k', n)$ denotes the crossband filter with crossband index k' in subband k,

$$\phi_{k,k'}[t] = e^{j2\pi k' t} \sum_{m=-K}^{K} \tilde{\psi}[m]\psi[t+m]e^{-j\frac{2\pi}{K}m(k-k')}, \tag{22}$$

with $\tilde{\psi}[t]$ and $\psi[t]$ denoting the STFT analysis and synthesis windows with length K. Using the crossband filter $u_m(k, k', n)$ and the integer frame-delay δ_m^{frame}, optimal TDOA compensation may be performed in the STFT-domain, equivalent to the optimal time-domain TDOA compensation filter $\underline{u}'_m[t]$ in (20).

We now apply optimal TDOA compensation to the prediction-delayed microphone signal $x_m(k, n - \tau)$, where τ is the integer prediction delay of the reference microphone. This leads to a non-integer microphone-dependent prediction delay $\tau_m = \tau + \text{TDOA}_m / L_{\text{shift}}$ in the WPE algorithm (see Fig. 1), i.e.

$$\boxed{x_m(k, n - \tau_m) = \sum_{k'=1}^{K} \sum_{l=-L_a}^{L_c} x_m(k', n - l - \tau_m^{\text{int}}) u_m(k, k', l),} \tag{23}$$

where L_a and L_c denote the number of acausal and causal taps in $u_m(k, k', n)$, and $\tau_m^{\text{int}} = \tau + \delta_m^{\text{frame}}$ denotes the integer component of the prediction delay τ_m.

4.3. Band-to-band approximation

Since the crossband filter $u_m(k, k', n)$ in (23) is computationally complex and it has been shown in [15] that crossband filters do not necessarily lower reconstruction error in $x_m(k, n - \tau_m)$, we propose to use a band-to-band approximation of the optimal TDOA compensation in (23), i.e.

$$\boxed{x_m(k, n - \tau_m) = \sum_{l=-L_a}^{L_c} x_m(k, n - l - \tau_m^{\text{int}}) u_m(k, l),} \tag{24}$$

where $u_m(k, n) = u_m(k, k, n)$ denotes the band-to-band filter in (21).

4.4. Integer prediction delays

In order to further reduce computational complexity, we only consider the integer frame-delay component in (24) and ignore the band-to-band filter $u_m(k, n)$. This corresponds to applying coarse TDOA compensation to the prediction-delayed microphone signal $x_m(k, n - \tau)$, leading to an integer microphone-dependent prediction delay $\tau_m = \tau_m^{\text{int}}$, i.e.

$$\boxed{x_m(k, n - \tau_m) = x_m(k, n - \tau_m^{\text{int}}).} \tag{25}$$

5. EXPERIMENTAL EVALUATION

In this section we evaluate the performance of the microphone-independent prediction delay and the proposed microphone-dependent prediction delays in the WPE algorithm for an acoustic sensor network with spatially distributed microphones, both using oracle as well as estimated TDOAs. In Section 5.1 we discuss the considered acoustic setup and the performance measures. In Section 5.2 we present the simulation results for the different TDOA compensation implementations proposed in Section 4.

5.1. Acoustic setup and performance measures

We consider an acoustic sensor network with $M = 9$ spatially distributed microphones and a single speech source in a room of dimensions 8m×8m×5m. Fig. 2 depicts the position of the microphones and the considered positions of the speech source inside the room. The microphones are placed on a 3×3 grid with equal spacing of dimensions 7.5m×7.5m at a height of 1.5m. The reference microphone is chosen as the bottom-left microphone and fixed for all considered source positions. In total 48 positions of the speech source are considered on a 7×7 grid with equal spacing of dimensions 7m×7m at a height of 1.5m, with one position of the speech source removed as it overlaps with a chosen microphone position. The minimum distance between any considered position of the speech source and the closest microphone is 0.5m.

