Das, Sneha; Bäckström, Tom

Enhancement by postfiltering for speech and audio coding in ad-hoc sensor networks

Published in:
JASA Express Letters

DOI:
10.1121/10.0003208

Published: 15/01/2021

Document Version
Publisher's PDF, also known as Version of record

JANUARY 15 2021

Enhancement by postfiltering for speech and audio coding in *ad hoc* sensor networks

Sneha Das; Tom Bäckström

JASA Express Lett 1, 015206 (2021)
https://doi.org/10.1121/10.0003208

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Enhancement by postfiltering for speech and audio coding in *ad hoc* sensor networks

Sneha Das and Tom Bäckström
Department of Signal Processing and Acoustics, Aalto University, 02230 Espoo, Finland
sneha.das@aalto.fi, tom.backstrom@aalto.fi

Abstract: Enhancement algorithms for wireless acoustic sensor networks (WASNs) are indispensable with the increasing availability and usage of connected devices with microphones. Conventional spatial filtering approaches for enhancement in WASNs approximate quantization noise with an additive Gaussian distribution, which limits performance due to the non-linear nature of quantization noise at lower bitrates. This work proposes a postfilter for enhancement based on Bayesian statistics to obtain a multidevice signal estimate, which explicitly models the quantization noise. The experiments using perceptual signal-to-noise ratio, perceptual evaluation of speech quality, and MUSHRA (multistimulus with hidden reference and anchors) scores demonstrate that the proposed postfilter can be used to enhance signal quality in *ad hoc* sensor networks.

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0003208

Received: 19 August 2020 **Accepted:** 20 October 2020 **Published Online:** 15 January 2021

1. Introduction

The emergence of connected and portable devices like smartphones and the rising popularity of voice user-interfaces and devices equipped with microphones enable the necessary infrastructure for *ad hoc* wireless acoustic sensor networks (WASNs). The dense, *ad hoc* positioning and collaboration in a WASN leads to efficient sampling of the acoustic space, thereby gaining higher quality signal estimates compared to single-channel estimates (Bertrand, 2011). Typical applications of *ad hoc* WASNs use microphones on low-resource devices, such that we need low-complexity methods that use bandwidth efficiently to compress and transmit the acoustic signals. This involves quantization at the encoder, whereby the received signal at the decoder is usually degraded by quantization noise (Bäckström and Fischer, 2016; Bäckström and Fischer, 2017; Bäckström, 2017; Dragotti and Gastpar, 2009; Pradhan and Ramchandran, 2003).

Past works on WASN often overlook the variability in maximum capacity of sensors (Zahedi *et al.*, 2015). However, rate-constrained spatial filtering like beamforming and multichannel Wiener filtering have been used in binaural hearing aids (HAs) (Doclo *et al.*, 2009; Dragotti and Gastpar, 2009; Roy and Vetterli, 2008; Srinivasan and Den Brinker, 2009a, 2009b). A study on rate-constrained optimal beamforming showed the advantage of using spatially separated microphones in HAs, although the method assumes that the joint statistics of signals are available at the processing nodes (Roy and Vetterli, 2008). Subsequently, sub-optimal strategies for noise reduction that do not use the joint statistics at the nodes have been proposed (Amini *et al.*, 2019a,b; Doclo *et al.*, 2009; Roy and Vetterli, 2008; Srinivasan and Den Brinker, 2009a, 2009b). While the above methods are effective in reducing noise, they are either limited to or are most efficient with two nodes (HAs) only. In a recent work on multinode WASN, a linearly constrained minimum variance beamformer was used to optimize rate allocation and sensor selection over nodes, based on spatial location and frequency content (Amini *et al.*, 2019a,b; Zhang *et al.*, 2017). However, due to the dynamic nature of an *ad hoc* WASN, information about sensor distribution, location, and number of targets and interference sources may be unavailable, or their exchange between nodes further adds to the bandwidth consumption and communication complexity. Further, the above methods assume an additive quantization noise model, which is accurate only at higher bitrates. Lastly, while all the above methods are optimized on Wyner–Ziv coding, their suitability in combination with existing speech and audio coding has not been demonstrated yet. Their performance in single-channel mode can therefore not compete with conventional single-channel codecs.

