Minimum Processing Beamforming

Zahedi, Adel; Pedersen, Mikael; Østergaard, Jan; Christiansen, Thomas; Bramsløw, Lars; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2021.3053411

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Zahedi, A., Pedersen, M., Østergaard, J., Christiansen, T., Bramsløw, L., & Jensen, J. (2021). Minimum Processing Beamforming. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2710-2724. Article 9332253. https://doi.org/10.1109/TASLP.2021.3053411

Minimum Processing Beamforming

Adel Zahedi, Michael Syskind Pedersen, Jan Østergaard, *Senior Member, IEEE,* Thomas Ulrich Christiansen, Lars Bramsløw, Jesper Jensen

***Abstract*—Most of the well-known classic beamformers have resulted from optimization problems that minimize a cost function such as the mean-square error (MSE) between the noisy speech and a reference clean speech. The rationale behind these formulations involves a speech-versus-noise dichotomy, where anything branded as noise shall be suppressed as much as possible. While leading to simple closed-form solutions and reasonably practical beamformers, this rationale has its own limitations, for instance, when the ambient noise provides context and is therefore not entirely undesirable. In this paper, we offer a new rationale, where the output of the beamformer is minimally processed with respect to a certain reference signal, as long as a given performance criterion is fulfilled. We provide a case study where the performance criterion is inspired by the Speech Intelligibility Index (SII), and the processing penalty is MSE. Regarding the reference signal, we consider two cases. In the first case, the reference signal is set to the unprocessed recording from a reference microphone, giving rise to a beamformer that limits the processing of the noisy signal to a minimum necessary for fulfilling the intelligibility requirement. For the second case, the reference signal is the output of an aggressive beamformer, yielding a beamformer that essentially eliminates the noise unless the concomitant distortion of the clean speech violates the intelligibility requirement. Through simulation studies, we demonstrate some of the benefits that each of the two cases offer in relevant contexts.**

***Index Terms*—Beamforming, speech intelligibility index, multichannel Wiener filter, MVDR beamformer, optimization**

I. Introduction

The multi-channel Wiener filter (MWF) [1] together with its variations [2]–[10] arguably make up the most commonly discussed beamformers in the acoustic signal processing community. The speech distortion weighted generalization of the MWF proposed in [2] (cf. Section II-B) covers a large and popular family of beamformers including the minimum variance distortion-less response (MVDR) beamformer [11] and the standard MWF. The principle underlying the rationale for this family of beamformers is the intrinsic undesirability of noise. The ideal, therefore, is to *remove* the noise such that only clean speech is left. This rationale can be limiting, and in some setups, even unrealistic.

There are numerous real-life scenarios, where noise provides context in terms of spatial perception, ambient awareness, etc. In such cases, it is desirable to reduce noise only to the extent that ensures sufficient intelligibility for the target speech. The above-mentioned rationale is clearly not suitable for this purpose. Another typical issue with the MWF and its generalizations is significant distortions of speech as a price for high levels of noise suppression.

In this paper, we propose a new rationale that allows for more general and flexible formulations, while covering the classic rationale as a special case. The proposed rationale is based on minimizing the distance between the beamformer output and a given reference signal subject to a certain performance constraint. In particular, we make a case study, where the distance measure is based on the mean-square error (MSE) and the performance criterion is an intelligibility estimator inspired by the speech intelligibility index (SII) [12]. Depending on the choice of the reference signal, the proposed rationale can lead to ambient-preserving beamformers or aggressive noise suppressing beamformers, or simply reduce to the existing family of MWF beamformers.

It should be noted that in addition to the MWF family of beamformers, which is the main focus of this paper, alternative approaches to beamforming have been proposed. Examples include robust beamforming [13], sparsity-based beamforming [14], DNN-based beamforming [15], and echo-aware beamforming [16]. Furthermore, this work is primarily focused on beamforming for human end users, e.g. hearing assistive devices. Other applications of beamforming such as automatic speech recognition (cf. [17] and references therein) are outside the scope of this work.

The motivation for this work and potential advantages of the new rationale in comparison to the existing one can be most easily understood in its native language; i.e. in signal processing symbols. For that reason, we start from the notations and signal model, followed by an overview of the existing beamformers, and finally the proposed concept.

II. Preliminaries

A. Notation and Signal Model

We denote matrices and vectors by boldface uppercase and lowercase letters, respectively. Covariance matrices are denoted by the letter $\mathbf{C}$ followed by an appropriate subscript as for example in $\mathbf{C}_{x_k}$ for the random vector $\mathbf{x}_k$. Similarly, variances of random variables are denoted by the symbol σ^2 with an appropriate subscript. Sets and functionals are denoted by Blackboard Bold and Calligraphic symbols, respectively, as in $\mathbb{A}$ and $\mathcal{F}$. The $M \times M$ identity matrix is denoted by $\mathbf{I}_M$, and $\mathbf{e}_r$ denotes a vector which is zero everywhere except for its r^{th} component, which is unity. We use the superscript ^{H} to denote the Hermitian transpose. For complex conjugate of scalars, we use the superscript *, not to be confused with the superscript *, which we use to mark the solutions to optimization problems. We denote the statistical expectation operation by $E[\cdot]$.

In this work, speech and noise signals are represented in the time-frequency domain. A frequency bin index k and a

This work was partially funded by Innovation Fund Denmark.

A. Zahedi and J. Jensen (email: {adza, jesj}@demant.com) are with Oticon A/S, Denamrk and the Department of Electronic Systems, Aalborg University, Denmark. J. Østergaard (email: jo@es.aau.dk) is with the Department of Electronic Systems, Aalborg University, and M. S. Pedersen, T. H. Christiansen (email: {micp,thch}@demant.com) and L. Bramsløw (email: labw@eriksholm.com) are with Oticon A/S.

time frame index l are thus needed to address a certain time-frequency tile. In this work, however, we drop the time frame index l to avoid confusing notation. It is therefore assumed by default, that we are considering a certain time frame l, unless otherwise is stated.

Denoting the number of microphones by M, without loss of generality, we arbitrarily select microphone r, $1 \leq r \leq M$ as the reference microphone. Suppose that $\mathbb{K} = \{1, ..., K\}$ is the set of all frequency bin indices. Stacking the signals acquired by all the microphones in one vector $\tilde{\mathbf{x}}_k \in \mathbb{C}^M$ for frequency bin k, we use the following speech in noise model:

$$\tilde{\mathbf{x}}_k = \tilde{s}_k \mathbf{d}_k + \tilde{\mathbf{v}}_k, \tag{1}$$

where all the variables are in general complex-valued. The M-dimensional random vectors $\tilde{\mathbf{v}}_k$ and $\tilde{\mathbf{x}}_k$ respectively represent the noise and noisy signals collected by the M microphones, and the random variable $\tilde{s}_k$ denotes the clean speech signal at the reference microphone. The M-dimensional vector $\mathbf{d}_k$ represents the relative transfer function [18] for the M microphones (with respect to the reference microphone), and its r^{th} component is therefore unity. We thus have $\mathbf{e}_r^H \mathbf{d}_k = 1$

In some applications of beamforming, e.g. in some hearing-assistive devices, the signal needs to be amplified or attenuated depending on the application. This means that the speech to be delivered to the listener's ear will be subject to an *insertion gain* g_k. Therefore, in ideal conditions, the clean speech at the output of the device is given by:

$$s_k = g_k \tilde{s}_k. \tag{2}$$

Obviously $g_k = 1$, when no gain is applied. Corresponding to (2), we define $\mathbf{x}_k \triangleq g_k \tilde{\mathbf{x}}_k$ and $\mathbf{v}_k \triangleq g_k \tilde{\mathbf{v}}_k$. Therefore, without any change in the form, (1) can be rewritten as:

$$\mathbf{x}_k = s_k \mathbf{d}_k + \mathbf{v}_k. \tag{3}$$

As common practice in the speech processing literature, we assume independence across the frequency bins, which is approximately valid, when the correlation time of the signals involved is short compared to the time-frequency analysis window size [19], [20]. Moreover, we assume that speech and noise signals are uncorrelated and zero-mean. Combining these assumptions, the covariance matrix $\mathbf{C}_{x_k}$ of $\mathbf{x}_k$ is given by:

$$\mathbf{C}_{x_k} = \mathbf{C}_{s_k} + \mathbf{C}_{v_k} = \sigma_{s_k}^2 \mathbf{d}_k \mathbf{d}_k^H + \mathbf{C}_{v_k}. \tag{4}$$

More generally, we define $\mathbf{C}_{x_k}^{(\mu)}$ as:

$$\mathbf{C}_{x_k}^{(\mu)} = \mathbf{C}_{s_k} + \mu \mathbf{C}_{v_k}, \tag{5}$$

where μ is a real-valued non-negative constant. We call $\mathbf{C}_{x_k}^{(\mu)}$ the generalized covariance matrix of $\mathbf{x}_k$.