The reverberant microphone signals were generated by convolving a subset of anechoic speech from the TIMIT database [17] with room impulse responses (RIRs) that were simulated using the randomized image method [18] with reverberation time $T_{60} \in \{500, 750, 1000\}$ ms. The signals were processed at a sampling rate of 16 kHz using an STFT framework with frame size $K = 1024$ samples, frame shift $L_{\text{shift}} = 256$ samples and square-root Hann analysis and synthesis windows in (22). The WPE algorithm was implemented with prediction filter length $L_g \in \{8, 12, 16\}$ (proportional to T_{60}), prediction delay $\tau = 2$ and sparsity-promoting parameter $p = 0.5$ in (15). Hence for the considered

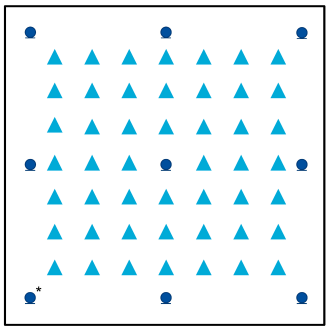


Fig. 2: Positions of $M = 9$ spatially distributed microphones ● with fixed reference microphone (*) and 48 considered positions of single speech source ▲

acoustic scenario, the possible microphone-dependent prediction delays τ_m lie in the range $\tau_m \in [0, 4]$.

To estimate the TDOAs of the speech source between the microphones, we used the popular generalised cross-correlation with phase transform (GCC-PHAT) algorithm [19] with a frame size of 2048 samples (i.e. twice as large as WPE) and a frame shift of 1024 samples, since the frame size needs to be at least twice as large as the largest possible TDOA.

Dereverberation performance is evaluated using the frequency-weighted segmental signal-to-noise-ratio (FWSSNR), the perceptual evaluation of speech quality (PESQ) and the cepstral distance (CD) measure [20]. The reference signal used in these measures is the direct component in the reference microphone. The above measures are averaged across the 48 considered positions of the speech source.

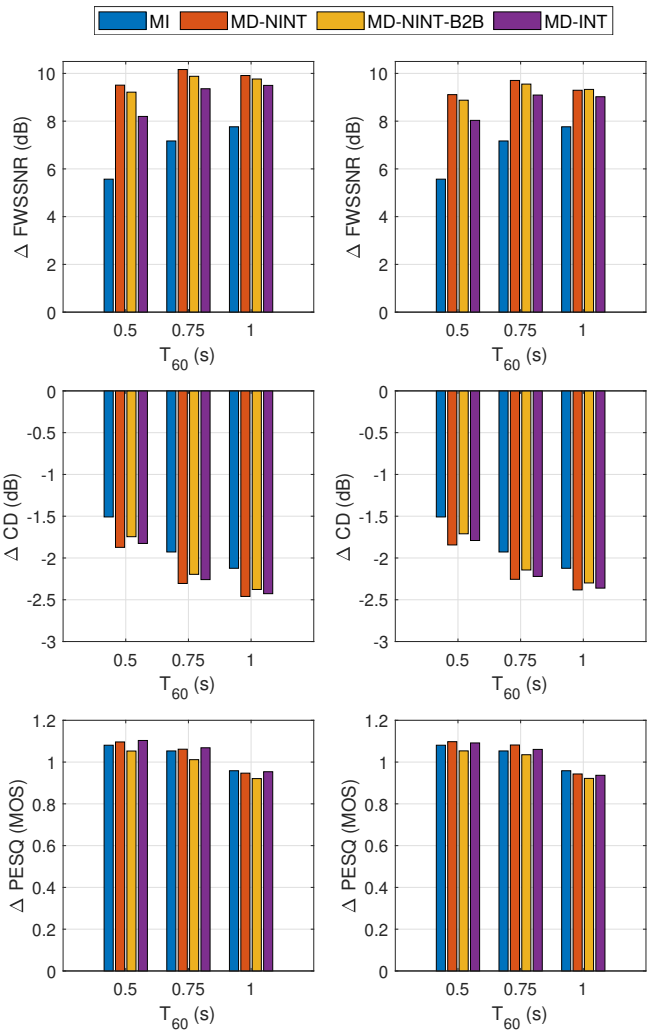
5.2. Simulation results

For the WPE algorithm, we consider the following prediction delays:

- MI: using a microphone-independent prediction delay τ
- MD-NINT: using a microphone-dependent prediction delay τ_m, performing optimal TDOA compensation using crossband filtering in (23)
- MD-NINT-B2B: using a microphone-dependent prediction delay τ_m with band-to-band approximation in (24)
- MD-INT: using a microphone-dependent integer prediction delay τ_m in (25)

For all considered WPE algorithms, Fig. 3 depicts the average performance for different reverberation times, both for oracle as well as for estimated TDOAs. First, for oracle TDOAs it can be observed for all reverberation times that in terms of ΔFWSSNR and ΔCD the performance using microphone-dependent prediction delays is significantly better than using a microphone-independent prediction delay. In terms of ΔPESQ, the performance is similar, although clear differences in speech quality are audible between using microphone-dependent prediction delays and using a microphone-independent prediction delay[1]. Comparing the different implementations of the microphone-dependent prediction delays, it can be observed that non-integer delays with crossband filtering (MD-NINT) clearly outperforms integer delays (MD-INT), while using the band-to-band approximation of non-integer delays (MD-NINT-B2B) is only marginally worse than using crossband filtering. Second, using estimated TDOAs for TDOA compensation a very similar performance can be achieved as using oracle TDOAs for all considered WPE algorithms. This is not very surprising as the estimated TDOAs were quite close to the oracle TDOAs, with an average correlation coefficient of 0.85.

[1]Audio examples available on uol.de/f/6/dept/mediphysik/ag/sigproc/audio/dereverb/mdpd.html



(a) Oracle TDOAs (b) Estimated TDOAs

Fig. 3: Average performance in terms of frequency-weighted segmental SNR improvement, CD improvement and PESQ improvement for all considered WPE algorithms for different reverberation times T_{60}, for (a) oracle TDOAs and (b) estimated TDOAs

6. CONCLUSION

In this paper we have presented a WPE algorithm with microphone-dependent prediction delays. We have proposed to use these microphone-dependent prediction delays to perform TDOA compensation, which is especially relevant in acoustic sensor networks with spatially distributed microphones. We have considered several versions to perform TDOA compensation in the STFT-domain, i.e. optimal TDOA compensation with non-integer delays using crossband filtering, approximate TDOA compensation using band-to-band filters or coarse TDOA compensation with integer-frame delays. The experimental evaluation showed that using microphone-dependent prediction delays to perform TDOA compensation improves the performance compared to a microphone-independent prediction delay, where the best dereverberation performance is achieved using non-integer delays with crossband filtering closely followed by the computationally less complex band-to-band approximation. Investigating the performance of the proposed schemes for acoustic scenarios with additive noise, a moving source or multiple sources are directions for future research.

7. REFERENCES

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] E. A. P. Habets and P. A. Naylor, "Dereverberation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, 2015.

[5] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, 2018.

[6] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 544–558, 2019.

[7] X. Li, L. Girin, S. Gannot, and R. Horaud, "Multichannel online dereverberation based on spectral magnitude inverse filtering," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1365–1377, 2019.

[8] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[9] J. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 171–175.

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

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

[12] J. Wung, A. Jukić, S. Malik, M. Souden, R. Pichevar, J. Atkins, D. Naik, and A. Acero, "Robust multichannel linear prediction for online speech dereverberation using weighted Householder least squares lattice adaptive filter," *IEEE Trans. on Signal Processing*, vol. 68, pp. 3559–3574, 2020.

[13] M. Witkowski and K. Kowalczyk, "Split Bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Processing Letters*, vol. 28, pp. 942–946, 2021.

[14] G. Huang, J. Benesty, I. Cohen, and J. Chen, "Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 1277–1289, 2022.

[15] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[16] J. Benesty, J. Chen, and Y. Huang, "Conventional beamforming techniques," in *Microphone Array Signal Processing*, J. Benesty and W. Kellermann, Eds. Springer, 2008.

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

[18] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 774–786, 2015.

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

[20] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, 2016.