In this paper, we propose a Bayesian postfilter for enhancement in *ad hoc* WASNs, which explicitly models the quantization noise within the optimization framework of the filter and can be applied on top of existing codecs with minimal modifications. Thus, the main contribution of the current work is the postfilter, which takes quantization into account through truncation while retaining the conventional assumption of additive Gaussian background noise, thereby resulting in a truncated Gaussian representation of the clean speech distribution. To evaluate the proposed methodology, we make the necessary assumptions that the devices are dominantly degraded either by background noise and reverberation or by coding noise due to quantization, and each device operates at its maximum capacity. In line with past works, we show that by distributing the total available bitrate between the two sensors, the output gain of the WASN signal estimate is

higher than the output gain of a low input SNR single sensor transmitting at full bitrate (Doclo *et al.*, 2009; Roy and Vetterli, 2008; Srinivasan and Den Brinker, 2009a,b). In addition, we present the advantages of incorporating the exact quantization noise models within the optimization framework. To focus on the effect of the postfilter on quantization noise, we apply the proposed method on the output of a codec (Bäckström *et al.*, 2018), which is specifically designed to address multidevice coding. To the best of our knowledge, this is the first time a complete WASN system has been evaluated with competitive performance also in a single-channel codec mode. Although we have not yet included models of spatial configuration of sensors, room impulse responses, or multiple sources, we show that the proposed method already yields large output gains.

2. Methodology

To focus on the novel aspects of the approach, we consider a simple WASN consisting of two devices with microphones: (1) a low-resource device A with high input SNR and (2) a high-resource device B with low input SNR, as illustrated in Fig. 1. An example application is a smartwatch that collaborates with a distant smart speaker. Let $x(k,t), n(k,t)$ be the perceptual domain representations of the speech and noise signal, respectively, at the frequency bin k and time frame t (Bäckström, 2017); the perceptual domain representations are computed by dividing the frequency domain signals by the perceptual envelope obtained from the codec (Bäckström, 2017). These signals can be approximated by zero-mean Gaussian distributions with variances σ_x^2 and σ_n^2, whereby the random variables are correspondingly $X \sim \mathcal{N}(0, \sigma_x^2), \quad N \sim \mathcal{N}(0, \sigma_n^2)$ (Kim and Shevlyakov, 2008). Under the assumption of uncorrelated, additive background noise, the noisy signal $y(k,t) = x(k,t) + n(k,t)$ is Gaussian distributed with $Y \sim \mathcal{N}(0, \sigma_y^2)$ and variance $\sigma_y^2 = \sigma_x^2 + \sigma_n^2$ (Kim and Shevlyakov, 2008). Our goal is to estimate the distribution of clean speech, conditioned over the noisy observation $P(X|Y)$, in other words, the *posterior distribution* (Särkkä, 2013). We obtain estimates for every time-frequency bin and shall omit the time and frequency subscripts in the rest of the section to aid readability. According to the Bayes rule, the posterior distribution can be written as

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)} \propto P(X)P(Y|X), \tag{1}$$

where $P(X)$ and $P(Y)$ are the *prior distributions* of the speech and observed signals and $P(Y|X)$ is the conditional likelihood. However, our quantized observation, $y_q(k,t)$, of the noisy signal gives more evidence about X; The true value of the noisy signal Y lies within the quantization bin limits, $y(k,t) \in [l(k,t), u(k,t)]$, and the lower and upper bin limits for the quantization levels in a frame $\{\mathbf{l}, \mathbf{u}\} \in \mathbb{R}^{K\times 1}$ are obtained from the observed quantized spectrum of a frame $\mathbf{y}_q \in \mathbb{R}^{K\times 1}$ (Das and Bäckström, 2018). Since the true noisy signal lies in the bounded field $l(k,t) \le Y \le u(k,t)$, we compute the summation of the likelihood over the quantization bin limits to obtain the posterior distribution of speech,

$$P(X)_{(l\le Y\le u)} \propto P(X) \int_l^u P(Y|X)dy, \tag{2}$$

where $\propto$ signifies equality up to a scaling factor. Eq. 2 can be rewritten as the difference between cumulative distributions, $P(X)_{(l\le Y\le u)} \propto P(X)(F(u) - F(l))$. The conditional likelihood can be represented as $P(Y|X) \sim \mathcal{N}(x, \sigma_n^2)$, thus resulting in the final equation for the posterior distribution,

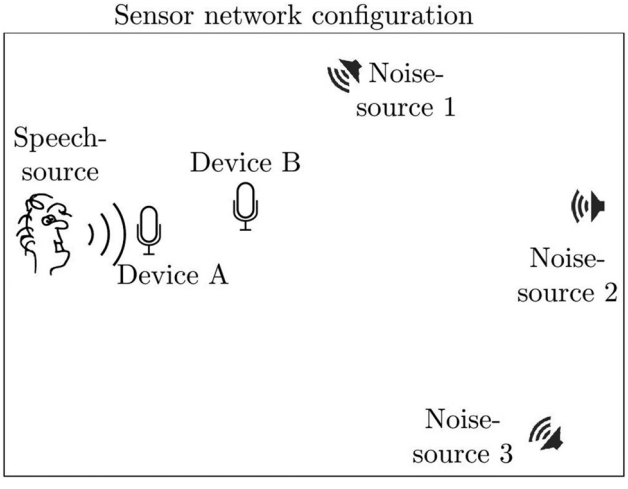

Fig. 1. Distribution of microphones in the *ad hoc* acoustic sensor network.