Throughout this work, we make the common assumption that $\mathbf{C}_{v_k}$ is invertible. Consequently, we exclude the rare case, where noise is only composed of less than M point sources.[1] In addition to $\mathbf{C}_{v_k}$, we will frequently refer to $\sigma_{v_k}^2$, which is the variance of the component of $\mathbf{v}_k$ at the reference microphone.

The proposed concept heavily relies on perceptually driven performance criteria, e.g. intelligibility or quality predictors. The most well-known examples of these predictors, such as PESQ [21], STOI [22] and ESTOI [23], HASPI [24] and HASQI [25], and SII [12] and ESII [26] are defined in subbands that are deliberately defined for compliance with the human perception of sound. Critical bands, octave bands, and fractional octave bands are a few examples. On the other hand, beamformers are typically derived and analyzed in the time-frequency domain using easy-to-invert time-frequency transformations such as the short-time Fourier transform (STFT). For the sake of generality, we make a distinction between the two: For the perceptually driven subband divisions in which a certain performance criterion is defined, we use the term *subband*, while for the time-frequency tiles where the beamformer weight vector is derived/applied, we use the term *frequency bin*. The case where the two are chosen to be the same is a special case of this general framework. Depending on how the subbands and frequency bins are defined, there may be multiple frequency bins contributing to the same subband and/or multiple subbands contributing to the same frequency bin, each with certain weights. Throughout this work, we use i to index subbands, and k to index frequency bins.

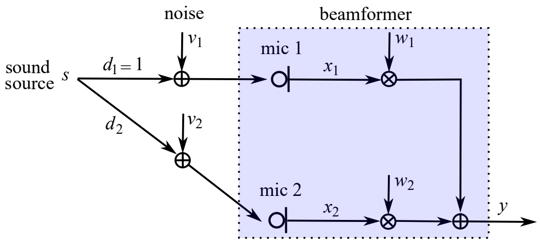

Fig. 1. A simple diagram of the signal model in a two-microphone beamforming system. The reference microphone is chosen to be microphone 1.

Suppose that we have n subbands, and $\mathbb{B}_i$ for $i = 1, ..., n$ is the set of all frequency bins k that contribute to subband i. As an example of how we use the correspondence between the subbands and frequency bins, the clean speech spectrum level for subband i is defined as:

$$P_{s_i} \triangleq \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \sigma_{s_k}^2, \tag{6}$$

where β_i is the bandwidth for subband i, and $\omega_{i,k}$ is a weight that specifies the contribution of frequency bin k to subband i (cf. Appendix A for more details).

Fig. 1 shows a simple diagram of a linear beamformer with our signal model for the special case of $M = 2$ microphones. Denoting the beamformer weight vector at frequency bin k by $\mathbf{w}_k$, the output of the beamformer is given by:

$$y_k = \mathbf{w}_k^H \mathbf{x}_k. \tag{7}$$

B. Multi-channel Wiener Filter

The standard form of MWF results from solving a minimum MSE problem which minimizes the following cost function:

$$\begin{aligned} \mathcal{MSE}(s_k, y_k) &= E\left[|s_k - y_k|^2\right] && (8) \\ &= (\mathbf{e}_r - \mathbf{w}_k)^H \mathbf{C}_{s_k} (\mathbf{e}_r - \mathbf{w}_k) + \mathbf{w}_k^H \mathbf{C}_{v_k} \mathbf{w}_k, && (9) \end{aligned}$$

[1] In practice, even in this case, the microphones add small uncorrelated noise terms, that ensure a full-rank covariance matrix.

where (9) follows from (7) and the assumption that speech and noise are uncorrelated. The solution is given by [1]:

$$\mathbf{w}_k^{\mathrm{MWF}} = \mathbf{C}_{x_k}^{-1}\mathbf{C}_{s_k}\mathbf{e}_r. \tag{10}$$

The first term on the right-hand side of (9) formulates the distortion introduced to the clean speech due to the enhancement, and the second term is the residual noise power. As seen in (9), the MSE criterion equally penalizes speech distortion and residual noise. A natural generalization of this cost function is to allow for different weights for these two terms. As proposed in [2], one such generalization is to use

$$\mathcal{MSE}_\mu(s_k, y_k) \triangleq (\mathbf{e}_r - \mathbf{w}_k)^H \mathbf{C}_{s_k}(\mathbf{e}_r - \mathbf{w}_k) + \mu \mathbf{w}_k^H \mathbf{C}_{v_k}\mathbf{w}_k, \tag{11}$$

with μ being a non-negative constant, resulting in the following generalized MWF:

$$\mathbf{w}_k^{\mu\mathrm{MWF}} = \left(\mathbf{C}_{x_k}^{(\mu)}\right)^{-1}\mathbf{C}_{s_k}\mathbf{e}_r. \tag{12}$$

It is well-known that MWF can be restated as a cascade of the MVDR beamformer and a Wiener postfilter [27]. It can be shown (cf. Appendix B), that the μMWF beamformer in (12) can similarly be restated as the cascade of the MVDR beamformer and the following generalized Wiener postfilter:

$$g_k^{(\mu)} = \frac{\xi_k}{\mu + \xi_k}, \tag{13}$$

where $\xi_k \triangleq \sigma_{s_k}^2 \mathbf{d}_k^H \mathbf{C}_{v_k}^{-1}\mathbf{d}_k$ is the SNR at the output of the MVDR beamformer. Fig. 2 shows the plot of $g_k^{(\mu)}$ as a function of ξ_k for $\mu = 1$, $\mu < 1$ and $\mu > 1$. For $\mu = 1$, it reduces to the well-known single-channel Wiener filter (SWF), leading to a beamformer that is optimal in MSE sense. For $\mu < 1$, the postfilter incurs a lower level of speech distortion compared to the standard Wiener filter at the cost of higher residual noise. In the limit when $\mu \to 0$, the μMWF beamformer reduces to the MVDR beamformer. On the contrary, $\mu > 1$ leads to an aggressive postfilter that suppresses more noise compared to the standard SWF in price of higher levels of speech distortion.

All the beamformers introduced so far are formulated with the aim of reconstructing the clean speech, i.e. complete suppression of noise as an ideal. In [4], [5], it was suggested that one may be interested in preserving a fraction of the noise in addition to the target speech, for instance to better preserve the spatial characteristics of noise in addition to the target speech. For that purpose, instead of the cost function in (8), one can minimize $\mathcal{MSE}(s_k + \alpha v_k, y_k)$ for a given positive constant α, which leads to the following solution [5]:

$$\mathbf{w}_k^{\mathrm{MWF\text{-}N}} = \mathbf{w}_k^{\mathrm{MWF}} + \alpha\mathbf{e}_r. \tag{14}$$

In effect, the MWF-N beamformer takes the output of an MWF beamformer, and adds a fraction of the unprocessed noisy speech from the reference microphone to it.

Finally, one can combine the μMWF and MWF-N beamformers to obtain the following generalized beamformer [5]:

$$\mathbf{w}_k^{\mu\mathrm{MWF\text{-}N}} = \mathbf{w}_k^{\mu\mathrm{MWF}} + \alpha\mathbf{e}_r. \tag{15}$$

This is especially useful when a large μ is chosen for the μMWF part; i.e. an aggressive beamformer with a high level of speech distortion. In this case, the resulting distortion of the clean speech can be partially compensated for by adding a fraction of the unprocessed signal to the output of the μMWF beamformer. The μMWF-N beamformer in (15) is the most general of the above-mentioned beamformers. All the other beamformers can be seen as special cases of (15) for certain choices of the parameters μ and α.

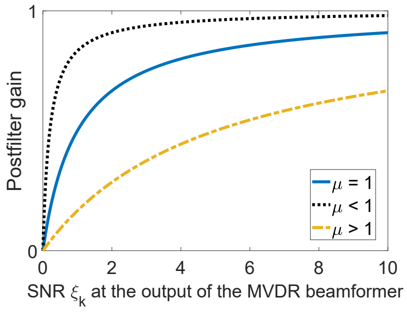

Fig. 2. Postfilter gain $g_k^{(\mu)}$ as a function of the SNR ξ_k for the μMWF beamformer with three different values of μ.