$$P(X)_{(l\le Y\le u)} \propto P(X)\left[0.5\left\{\operatorname{erf}\left(\frac{u-x}{\sigma_n\sqrt{2}}\right)-\operatorname{erf}\left(\frac{l-x}{\sigma_n\sqrt{2}}\right)\right\}\right], \tag{3}$$

where erf(·) is the error function. Note that due to the use of the exact quantization bin limits, $P(X)_{(l\le Y\le u)}$ corresponds to a *truncated Gaussian* (Barr and Sherrill, 1999). This is in contrast to past works, where the quantization noise is *approximated* by an additive Gaussian distribution, which is an accurate approximation only at high bitrates (Amini *et al.*, 2019a,b).

From Eq. (3), the single-channel posterior probability distribution function (PDF) of the clean speech in spatial channel i is as follows:

$$P_i(X)_{(l_i\le Y_i\le u_i)} \propto P_i(X)\left[0.5\left\{\operatorname{erf}\left(\frac{u_i-x}{\sigma_{n_i}\sqrt{2}}\right)-\operatorname{erf}\left(\frac{l_i-x}{\sigma_{n_i}\sqrt{2}}\right)\right\}\right]. \tag{4}$$

Here we assume that the speech and noise energies at each channel are estimated in a pre-processing stage, for example, using voice activity detection and minimum statistics (Martin, 2001). Additionally, to focus on the advantage of the proposed enhancement approach, we assumed that the time-delay between microphones with respect to the desired sources was known at the decoder, whereby the signals from the microphones were synchronized. We shall include time-delay estimation within the enhancement framework in future work. Based on our setup, the environmental degradation and the bitrate are different for the two channels. Hence, we can assume that $N_i \sim \mathcal{N}(\mu_{n_i}, \sigma^2_{n_i})$, and the quantization bin $\{\mathbf{l},\mathbf{u}\}_i$ offsets are uncorrelated and independent between the two channels. Therefore, when conditioned on Y, due to conditional independence between the channels, the joint posterior PDF of speech over the network can be represented as $P(X)_Y \propto \prod_{i=1}^{M} P_i(X)_{(l_i\le Y_i\le u_i)}$, where M is the number of microphones in the WASN. The posterior PDF of speech in a two microphone network is thus

$$P(X)_Y \propto \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{2}\left(x-\mu_{s_i}/\sigma_{s_i}\right)^2\right)}{8\pi\prod_{i=1}^{2}\sigma_{s_i}}\prod_{i=1}^{2}\operatorname{erf}\left(\frac{u_i-x}{\sigma_{n_i}\sqrt{2}}\right)+\prod_{i=1}^{2}\operatorname{erf}\left(\frac{l_i-x}{\sigma_{n_i}\sqrt{2}}\right)-\sum_{i=1}^{2}\left[\operatorname{erf}\left(\frac{u_i-x}{\sigma_{n_i}\sqrt{2}}\right)\operatorname{erf}\left(\frac{l_{3-i}-x}{\sigma_{n_{3-i}}\sqrt{2}}\right)\right]. \tag{5}$$

We obtain the multidevice signal estimate $\hat{x}_{\mathrm{MC}}$, optimal in minimum mean squared error (MMSE) sense (Särkkä, 2013) by computing the expectation of the PDF obtained from Eq. (5). Due to the product of error functions in Eq. (5), the expectation does not have a known analytical formulation. Therefore, we approximate the expectation of the PDF via numerical integration (McLeod, 1980); computing the Riemann sum using the midpoint rule over intervals $n = 200$ provided an approximate with sufficient accuracy in our experiments. Hence, the final equation is

$$\hat{x}_{\mathrm{MC}} = E[X_{\mathrm{MC}}] \approx \sum_{j=1}^{n} x_j P(X = x_j)_Y, \quad \min_{i=1,2}\{l_i\} \le x \le \max\{u_i\}. \tag{6}$$

The system block diagram is depicted in Figs. 2(a) and 2(b), where Fig. 2(a) is the overview of the entire system, from acoustic signal acquisition at the sensors to obtaining the time-domain estimate from multidevice signals. Note that the postfilter is placed at the fusion center, directly after the decoder, which provides the decoded perceptual domain

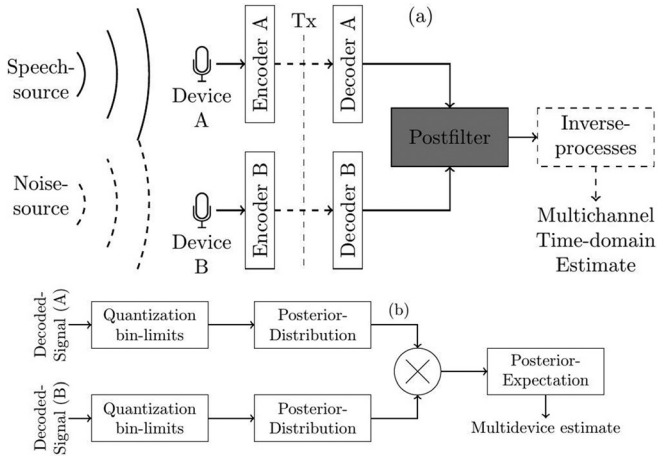

Fig. 2. Block diagrams showing the overall system structure with the location of the postfilter (a) and an overview of the postfilter (b).

signals to the postfilter. Fig. 2(b) shows the internal structure of the postfilter. After receiving the quantization bin limits from the decoded signals, we compute the truncated Gaussian distribution for each channel and then compute the joint posterior distribution as the product of the truncated distributions of the channels. The final point estimate, obtained as the expectation of the posterior distribution, yields the multidevice signal estimate.

3. Experiments and results

To evaluate the performance of the proposed postfiltering approach, we determined the perceptual signal-to-noise ratio (PSNR) and PESQ scores (Bäckström, 2017) and conducted a subjective listening test using MUSHRA (multistimulus with hidden reference and anchors) (ITU-R, 2014; Schoeffler *et al.*, 2015). We considered two categories of degradation: (1) additive background noise and (2) background noise with reverberation. For the background noises, from the QUT dataset, we extracted the cafeteria scenario with babble noise (Dean *et al.*, 2010). The clean speech samples were obtained from the test set of the TIMIT dataset (Zue *et al.*, 1990). We encoded the noisy samples and applied the proposed postfilter to the decoded samples. Hence, the output signal is corrupted by both coding and environmental artefacts. To generate noisy speech with reverberation, we considered a room of dimensions $7.5 \times 5 \times 2$ m^3, with one speech source at coordinates (1, 2.5, 0.5) m and three noise sources placed at (6.5, 2.85, 0.5) m, (3.5, 4.5, 0.5) m, and (6, 0, 0.5) m. The locations of the near and distant microphones are, respectively, (1.05, 2.55, 0.5) m and (2.25, 2.85, 0.5) m. An illustration of the setup is presented in Fig. 1. The signals at the microphones for the described acoustic scenario were simulated using Pyroomacoustics (Scheibler *et al.*, 2018).

Let ρ and γ represent the PSNR and PESQ scores, respectively, and R the total bitrate. The postfilter is applied on the output of a codec that is specifically suitable for multidevice coding (Bäckström *et al.*, 2018). For a fair evaluation, the single-channel enhancements from Eq. (6) are used as baselines. Furthermore, we employ the conventional multichannel Wiener filter (MWF) with diagonalized covariance matrix to evaluate the advantage of the proposed method with respect to a conventionally accepted baseline (Doclo and Moonen, 2002). The notations and their definitions are as follows. (1) $\hat{x}_{\text{MC}}$ is the multidevice estimate using device A at the bitrate $= \frac{1}{4}R$ and device B at the bitrate $= \frac{3}{4}R$; the PSNR and PESQ scores of the estimate are ρ_{MC} and γ_{MC}, respectively. (2) $\hat{x}_{\text{BL\_B}}$ is the baseline posterior estimate (from Eq. 6) at distant device B, encoding at full bitrate $= R$; $\rho_{\text{BL\_B}}$ and $\gamma_{\text{BL\_B}}$ are the objective measures. (3) $\hat{x}_{\text{BL\_A}}$ is the baseline posterior estimate (from Eq. 6) at device A using bitrate $= \frac{1}{4}R$, and $\rho_{\text{BL\_A}}$ and $\gamma_{\text{BL\_A}}$ are the objective measures. (4) $\hat{x}_{\text{MWF}}$ is the multichannel Wiener filter using noisy signals from device A and device B, and ρ_{MWF} and γ_{MWF} are the objective measures. We show the advantage of the proposed postfilter over the baseline methods using differential PSNR and PESQ scores; their definitions are (1) $\rho_{(\text{MC}-\text{BL\_B})} = \rho_{\text{MC}} - \rho_{\text{BL\_B}}$, (2) $\rho_{(\text{MC}-\text{BL\_A})} = \rho_{\text{MC}} - \rho_{\text{BL\_A}}$, (3) $\rho_{(\text{MC}-\text{MWF})} = \rho_{\text{MC}} - \rho_{\text{MWF}}$, (4) $\gamma_{(\text{MC}-\text{BL\_B})} = \gamma_{\text{MC}} - \gamma_{\text{BL\_B}}$, (5) $\gamma_{(\text{MC}-\text{BL\_A})} = \gamma_{\text{MC}} - \gamma_{\text{BL\_A}}$, and (6) $\gamma_{(\text{MC}-\text{MWF})} = \gamma_{\text{MC}} - \gamma_{\text{MWF}}$.