III. Minimum Processing Beamforming

A. Proposed Concept

Suppose that s_k^{R} is a given reference signal (not to be confused with the clean speech at the reference microphone). Consider a certain subband i. We stack all s_k^{R} for $k \in \mathbb{B}_i$ in a vector denoted by $\mathbf{s}_i^{\mathrm{R}}$. Similarly, we stack all y_k, s_k and v_k for $k \in \mathbb{B}_i$ into vectors $\mathbf{y}_i$, $\mathbf{s}_i$ and $\mathbf{v}_i$, respectively. Also, consider the two finite non-negative functionals $\mathcal{D}(\cdot,\cdot)$ and $\mathcal{I}(\cdot,\cdot)$. We define the minimum-processing beamformer in subband i as the solution to the following optimization problem:

$$\min_{\mathbf{w}_k, k\in\mathbb{B}_i} \mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i) \quad \text{s.t.} \quad \mathcal{I}(\mathbf{y}_i, \mathbf{s}_i) \geq I_i', \tag{16}$$

where $\mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i)$ measures the distance (processing penalty) between the reference signal and the beamformer output, $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i)$ is an estimator of performance for the beamformer output in subband i in a certain sense, e.g. speech intelligibility, sound quality, etc. The term I_i' in (16) is defined as:

$$I_i' \triangleq \min\left(I_i, I_i^{\max}\right), \tag{17}$$

where I_i is a given minimum requirement on the beamformer performance $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i)$, and $I_i^{\max}$ is the maximum achievable performance which is obtained when the processing penalty $\mathcal{D}(\mathbf{y}_i^{\mathrm{R}}, \mathbf{y}_i)$ is disregarded, and the performance $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i)$ is maximized in an unconstrained manner.

In (16), dependency of $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i)$ on the clean speech $\mathbf{s}_i$ is implied by the notation for generality. In many practical situations, performance is estimated from the beamformer output alone, and we have $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i) = \mathcal{I}(\mathbf{y}_i)$.

A special case of (16), where $\mathbf{s}_i^{\mathrm{R}} = \mathbf{s}_i + \alpha\mathbf{v}_i$, the processing penalty $\mathcal{D}$ is chosen to be the $\mathcal{MSE}_\mu$ defined in (11), and the constraint is annihilated by setting $I_i = 0$, leads to the generalized μMWF-N beamformer in (15). This demonstrates the generality of the formulation in (16). In this paper, we make a case study, where the processing penalty $\mathcal{D}$ is similar to the $\mathcal{MSE}_\mu$ criterion, and the performance criterion $\mathcal{I}(\cdot,\cdot)$ is an intelligibility estimator based on the SII [12]. We solve the problem analytically for any given reference signal s_k^{R}. Next, we study two special cases:

1) Ambient-preserving mode: In this mode of operation, the unprocessed signal from the reference microphone $\mathbf{e}_r^H \mathbf{x}_k$ is chosen as the reference signal s_k^{R}. This leads to a beamformer that attempts to retain as much of the clean speech and noise as possible by keeping the processing of the noisy speech to the minimum amount necessary for achieving the given intelligibility requirement.

2) Aggressive mode: In this mode, the reference signal s_k^{R} is the output of a reference beamformer $\mathbf{w}_k^{\mathrm{R}}$. This leads to a beamformer that inherits the (presumably desirable) properties of the reference beamformer, except for the situations, where this violates the intelligibility requirement. In particular, we study the case where the reference beamformer is the aggressive form of the μMWF beamformer.

B. Motivation

Existing research [28], [29] (as well as our experience) show that directional hearing aids in some situations tend to over-suppress the natural ambient noise, leaving the users with a feeling of isolation or exclusion. While not downplaying the crucial role of sufficient speech intelligibility, it seems reasonable that if any suppression of the ambient noise takes place, it should be limited to the minimum necessary amount that precludes any compromise of speech intelligibility. This can be formulated by setting the reference signal in (16) equal to the unprocessed signal at the reference microphone, and choosing a speech intelligibility estimator as the performance criterion $\mathcal{I}(\cdot,\cdot)$. In other words, we apply a minimum processing principle to modify the noisy signal as little as possible in order to obtain a desired level of intelligibility. This was indeed our initial motivation for this work, as evident from the title of this paper. The concept was later generalized from the noisy signal at the reference microphone to any given reference signal as in (16). An example of special interest is when the reference signal is the output of a certain beamformer $\mathbf{w}_k^{\mathrm{R}}$. This can be useful when the reference beamformer $\mathbf{w}_k^{\mathrm{R}}$, within a certain context or for a certain application, has particularly desirable properties that are compromised by pronounced drawbacks. As an example, the μMWF beamformer in (12) with aggressive noise suppression properties ($\mu \gg 1$) can effectively suppress noise at the cost of distorting speech. By choosing it as the reference beamformer in (16), while opting for a speech preserving performance criterion $\mathcal{I}(\cdot,\cdot)$, we obtain a beamformer that does an outstanding job of suppressing the noise, whenever it would not harm the speech to more than a certain extent.

IV. Theory

A. Processing Penalty

Our starting point for defining the processing penalty $\mathcal{D}(\cdot,\cdot)$ is the MSE criterion. Writing it in subbands rather than frequency bins for the sake of compatibility with the formulation in (16), it takes the following form:

$$\hat{\mathcal{D}}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i) = \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} E\left[\left|s_k^{\mathrm{R}} - y_k\right|^2\right]. \tag{18}$$

Let us define the vectors $\mathbf{r}_k$ and $\mathbf{u}_k$ as:

$$\mathbf{r}_k \triangleq E\left[\mathbf{x}_k \left(s_k^{\mathrm{R}}\right)^*\right], \tag{19}$$

$$\mathbf{u}_k \triangleq \mathbf{C}_{x_k}^{-1} \mathbf{r}_k. \tag{20}$$

Expanding the terms in (18) and subtracting and adding $\mathbf{r}_k^H \mathbf{C}_{x_k}^{-1} \mathbf{r}_k$ on the right side, we obtain:

$$\begin{aligned} \mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i) &= \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \left(\sigma_{s_k^{\mathrm{R}}}^2 - \mathbf{r}_k^H \mathbf{C}_{x_k}^{-1} \mathbf{r}_k\right) \\ &+ \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \left(\mathbf{w}_k - \mathbf{u}_k\right)^H \mathbf{C}_{x_k} \left(\mathbf{w}_k - \mathbf{u}_k\right). \end{aligned} \tag{21}$$

The first term on the right-hand side of (21) is independent of the weight vectors $\mathbf{w}_k$. It thus has no impact on the solution to the optimization problem (16). Discarding this term, and substituting $\mathbf{C}_{x_k}$ with $\mathbf{C}_{x_k}^{(\mu)}$ in (21) for more generality, we obtain the final form of our processing penalty as follows:

$$\mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i) = \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} (\mathbf{w}_k - \mathbf{u}_k)^H \mathbf{C}_{x_k}^{(\mu)} \left(\mathbf{w}_k - \mathbf{u}_k\right). \tag{22}$$

B. Performance Criterion

For the performance criterion, we use an estimation of speech intelligibility based on the SII. It is evaluated on a per-frame basis, making it similar to the ESII [26]. Assuming normal vocal effort and thus no speech level distortion, the SII is given by a weighted sum of the so-called band audibility functions over all the subbands [12]. Since (16) is defined for a certain subband, we define a band audibility constraint for each subband instead of setting one single intelligibility constraint for the entire signal. Moreover, we disregard the spectral masking effects [12] to avoid unnecessary complications, as our experience suggests that for most cases of practical interest, it has an insignificant effect on the resulting score.