The input SNR at device A was fixed to 40 dB and at device B, we used a range of input SNRs $\in \{-5, 0, 5\ldots, 30$ dB$\}$. From the test set of the TIMIT dataset, we randomly selected 100 speech samples (50 male and 50 female) and tested the postfilter over all the combinations of the bitrates, $R \in \{16, 24, 32, 48, 64, 80, 96$ kbps$\}$, and the input SNRs for each speech sample. The objective results for the additive noise scenario are presented in Figs. 3(a)

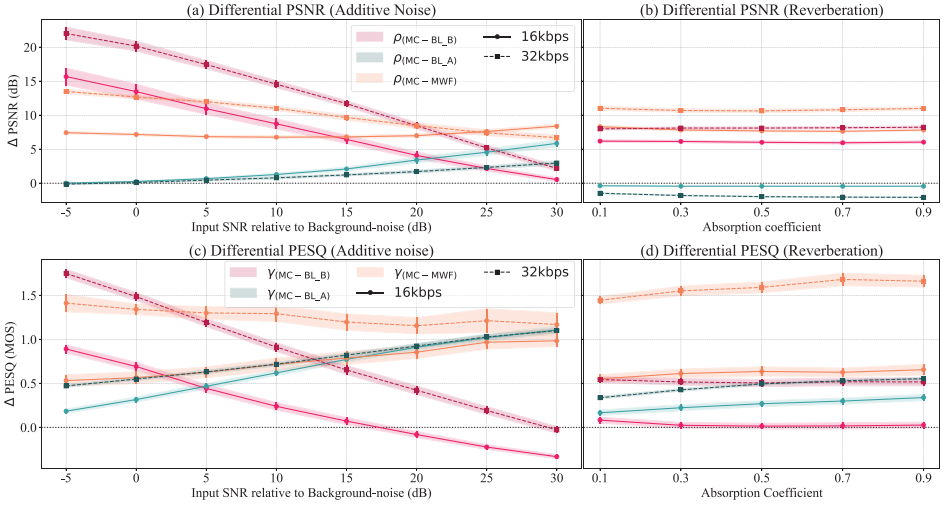

Fig. 3. Illustration of differential PSNR and PESQ scores between the proposed multidevice estimate and single-channel baseline and multichannel Wiener filter at $R = \{16, 32$ kbps$\}$ with 95% confidence intervals. $\rho_{(\text{MC}-\text{BL\_B})}$ and $\gamma_{(\text{MC}-\text{BL\_B})}$ are the differential PSNR and PESQ of the proposed multidevice estimate with respect to single-channel estimate of device B; $\rho_{(\text{MC}-\text{BL\_A})}$ and $\gamma_{(\text{MC}-\text{BL\_A})}$ are the differential scores of the multidevice estimate with respect to single-channel estimate of device A; $\rho_{(\text{MC}-\text{MWF})}$ and $\gamma_{(\text{MC}-\text{MWF})}$ are the differential scores of the multidevice estimate with respect to the multichannel Wiener filter.