With ζ_i being the speech to disturbance ratio for subband i, the audibility function $\Psi(\zeta_i)$ for subband i is given by the following function [12]:

$$\Psi(\zeta_i) = \begin{cases} 0, & \text{if } (10 \log \zeta_i) < -15 \\ 1, & \text{if } (10 \log \zeta_i) > +15 \\ \frac{10 \log \zeta_i + 15}{30}, & \text{otherwise.} \end{cases} \tag{23}$$

This function is plotted in Fig. 3. With the performance estimator chosen to be $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i) = \Psi(\zeta_i)$, the performance criterion in (16) is given by:

$$\Psi(\zeta_i) \geq I_i'. \tag{24}$$

To calculate ζ_i, we first obtain the total error power in subband i at the output of beamformers $\mathbf{w}_k, k \in \mathbb{B}_i$. This is calculated, in a manner similar to (11), as the sum of the speech distortion and noise power:

$$N_i = \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} (\mathbf{e}_r - \mathbf{w}_k)^H \mathbf{C}_{s_k} (\mathbf{e}_r - \mathbf{w}_k) + \mu \omega_{i,k} \mathbf{w}_k^H \mathbf{C}_{v_k} \mathbf{w}_k, \tag{25}$$

where normalization by bandwidth β_i is in accordance with the ANSI standard [12]. Let Λ_i denote the *equivalent internal noise level* (cf. [12]) for subband i, modelling the threshold of hearing. For normal-hearing listeners, Λ_i follows from

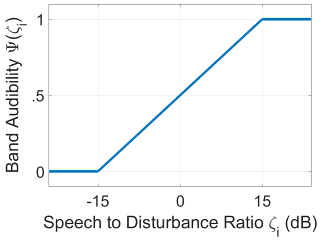

Fig. 3. ANSI recommendation for the relationship between band audibility and speech-to-disturbance ratio [12].

the threshold of hearing in quiet for the average normal-hearing person. For the hearing-impaired, the threshold must be elevated based on the individual's pure-tone audiogram. Using N_i and Λ_i, the equivalent disturbance spectrum for subband i is calculated as [12]:

$$D_i = \max\left(\Lambda_i, N_i\right). \tag{26}$$

Finally, we calculate the speech to disturbance ratio ζ_i using the following formula:

$$\zeta_i = \frac{P'_{s_i}}{D_i}, \tag{27}$$

where P'_{s_i} is defined as

$$P'_{s_i} \triangleq P_{s_i} - \Delta_i, \tag{28}$$

with P_{s_i} given in (6), and Δ_i modelling a possible loss of the clean speech power at the output of the beamformer. We elaborate on this issue in Section V-B.

The fact that the threshold of hearing Λ_i, as well as the insertion gain g_k (cf. (26) and (2), respectively) are taken into account, makes our framework suitable for hearing-impaired as well as normal-hearing users.

C. Problem Formulation and Solution

Combining the results in Sections IV-A and IV-B, the optimization problem (16) can be written as follows:

$$\begin{aligned} &\min_{\mathbf{w}_k, k\in\mathbb{B}_i} \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k} \left(\mathbf{w}_k - \mathbf{u}_k\right)^H \mathbf{C}_{x_k}^{(\mu)} \left(\mathbf{w}_k - \mathbf{u}_k\right) \\ &\text{s.t.} \quad \begin{cases} \max\left(\Lambda_i, N_i\right) \le P'_{s_i} 10^{-3\left(I'_i - \frac{1}{2}\right)}, \\ P_{s_i} 10^{-\frac{3}{2}} \le \max\left(\Lambda_i, N_i\right) \le P'_{s_i} 10^{\frac{3}{2}}, \end{cases} \end{aligned} \tag{29}$$

where the first constraint reflects the third condition in (23), and the second constraint is corresponding to the first two boundary conditions in (23). Before presenting the solution, we first need to make a number of definitions. In particular, we define the two parameters N_i^{R} and h_i as follows:

$$N_i^{\mathrm{R}} \triangleq \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k} (\mathbf{e}_r - \mathbf{u}_k)^H \mathbf{C}_{s_k} (\mathbf{e}_r - \mathbf{u}_k) + \mu\omega_{i,k} \mathbf{u}_k^H \mathbf{C}_{v_k} \mathbf{u}_k, \tag{30}$$

$$h_i \triangleq \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k} \left(\mathbf{u}_k - \mathbf{w}_k^{\mu\mathrm{MWF}}\right)^H \mathbf{C}_{x_k}^{(\mu)} \left(\mathbf{u}_k - \mathbf{w}_k^{\mu\mathrm{MWF}}\right). \tag{31}$$

We will show later, that these parameters can be interpreted depending on the choice of the reference signal. In addition, we define the two constants $I_i^{\min}$ and $I_i^{\max}$ as follows (details can be found in Appendix C):

$$I_i^{\min} \triangleq \min\left(1, \frac{1}{2} + \frac{1}{3} \max\left(-\frac{3}{2}, \log \frac{P'_{s_i}}{\max(N_i^{\mathrm{R}}, \Lambda_i)}\right)\right), \tag{32}$$

$$I_i^{\max} \triangleq \min\left(1, \frac{1}{2} + \frac{1}{3} \max\left(-\frac{3}{2}, \log \frac{P'_{s_i}}{\max(N_i^{\mathrm{R}} - h_i, \Lambda_i)}\right)\right). \tag{33}$$

Finally, we define the constant $\alpha_i^{\min}$:

$$\alpha_i^{\min} \triangleq \sqrt{\max\left(0, 1 - \frac{N_i^{\mathrm{R}} - \Lambda_i}{h_i}\right)}. \tag{34}$$

In Appendix C, we prove the following results:

1) The *minimum processing beamformer*; i.e. the solution $\mathbf{w}_{k,i}^{\mathrm{MP}}$ to (29) is given by:

$$\mathbf{w}_{k,i}^{\mathrm{MP}} = \alpha_i \mathbf{u}_k + (1 - \alpha_i) \mathbf{w}_k^{\mu\mathrm{MWF}}, \tag{35}$$

where α_i (henceforth called the combination weights) are calculated as follows: If $N_i^{\mathrm{R}} \le \Lambda_i$, then $\alpha_i = 1$; otherwise we have:

$$\alpha_i = \begin{cases} \alpha_i^{\min}, & \text{if } I_i \ge I_i^{\max} \\ 1, & \text{if } I_i \le I_i^{\min} \\ \sqrt{\max\left(0, 1 - \max\left(0, \frac{N_i^{\mathrm{R}} - P'_{s_i} 10^{-3\left(I_i - \frac{1}{2}\right)}}{h_i}\right)\right)}, & \text{otherwise.} \end{cases} \tag{36}$$

2) Maximum performance (in terms of band audibility), which is obtained by disregarding the processing penalty $\mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i)$ and maximizing $\mathcal{I}(\mathbf{y}_i, \mathbf{s}_i) = \Psi(\zeta_i)$, is given by (33).

3) Minimum performance, which is obtained by disregarding the performance constraint $\Psi(\zeta_i) \ge I'_i$ and minimizing the processing penalty $\mathcal{D}(\mathbf{s}_i^{\mathrm{R}}, \mathbf{y}_i)$, is given by (32).

Depending on the type of correspondence considered between the frequency bins and subbands, there can be overlap between the subbands; i.e., a single frequency bin can contribute to more than one subband. For that reason, we have assumed dependency both on the frequency bin index k and the subband index i in the beamformer weight vector $\mathbf{w}_{k,i}^{\mathrm{MP}}$. Let $\mathbb{F}_k$ denote the set of all subbands to which the frequency bin k contributes, and $\eta_{i,k}$ be the weight that accounts for the impact of this contribution on the beamformer weight vector. The beamformer weight vector at frequency bin k is given by:

$$\mathbf{w}_k^{\mathrm{MP}} \triangleq \sum_{i\in\mathbb{F}_k} \eta_{i,k} \mathbf{w}_{k,i}^{\mathrm{MP}}. \tag{37}$$

In Appendix A, we provide more details on the calculation of $\eta_{i,k}$ and other considerations related to the correspondence between the subbands and frequency bins.

D. Reference Signal

For the sake of case study, we confine ourselves to two choices of the reference signal with two different goals in mind. Obviously, for any other relevant scenario, one has to define the reference signal that suits the application.

1) Ambient noise preserving mode: In applications, such as hearing assistive devices, when sounds other than the target speech potentially convey useful information (e.g. traffic noise alarms, etc.) or are of interest (e.g. background music), it is desirable to preserve them fully or in part, with the criterion being an uncompromised level of intelligibility for the target speech. Setting the reference signal s_k^{R} equal to the unprocessed signal from the reference microphone $\mathbf{e}_r^H \mathbf{x}_k$ allows for this mode of operation. Substituting in (19) and the result in (20), we obtain:

$$\mathbf{u}_k = \mathbf{e}_r. \tag{38}$$

Following (35), we thus have:

$$\mathbf{w}_{k,i}^{\mathrm{MP}} = \alpha_i \mathbf{e}_r + (1-\alpha_i)\mathbf{w}_k^{\mu\mathrm{MWF}}. \tag{39}$$

This beamformer is similar to (15), with the important difference that here the coefficient α_i is signal dependent. More particularly, α_i adapts to the situation depending on how noisy the speech is in the given time frame and subband, cf. (36).

Substituting (38) and (39) in (30), we have:

$$N_i^{\mathrm{R}} = \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k}\left(\mu\sigma_{v_k}^2\right). \tag{40}$$

In other words, N_i^{R} is the noise power in subband i. Similarly, substituting (38) and (39) in (31), and using (12), we obtain:

$$h_i = \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k}\mathbf{e}_r^H \left(\mu\mathbf{C}_{v_k}\right)^H \left(\mathbf{C}_{x_k}^{(\mu)}\right)^{-1} \left(\mu\mathbf{C}_{v_k}\right)\mathbf{e}_r. \tag{41}$$

Using (5), applying the Sherman-Morrison formula [30], and simplifying the result, (41) reduces to the following:

$$h_i = N_i^{\mathrm{R}} - \frac{1}{\beta_i} \sum_{k\in\mathbb{B}_i} \omega_{i,k}\mu\sigma_{o,v_k}^2 g_k^{(\mu)}, \tag{42}$$

where $g_k^{(\mu)}$ is the generalized Wiener postfilter given by (13), and $\sigma_{o,v_k}^2 \triangleq \mathbf{d}_k^H \mathbf{C}_{v_k}^{-1}\mathbf{d}_k$ is the noise variance at the output of the MVDR beamformer.

2) Aggressive mode: This mode of operation is suitable for circumstances, where maximum suppression of noise is desired, without severely damaging the target speech. The reference signal is chosen to be the output of a reference beamformer $\mathbf{w}_k^{\mathrm{R}}$. We thus have $s_k^{\mathrm{R}} = \left(\mathbf{w}_k^{\mathrm{R}}\right)^H \mathbf{x}_k$. Substituting in (19) and the result in (20), we obtain:

$$\mathbf{u}_k = \mathbf{w}_k^{\mathrm{R}}. \tag{43}$$

Consequently, (35) takes the following form:

$$\mathbf{w}_{k,i}^{\mathrm{MP}} = \alpha_i \mathbf{w}_k^{\mathrm{R}} + (1-\alpha_i)\mathbf{w}_k^{\mu\mathrm{MWF}}. \tag{44}$$

One viable choice of the reference beamformer is the μMWF beamformer (12) with $\mu \gg 1$. This beamformer can do an outstanding job of suppressing the noise, but at the same time, it significantly distorts the target speech. In time frames and subbands where the SNR is not particularly high, these distortions will be very severe, giving rise to an overall output speech that is more audibly distorted than desired. We attempt to obtain a performance as close as possible to the μMWF beamformer (with $\mu \gg 1$) in terms of noise suppression by choosing it as the reference beamformer. On the other hand, for the second term on the right-hand side of (44), we set $\mu \ll 1$ to obtain a speech-preserving beamformer that precludes excessive distortions of speech in unfavourable conditions. This yields:

$$\mathbf{w}_{k,i}^{\mathrm{MP}} = \alpha_i \mathbf{w}_k^{\mu_1\mathrm{MWF}} + (1-\alpha_i)\mathbf{w}_k^{\mu_2\mathrm{MWF}}, \tag{45}$$

where $\mu_1 \gg 1$ and $\mu_2 \ll 1$.

Next, we calculate N_i^{R} and h_i for the present case. Substituting (43) in (30) yields:

$$\begin{aligned} N_i^{\mathrm{R}} &= \frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i} \omega_{i,k}\left(\mathbf{e}_r - \mathbf{w}_k^{\mathrm{R}}\right)^H \mathbf{C}_{s_k}\left(\mathbf{e}_r - \mathbf{w}_k^{\mathrm{R}}\right) + \mu\omega_{i,k}\left(\mathbf{w}_k^{\mathrm{R}}\right)^H \mathbf{C}_{v_k}\mathbf{w}_k^{\mathrm{R}} \\ &= N_{s,i}^{\mathrm{R}} + \mu N_{v,i}^{\mathrm{R}}. \end{aligned} \tag{46}$$

It thus becomes clear that N_i^{R} is the total error at the output of the reference beamformer in subband i, and can be written as the sum of the noise power $\mu N_{v,i}^{\mathrm{R}}$ and speech distortion $N_{s,i}^{\mathrm{R}}$ at the output of the reference beamformer. To calculate h_i using (31), we rewrite the two μMWF beamformers in (45) as the series of the MVDR beamformer and a generalized Wiener postfilter to obtain:

$$\begin{aligned} h_i &= \frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i} \omega_{i,k}\left(g_k^{(\mu_1)} - g_k^{(\mu_2)}\right)^2 \left(\mathbf{w}_k^{\mathrm{MVDR}}\right)^H \mathbf{C}_{x_k}^{(\mu_2)} \mathbf{w}_k^{\mathrm{MVDR}} \\ &= \frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i} \omega_{i,k}\left(g_k^{(\mu_1)} - g_k^{(\mu_2)}\right)^2 \left(\sigma_{s_k}^2 + \mu_2\sigma_{o,v_k}^2\right), \end{aligned} \tag{47}$$

where (47) follows from $\mathbf{C}_{x_k}^{(\mu_2)} = \sigma_{s_k}^2 \mathbf{d}_k\mathbf{d}_k^H + \mu_2\mathbf{C}_{v_k}$ and $\mathbf{w}_k^{\mathrm{MVDR}} = \mathbf{C}_{v_k}^{-1}\mathbf{d}_k/\sigma_{o,v_k}^2$.

V. Practical Considerations

There are practical matters that are crucial for optimal operation of the proposed beamformers in real-life scenarios. In this section, we address these considerations.

A. Time Averaging for Combination Weights

The value of α_i given by (36) can change abruptly across the time frames, leading to audible distortions of the speech. To avoid this, we perform a recursive averaging of α_i across the time frames as follows:

$$\bar{\alpha}_i(l) = (1-b)\ \bar{\alpha}_i(l-1) + b\,\alpha_i(l), \tag{48}$$

where l and $l-1$ index the current and previous time frames, respectively, and b is calculated from a time constant τ using the following formula:

$$b = 1 - e^{-\frac{1}{R\tau}}, \tag{49}$$

where R is the frame rate.

B. Target Loss Effects

Applying a beamformer to a noisy signal $\mathbf{x}_k$ generally results in a suppression of the target signal s_k at the output, i.e., a target loss. Formulation of the target loss requires a model for the speech distortion that is introduced by the beamformer. The simplest model is the additive noise model, i.e. speech distortion treated as additive noise uncorrelated with both speech and noise. With the additive noise model, the target loss Δ_i in (28) is zero, and speech distortion is accounted for by adding it to the residual noise power as in (25). An alternative is to subtract the speech distortion from the clean speech power in addition to treating it as residual noise power. In this case, we have:

$$\Delta_i = \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \left(\mathbf{e}_r - \mathbf{w}_k\right)^H \mathbf{C}_{s_k} \left(\mathbf{e}_r - \mathbf{w}_k\right), \qquad (50)$$

which suggests that Δ_i depends on the weight vector $\mathbf{w}_k$. This renders the resulting optimization problem in (16) difficult to solve analytically. To mitigate this problem, we notice that due to the averaging with a large time constant (cf. Sections V-A and VI), we have $\Delta_i(l) \approx \Delta_i(l-1)$, making it independent of $\mathbf{w}_k(l)$. In practice, we did not observe any significant difference in the performances between the additive noise and the subtractive models. For the rest of this paper, we use the latter, and thus provide an analysis of it in the sequel.

Substituting (35) in (50) and using $\mathbf{C}_{s_k} = \sigma^2_{s_k} \mathbf{d}_k \mathbf{d}_k^H$ yields:

$$\begin{aligned}\Delta_i &= \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \sigma^2_{s_k} \left| 1 - \alpha_i \mathbf{u}_k^H \mathbf{d}_k - (1-\alpha_i) \left(\mathbf{w}_k^{\mu\text{MWF}}\right)^H \mathbf{d}_k \right|^2 \\ &= \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \sigma^2_{s_k} \left| (1-\alpha_i) \left(1 - g_k^{(\mu)}\right) + \alpha_i (\mathbf{e}_r - \mathbf{u}_k)^H \mathbf{d}_k \right|^2, \qquad (51)\end{aligned}$$

where in (51), we have made use of the facts that $\mathbf{w}_k^{\mu\text{MWF}} = g_k^{(\mu)} \mathbf{w}_k^{\text{MVDR}}$ and $\left(\mathbf{w}_k^{\text{MVDR}}\right)^H \mathbf{d}_k = \mathbf{e}_r^H \mathbf{d}_k = 1$. As seen in (51), dependency of Δ_i on the weight vector is reflected by the presence of α_i. From (51) and (28), one needs the knowledge of α_i to calculate P'_{s_i}. On the other hand, P'_{s_i} has to be known in order to calculate α_i in (36). As suggested above, to cope with this, we make use of the approximation $\bar{\alpha}_i(l) \approx \bar{\alpha}_i(l-1)$, i.e. we use $\bar{\alpha}_i(l-1)$ to calculate $\Delta_i(l)$ and $P'_{s_i}(l)$ in (51) and (28), respectively, and then update $\bar{\alpha}_i(l)$ using $P'_{s_i}(l)$.