and 3(c). $\rho_{(\mathrm{MC-BL\_A})}$, $\rho_{(\mathrm{MC-BL\_B})}$, and $\rho_{(\mathrm{MC-MWF})}$ are shown in Fig. 3(a) for the listed SNRs and the total bitrate $\in \{16, 32\ \mathrm{kbps}\}$; We found that the PSNR of the proposed method was better than all three baselines over all SNRs and bitrates. For $\hat{x}_{\mathrm{MC}}$ relative to the single-channel estimate $\hat{x}_{\mathrm{BL\_B}}$, the highest differential PSNR is $\rho_{(\mathrm{MC-BL\_B})} \approx 22.5$ dB. With respect to $\hat{x}_{\mathrm{BL\_A}}$, the highest $\rho_{(\mathrm{MC-BL\_A})} \approx 6$ dB is obtained at 30 dB input SNR and 16 kilobits/s (kbps). In addition, we observe that $\rho_{(\mathrm{MC-BL\_B})}$ decreases with the increase in the input SNR at device B; also, it increases with an increase in total bitrate due to lower degradation from coding noise, specifically at device A. In contrast, $\rho_{(\mathrm{MC-BL\_A})}$ increases with an increase in the input SNR at device B but decreases with increase in the total bitrate. In terms of PESQ, the largest differential PESQ for $\hat{x}_{\mathrm{MC}}$ relative to $\hat{x}_{\mathrm{BL\_B}}$ is $\gamma_{(\mathrm{MC-BL\_B})} \approx 1.8$ Mean Opinion Score (MOS), attained at -5 dB and 32 kbps. However, at 16 kbps and above 15 dB, the negative MOS implied a decrease in quality. With respect to $\hat{x}_{\mathrm{BL\_A}}$, largest value is $\gamma_{(\mathrm{MC-BL\_B})} \approx 1.1$ MOS at 30 dB input SNR at device B. Furthermore, the variations of $\gamma_{(\mathrm{MC-BL\_A})}$ and $\gamma_{(\mathrm{MC-BL\_A})}$ relative to the input SNR and bitrate follow similar trends as differential PSNR. Without exception, we observed similar trends for all the listed bitrates. The inverse variations of the differential scores with respect to $\hat{x}_{\mathrm{BL\_A}}$ and $\hat{x}_{\mathrm{BL\_B}}$ support our expectation that the proposed postfilter optimally merges information from the two channels to obtain an enhanced multidevice estimate.

The test was repeated to include reverberation over a range of absorption coefficients, $\alpha = \{0.1, 0.3, \ldots 0.9\}$. The results for $R \in \{16, 32\ \mathrm{kbps}\}$ are illustrated in Figs. 3(b) and 3(d). While $\rho_{(\mathrm{MC-BL\_B})}$ is positive for both bitrates over all the listed absorption coefficients, $\rho_{(\mathrm{MC-BL\_A})}$ is consistently negative. One reason for this could be that while the postfilter reduces environment noise, as is reflected in the improvement with respect to $\hat{x}_{\mathrm{BL\_B}}$, it may introduce some speech distortion or be unable to completely remove reverberation due to the lack of reverberation model, which shows as a drop in the PSNR with respect to $\hat{x}_{\mathrm{BL\_A}}$. Nevertheless, both $\gamma_{(\mathrm{MC-BL\_A})}$ and $\gamma_{(\mathrm{MC-BL\_B})}$ are positive over both the bitrates and all α, and they follow similar variation trends as in the additive noise scenario. Lastly, the positive differential objective scores for both noise types with respect to the MWF indicate that the PSNR and PESQ gains of the proposed postfilter are larger than the gains obtained using the multichannel Wiener filter. This supports our informal observation that Wiener filtering is inefficient in capturing the essential features of speech signals.

The subjective MUSHRA listening test contained eight test items (four male and four female), four of which included background noise with reverberation at $\alpha = 0.3$, while the remaining items were comprised of background noise only at SNR $= 15$ dB. Each test item consisted of five test conditions and the reference clean speech signal; a hidden reference and a lower anchor, which was the 3.5 kHz low-pass version of the reference signal, $\hat{x}_{\mathrm{MC}}$, $\hat{x}_{\mathrm{BL\_B}}$, and $\hat{x}_{\mathrm{BL\_A}}$ were presented as the test conditions; total bitrate was $R = 32$ kbps. As post-screening, we retained the responses from only those subjects who rated the hidden reference at more than 90 MUSHRA points for all items. Figure 4 presents the consolidated differential MUSHRA, represented as η, from 13 participants who passed the post-screening; the boxplots show the median and interquartile range of η. The background noise with reverberation is presented in items $\{1, 2, 3, 4\}$, and the background-noise-only samples are items $\{5, 6, 7, 8\}$. Items $\{1, 2, 5, 6\}$ are female, and the rest are male. $\eta_{(\mathrm{MC-BL\_A})}$ was positive for all items, indicating that most subjects preferred $\hat{x}_{\mathrm{MC}}$ over $\hat{x}_{\mathrm{BL\_A}}$. With respect to $\hat{x}_{\mathrm{BL\_B}}$, the variations were found to be gender dependent. While the median $\eta_{(\mathrm{MC-BL\_B})}$ points were positive for most male items (mean-M), they were negative for females (mean-F). Further analysis of the samples revealed that while background noise was attenuated in the $\hat{x}_{\mathrm{MC}}$, speech distortions were introduced into the estimate, and those distortions were more prominent in the female samples. This problem could potentially be addressed by using more informative speech priors and modifying the signal model to incorporate the effects of reverberation.