1) Ambient-preserving mode: In this mode of operation, we have $\mathbf{u}_k = \mathbf{e}_r$. Substitution in (51) yields:

$$\Delta_i = \frac{(1-\alpha_i)^2}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \sigma^2_{s_k} \left(1 - g_k^{(\mu)}\right)^2. \qquad (52)$$

2) Aggressive mode: In the aggressive mode, we have $\mathbf{u}_k = \mathbf{w}_k^{\text{R}} = g_k^{(\mu_2)} \mathbf{w}_k^{\text{MVDR}}$. Substituting in (51), we obtain:

$$\Delta_i = \frac{1}{\beta_i} \sum_{k \in \mathbb{B}_i} \omega_{i,k} \sigma^2_{s_k} \left| (1-\alpha_i) \left(1 - g_k^{(\mu_1)}\right) + \alpha_i \left(1 - g_k^{(\mu_2)}\right) \right|^2. \qquad (53)$$

VI. Performance Evaluation

In this section, we evaluate the performance of the minimum processing beamformer in the aggressive mode (45) as well as the ambient preserving mode (39) in a hearing aid setup. Although there is no theoretical barrier to apply the proposed beamforming framework in a binaural setting [31], [32], we prefer to confine ourselves to the monaural case in order to keep the evaluation simple and intuitive.

A dummy head with a hearing aid equipped with $M = 2$ microphones is placed at the center of a measurement room with arrays of loudspeakers. We measure the head-related transfer functions (HRTFs) from each loudspeaker to each one of the microphones [33]. The measured HRTFs are then used for calculating the relative acoustic impulse responses $\mathbf{d}_k$. We thus assume in all the simulations, that $\mathbf{d}_k$ are known, and the target speaker is frontal (zero degree azimuth) and in the same plane as the center of the dummy head (zero degree elevation).

The speech material is composed of excerpts of speech, each a few seconds in duration, randomly chosen from a database of recordings of 29 Danish speakers (15 females), reading through randomly selected excerpts of Danish news. The original database is calibrated to ensure a long-term spectrum that matches the ANSI standard speech spectrum level at normal vocal effort (cf. Table 3 in [12]).

At each trial, a speaker is chosen randomly. Next, two excerpts are randomly cut out from his/her speech, ensuring that each excerpt starts with and ends in at least 0.3 seconds of silence to avoid cutting the words. Finally, the two excerpts are united to form the speech material for the current trial. Evaluation of the beamformer performance will, however, be only based on the second of the two excerpts in order to ensure that transient behaviours will not distort the evaluation.

For creating the noise fields, we use both synthetic and realistic setups. For the synthetic setup, we measure the HRTFs from each loudspeaker to each hearing aid in a room where the dummy head is located at the center of a planar circular array of 24 equally distanced loudspeakers. We use this setup to create approximately isotropic multi-talker babble noise fields by playing independent speech realizations from each loudspeaker [33]. For the realistic setup, we use recordings of sound fields from a spherical array of 32 microphones in Oticon headquarters cafeteria in Denmark during the lunch hours. The sound field is then recreated in the measurement room with the dummy head located at the center of three circular arrays with 6, 16, and 6 equaly-distanced loudspeakers at elevations of -45, 0, and 45 degrees, respectively, measured from the center of the dummy head.

Speech and noise are transformed to the time-frequency domain using the STFT with a modified Hann window of size 128 samples with 50 percent overlap at a sampling frequency of 20 kHz. For the subband filters, we use a set of $n = 18$ one-third octave band Butterworth filters. Throughout all the experiments, the value of the time constant τ in (49) is kept fixed at $\tau = 2$ seconds, and the target audibility is set to $I_i = 0.8$ for all subbands, unless otherwise is stated. For the ambient-preserving mode (39), we set $\mu = 1$, and for the aggressive mode (45), we set $\mu_1 = 5$ and $\mu_2 = 0$.

In order to estimate the spectra of speech and noise, we used the maximum likelihood estimation technique in [34], [35], assuming that an ideal voice activity detector is available. We compare the proposed beamformer with the following beamformers: the μMWF-N beamformer in (15) with $\mu = 5$ and $\alpha = 0.2$,[2] the standard MWF, and the MVDR beamformer. For comparing the quality of the beamformer output sounds we use the perceptual evaluation of speech quality (PESQ) [21]. For speech intelligibility, we use the short-term objective intelligibility (STOI) [22], even though the proposed method is

[2] Although these values (taken from [4]) are originally meant for binaural beamforming, we have observed that they are also reasonable for the monaural case.

based on SII, the reason being the general agreement on STOI as a more general predictor of speech intelligibility [23].

A. Idealized Case

To validate the solution offered in Section IV-C with regard to the theoretical framework of processing penalty-versus-performance, it is desirable to study the behaviour of the beamformer under idealized situations. The SII-based performance criterion and MSE-based processing penalty introduced in Section IV are based on statistical expectations. In order to make a reasonable estimation of the actual performance and processing at the output of the beamformer, one needs to average across a large number of time frames. For that reason, the ideal condition for evaluating the behaviour of the beamformer is when the "speech" and "noise" signals are stationary processes drawn from known distributions. For this purpose, we create stationary Gaussian signals and use them as speech and noise. To keep the simulations minimalistic and intuitive, we shape the spectrum of the noise in a manner that matches the spectrum of the speech after having been convolved with the HRTFs. We then adjust the SNR by scaling the noise. This rules out any impact due to the differences between the subbands, and thus "audibility" per subband will be equal to the overall "intelligibility". Moreover, we set the hearing threshold (modelled by Λ_i) to zero to further simplify the simulations. We focus on the ambient-preserving mode of operation, since it more naturally lends itself to interpretation.

Fig. 4 shows the plots of the achieved "intelligibility" and the corresponding processing penalty versus SNR at the reference microphone for different values of the target intelligibility. The two extreme cases with $I_i = 0$ and $I_i = 1$ are particularly informative, since they correspond to no processing and full processing (since $\Lambda_i = 0$). As seen in the top panel of Fig. 4, for $I_i = 0$ (target SII = 0), the psychometric function in Fig. 3 is reproduced as expected. For $I_i = 1$, the achieved "intelligibility" is greater than that of the unprocessed noisy speech ($I_i = 0$) in a wide range of SNRs. This is in price of a higher processing penalty as seen in the lower panel of Fig. 4. Note that we do not achieve a processing penalty of $-\infty$ in dB for $I_i = 0$ due to round-off errors. For $0 < I_i < 1$, as long as the target I_i is not achievable even with full processing, the plots coincide with that of $I_i = 1$. Once I_i is achieved, there will be a transition, where as the SNR increases, we move from full processing to no processing.

B. Aggressive Mode

Figs 5 and 6 plot STOI and PESQ scores as a function of the global input SNR for the proposed beamformer as well as three existing beamformers for approximately isotropic multi-talker babble and cafeteria noises, respectively. In addition to the conventional PESQ score, we have also calculated the so-called *speech PESQ* score, which is the PESQ score when the clean speech part of a given beamformer output is compared to the clean speech at the reference microphone. Hence, speech PESQ quantifies processing distortions imposed on the target speech signal. The conventional PESQ score is penalized both by the existence of residual noise and distortion of target speech. However, in a beamforming setup, the residual noise is by far the dominating part in a wide range of input SNRs, such that the distortion of the clean speech is unlikely to be clearly represented by the conventional PESQ score. For that reason, we have also plotted the speech PESQ scores versus SNR to exclusively focus on distortion of the speech, regardless of the residual noise. Moreover, we have also shown *difference plots*, where the horizontal axes represent the scores for the standard MWF.

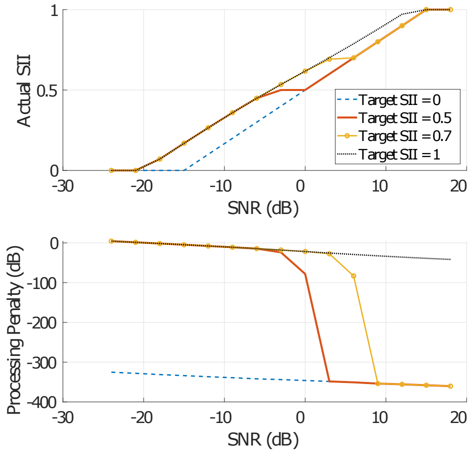

Fig. 4. Achieved SII and the corresponding processing versus input SNR for the idealized stationary Gaussian case for various choices of the target intelligibility. The legends on the top panel apply to both panels.