To study the region of optimal performance of the postfilter, we analyzed the average $\gamma_{(\mathrm{MC-BL\_B})}$ as a function of bitrate and input SNRs and absorption coefficient α; the resulting contour plots are depicted in Fig. 5. For the additive background noise scenario, the highest gains are at higher bitrates and low input SNRs. Furthermore, the negative

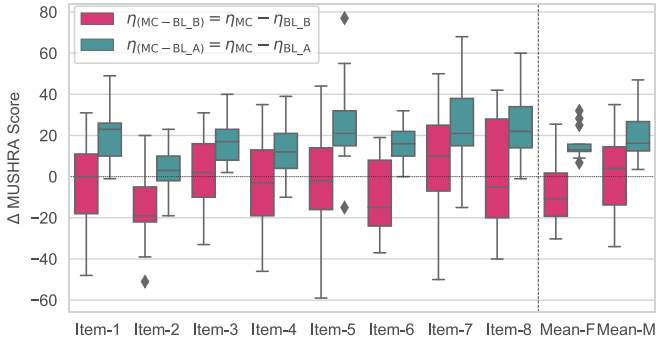

Fig. 4. Distribution of ΔMUSHRA points from the subjective listening test. $\eta_{(\mathrm{MC-BL\_B})}$ and $\eta_{(\mathrm{MC-BL\_A})}$ are the differential MUSHRA of multidevice estimate with respect to signal-channel estimates at device B and device A, respectively. Mean-F and Mean-M are the average differential scores over the female and male items, respectively.

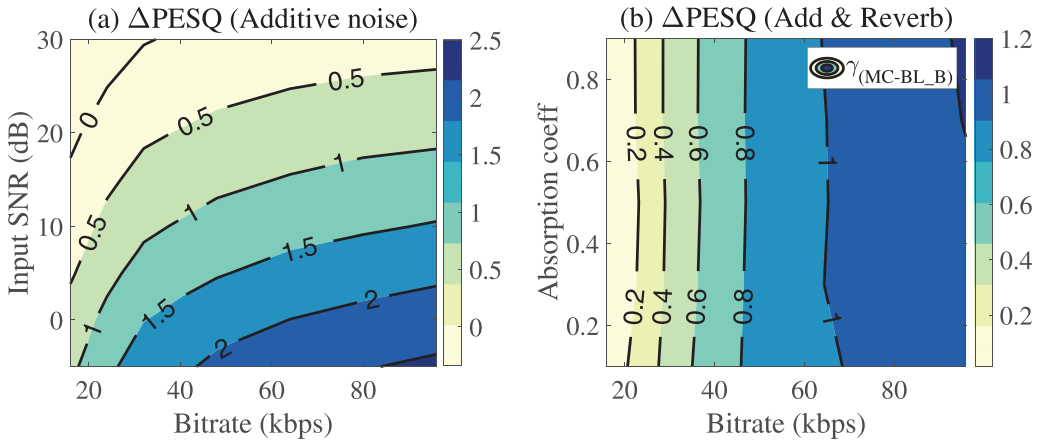

Fig. 5. Contour plot showing the differential PESQ, $\gamma_{(\mathrm{MC-BL\_B})}$ jointly over bitrates, input SNR, and absorption coefficients.

$\gamma_{(\mathrm{MC-BL\_B})}$ over 20 dB input SNR and below 32 kbps implies that the postfilter performs sub-optimally in this region; in other words, we gain from a multidevice signal estimate when the additive degradation level is below 20 dB and the total bitrate is greater than 32 kbps. In the presence of reverberation, we observed that while the total bitrate had an impact on $\gamma_{(\mathrm{MC-BL\_B})}$, the improvement was fairly constant over the range of α at an arbitrary bitrate, and the improvement was positive over the considered input SNR range. This implies that the proposed postfilter can also be used to enhance signals degraded by reverberation and is not especially sensitive to the amount of reverberation, despite the fact that the signal model did not explicitly account for distortions from reverberation.

4. Conclusion

In this work, we proposed a postfilter to enhance speech in an *ad hoc* sensor network. The method explored the feasibility of using sources degraded by two distinct noise types to obtain an enhanced estimate of the clean speech signal. We demonstrated that by distributing the total available bandwidth between two sensors, we can achieve signal quality that is higher than a single-channel estimate operating at full bitrate. Further work is needed to address the classic noise reduction vs speech distortion problem, by incorporating a signal model that takes into account the effects of reverberation, although the objective and subjective results are already encouraging.

References and links

Amini, J., Hendriks, R. C., Heusdens, R., Guo, M., and Jensen, J. (**2019a**). "Asymmetric coding for rate-constrained noise reduction in binaural hearing aids," IEEE/ACM Trans. Audio Speech Language Process. **27**(1), 154–167.