As seen in Figures 5 and 6 (and as expected), the MVDR beamformer does not distort the speech in price of relatively poor performance in terms of noise reduction, as clear from the PESQ scores. On the contrary, the standard MWF achieves a higher level of noise reduction compared to the MVDR beamformer in price of severely distorting the speech. The μMWF-N beamformer achieves a slightly higher PESQ score compared to the standard MWF, especially at higher SNRs, supposedly because of the aggressive noise reduction with $\mu = 5$. The distortion of the clean speech is also less severe compared to the standard MWF (beacuse of adding a portion of the unprocessed signal to the beamformer output), but it is still significant. Moreover, the STOI scores at lower SNRs (where it matters the most) are slightly lower compared to all the other methods. Because of the dynamic combination of the two beamformers in (45), the proposed method seems to achieve "the best of both worlds". At lower SNRs, where the noise impact is so severe that PESQ scores are the same for all the beamformers, and STOI scores matter the most, the proposed method achieves the highest STOI scores. At higher SNRs, where full intelligibility is already achieved, and the PESQ scores matter the most, the proposed method achieves the highest PESQ scores. In addition to these, the proposed method keeps the speech essentially distortionless within a wide range of SNRs.

C. Ambient-Preserving Mode

Unlike the traditional conception of beamforming, where noise is to be suppressed as much as possible, the goal here is to retain noise in addition to speech, whenever a certain level

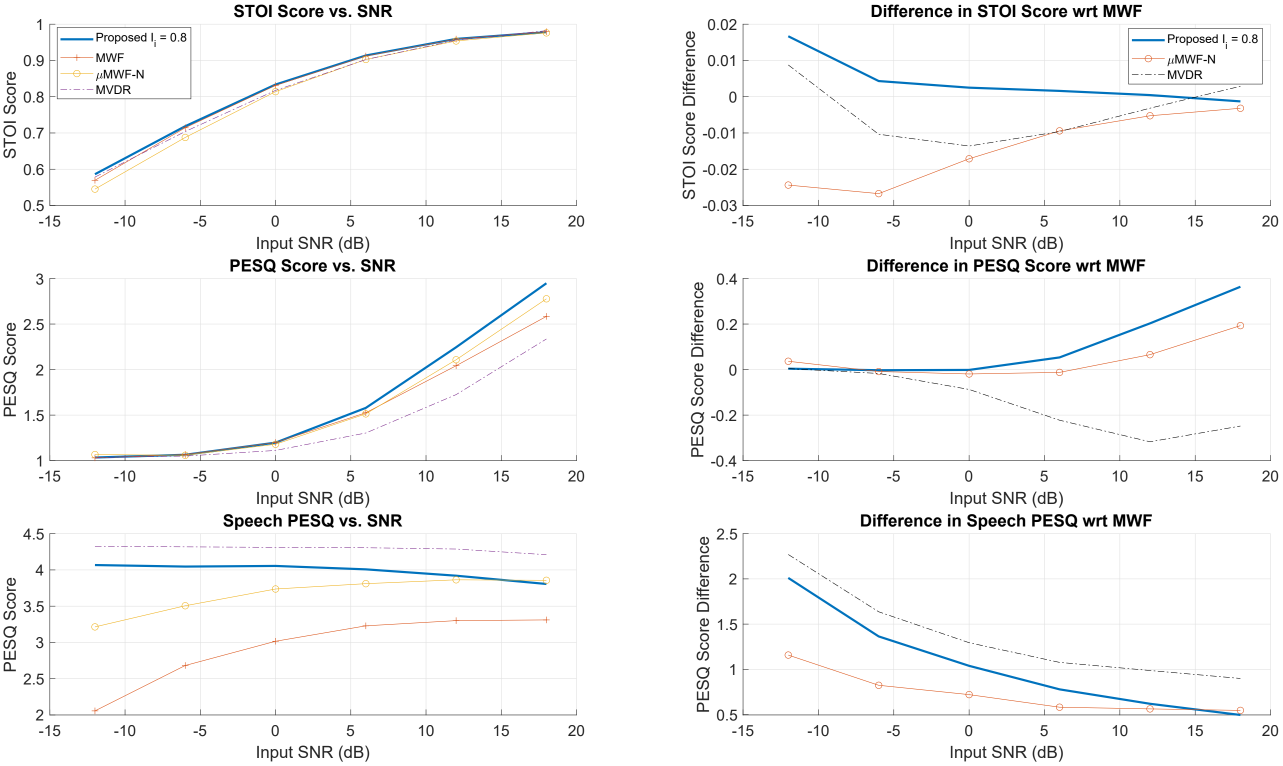

Fig. 5. STOI, PESQ and speech PESQ scores in approximately isotropic babble noise for the proposed method in the aggressive mode and the existing methods. Left panel shows the absolute scores, and right panel shows the difference scores with respect to the standard MWF. Legends on the top panels apply to the lower panels, too.

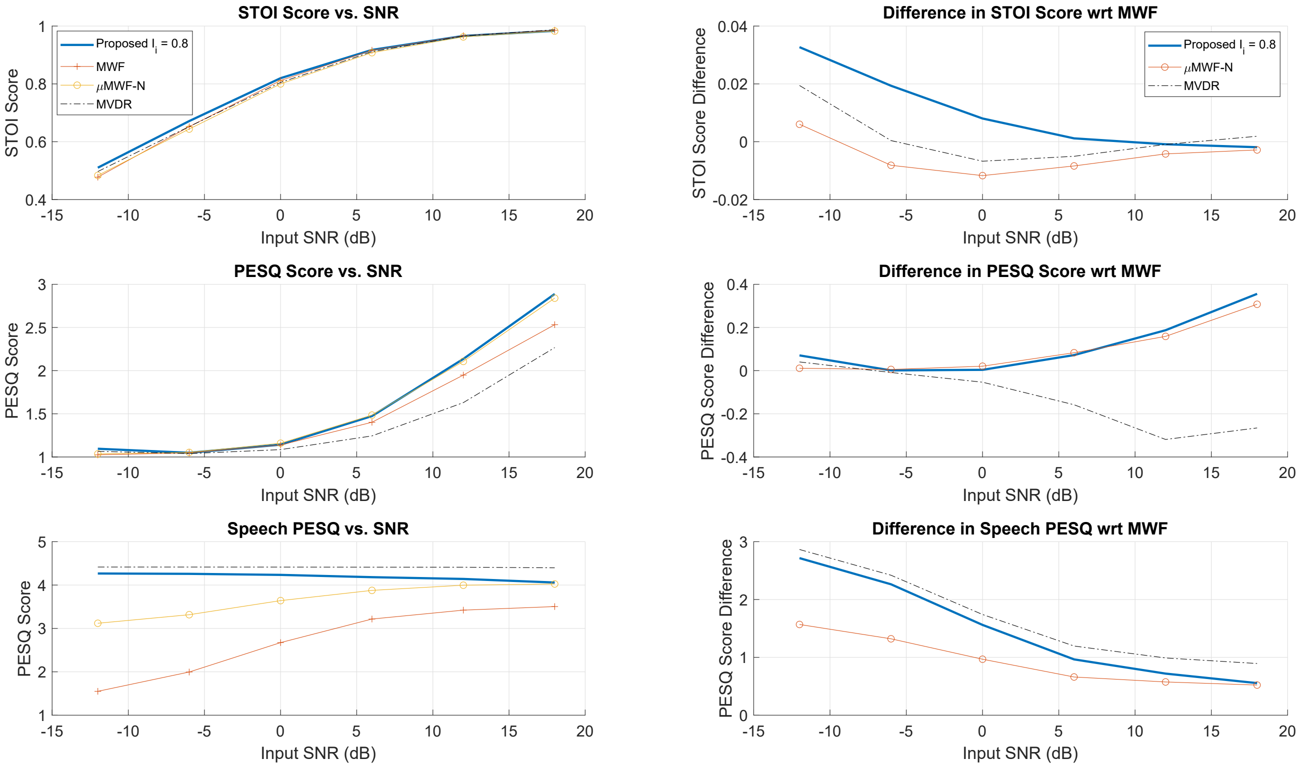

Fig. 6. STOI, PESQ and speech PESQ scores in cafeteria noise for the proposed method in the aggressive mode and the existing methods. Left panel shows the absolute scores, and right panel shows the difference scores with respect to the standard MWF. Legends on the top panels apply to the lower panels, too.