Amini, J., Hendriks, R. C., Heusdens, R., Guo, M., and Jensen, J. (**2019b**). "Rate-constrained noise reduction in wireless acoustic sensor networks," IEEE/ACM Trans. Audio Speech Language Process. **28**, 1–12.

Bäckström, T. (**2017**). *Speech Coding with Code-Excited Linear Prediction* (Springer, New York).

Bäckström, T., and Fischer, J. (**2016**). "Coding of parametric models with randomized quantization in a distributed speech and audio codec," in *Proceedings of the 12. ITG Symposium on Speech Communication*, October 5–7, Paderborn, Germany, pp. 1–5.

Bäckström, T., and Fischer, J. (**2017**). "Fast randomization for distributed low-bitrate coding of speech and audio," IEEE/ACM Trans. Audio Speech Language Process. **26**(1), 19–30.

Bäckström, T., Fischer, J., and Das, S. (**2018**). "Dithered quantization for frequency-domain speech and audio coding," in *Proceedings of Interspeech 2018*, September 2–6, Hyderabad, India, pp. 3533–3537.

Barr, D. R., and Sherrill, E. T. (**1999**). "Mean and variance of truncated normal distributions," Am. Stat. **53**(4), 357–361.

Bertrand, A. (**2011**). "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proceedings of the 2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, November 22–23, Ghent, Belgium, pp. 1–6.

Das, S., and Bäckström, T. (**2018**). "Postfiltering with complex spectral correlations for speech and audio coding," in *Proceedings of Interspeech 2018*, September 2–6, Hyderabad, India, pp. 3538–3542.

Dean, D. B., Sridharan, S., Vogt, R. J., and Mason, M. W. (**2010**). "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proceedings of Interspeech 2010*, September 26–30, 2010, Makuhari, Japan, pp. 3110–3113.

Doclo, S., and Moonen, M. (**2002**). "GSVD-based optimal filtering for single and multimicrophone speech enhancement," IEEE Trans. Signal Process. **50**(9), 2230–2244.

Doclo, S., Moonen, M., Van den Bogaert, T., and Wouters, J. (**2009**). "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," IEEE Trans. Audio Speech Language Process. **17**(1), 38–51.

Dragotti, P. L., and Gastpar, M. (**2009**). *Distributed Source Coding: Theory, Algorithms and Applications* (Academic Press, New York).

ITU-R (**2014**). "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union Radiocommunication Assembly.
Kim, K., and Shevlyakov, G. (**2008**). "Why Gaussianity?," IEEE Signal Process. Mag. **25**(2), 102–113.
Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process. **9**(5), 504–512.
McLeod, R. M. (**1980**). *The Generalized Riemann Integral* (Mathematical Association of America, Washington, DC).
Pradhan, S. S., and Ramchandran, K. (**2003**). "Distributed source coding using syndromes (DISCUS): Design and construction," IEEE Trans. Inf. Theory **49**(3), 626–643.
Roy, O., and Vetterli, M. (**2008**). "Rate-constrained collaborative noise reduction for wireless hearing aids," IEEE Trans. Signal Process. **57**(2), 645–657.
Särkkä, S. (**2013**). *Bayesian Filtering and Smoothing* (Cambridge University Press, Cambridge, UK).
Scheibler, R., Bezzam, E., and Dokmanić, I. (**2018**). "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15–20, 2018, Calgary, Alberta, Canada, pp. 351–355.
Schoeffler, M., Stöter, F. R., Edler, B., and Herre, J. (**2015**). "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *Proceedings of 1st Web Audio Conference*, January 26–28, Paris, France.
Srinivasan, S., and Den Brinker, A. C. (**2009a**). "Analyzing rate-constrained beamforming schemes in wireless binaural hearing aids," in *Proceedings of 2009 17th European Signal Processing Conference*, August 24–28, Glasgow, Scotland, pp. 1854–1858.
Srinivasan, S., and Den Brinker, A. C. (**2009b**). "Rate-constrained beamforming in binaural hearing aids," EURASIP J. Adv. Signal Process. **2009**(1), 257197.
Zahedi, A., Østergaard, J., Jensen, S. H., Naylor, P., and Bech, S. (**2015**). "Coding and enhancement in wireless acoustic sensor networks," in *Proceedings of 2015 Data Compression Conference*, April 7–9, 2015, Snowbird, UT, pp. 293–302.
Zhang, J., Chepuri, S. P., Hendriks, R. C., and Heusdens, R. (**2017**). "Microphone subset selection for MVDR beamformer based noise reduction," IEEE/ACM Trans. Audio Speech Language Process. **26**(3), 550–563.
Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**(4), 351–356.