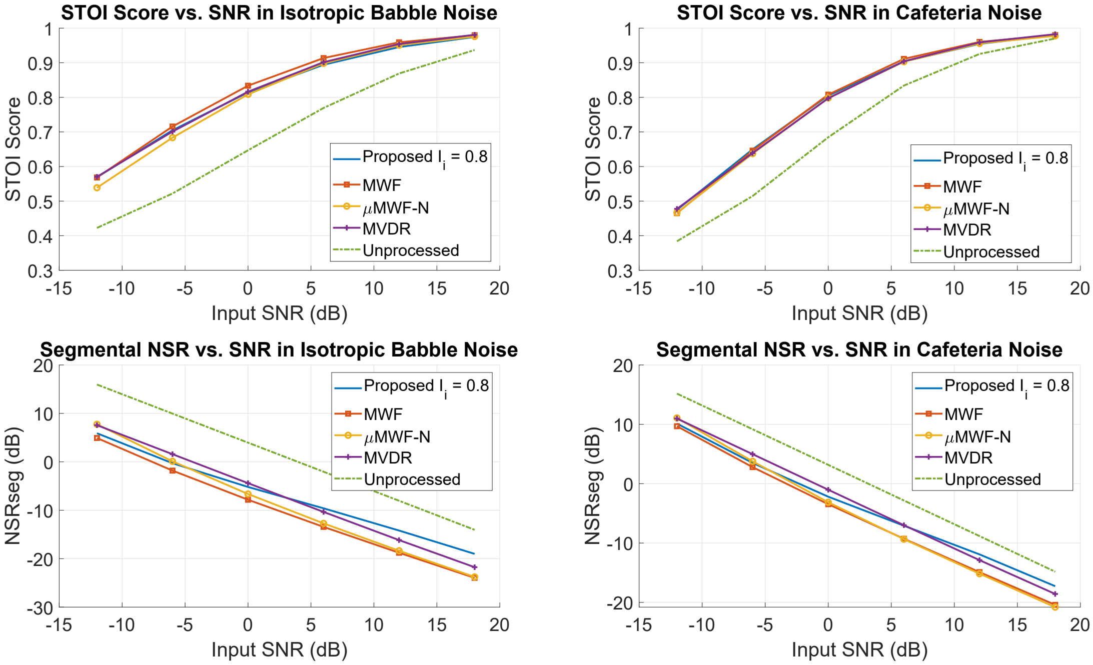

Fig. 7. STOI and NSRsig scores versus input SNR for approximately isotropic babble and cafeteria noises.

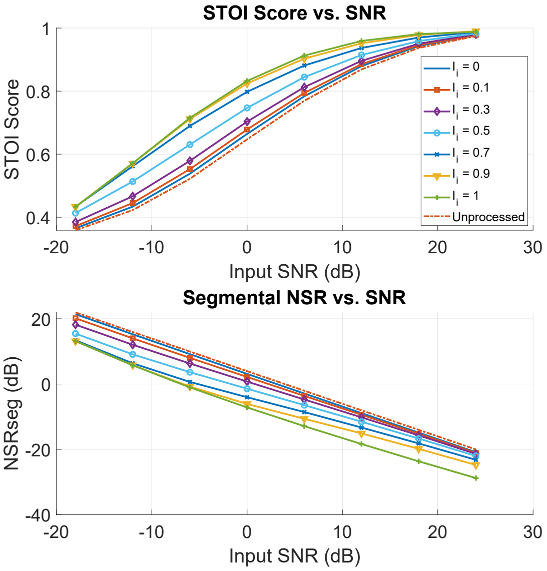

Fig. 8. STOI and NSRsig scores versus input SNR for the proposed beamformer with various choices of the target audibility in approximately isotropic babble noise.

of speech intelligibility is nevertheless achieved. This implies that standard evaluation methods used in Section VI-B are not suitable here. We therefore need to take a different approach, while trying to keep it as simple as possible to minimize the inherent unorthodoxy. Since intelligibility is key here, we use STOI in a similar manner to Section VI-B. At the same time, we need to quantify the beamformers ability to preserve noise. Since noise does not necessarily only consist of unwanted speech, it is not clear how one can quantify noise quality using the existing measures. A simple choice is the segmental noise to signal ratio (NSRseg), which is calculated during speech present intervals in a similar manner to segmental SNR, but with swapping the roles of speech and noise. The plots are shown in Fig. 7 for approximately isotropic babble and cafeteria noises. As seen, at lower input SNRs, the proposed method suppresses noise similar to the other beamformers. At higher SNRs, where noise does not interfere with speech intelligibility, the proposed beamformer retains a larger portion of noise without any loss in the STOI score.

It is important to notice that the target audibility for the proposed beamformer was deliberately set very high ($I_i = 0.8$) to ensure no loss in the STOI scores compared to the existing beamformers. However, STOI scores have a nonlinear relationship with speech intelligibility [22]. For instance, at a STOI score of around 0.8, full intelligibility is already achieved for the Dantale II [36] and IEEE (cf. [20]) databases in the presence of different types of artifacts [22]. The STOI scores achieved in Fig. 7 may therefore be unnecessarily high. To illustrate how one can control the compromise between intelligibility and residual noise by adjusting the target audibility, we plot the STOI and NSRseg scores as a function of SNR for the proposed beamformer with different target audibilities in approximately isotropic babble noise. The result is shown in Fig. 8. The plots suggest that sufficient intelligibility may be achievable with lower target audibilities, therefore preserving a larger portion of the background noise compared to Fig. 7. In practice, choice of I_i should depend on the application.

VII. Conclusions

We proposed a generalized beamforming rationale, where the beamformer output is optimized to take the minimum distance (processing) from a given reference signal, with the restriction that a certain performance criterion is fulfilled. We showed that the classic speech-versus-noise beamforming rationale is covered as a special case of the proposed rationale, while in general, it goes beyond this dichotomy. We provided a full analysis of the proposed concept, when the performance criterion is based on the speech intelligibility index, and the processing measure is based on the mean-square error. For the resulting beamformer, we studied two modes of operation, both theoretically and experimentally: an ambient preserving mode that keeps the processing of the noisy speech to a minimum necessary to achieve the desired intelligibility, and an aggressive mode that maximally suppresses noise as long as it does not compromise speech quality. Experimental studies verified the advantages of each mode of operation within its own relevant context.

Appendix A

The simplest way to make a correspondence between the frequency bins and subbands is to assign all the frequency bins k whose frequencies fall within the range of a certain subband i to that subband. This is equivalent to a partitioning of the set of all frequency bins $\{1, 2, ..., K\}$ into n disjoint subsets $\mathbb{B}_1, ..., \mathbb{B}_n$. In this case, $\omega_{i,k} = 1$, if $k \in \mathbb{B}_i$, and $\omega_{i,k} = 0$, otherwise. The advantage of such a mapping is that, there is no overlap between the subbands; i.e. each frequency bin contributes to only one subband. In many practical applications, however, such as in hearing aids, there are strict limits on the maximum tolerable processing delay. This means that the time frames must be short, leading to low-resolution frequency bins. On the other hand, perceptually motivated subband divisions, such as the critical or fractional octave bands, tend to have a relatively high frequency resolution in lower frequency subbands. Consequently, no frequency bins may be assigned to some of the lower frequency subbands, when the subbands have a higher resolution than the frequency bins. This means that such subbands will be ignored altogether. Moreover, the overall correspondence between the frequency bins and subbands will be imprecise.

In this work, we have developed and used another correspondence, which does not suffer from the above-mentioned issue in price of engendering overlap between the neighbouring subbands. A diagram of this method is shown in Fig. 9. A long sequence of white Gaussian noise [3] is taken to the time-frequency domain using the same transformation applied to the microphone signals for representing them in frequency bins. The component at a certain frequency bin k is then isolated and returned to the time domain. The resulting sequence u_k is then processed using the filter bank that defines the subbands. For a given subband i, the variance $\sigma^2_{k,i}$ of the output of the subband filter is then estimated. Finally, the contribution $\omega_{i,k}$ of frequency bin k to subband i is calculated as:

$$\omega_{i,k} = \frac{\sigma^2_{k,i}}{\sum_{j=1}^{n} \sigma^2_{k,j}} \frac{\sum_{l=1}^{K} \sigma^2_{l,i}}{\sigma^2_{u_k}}. \tag{54}$$

The first term in (54) is a normalization to ensure that the weights $\omega_{i,k}$ for $i = 1, ..., n$ give the relative contribution of the bin k to all the subbands. The second term is the output to input power ratio for subband filter i, when frequency bin k in isolation is under consideration. This term takes account of any power dissipation due to subband filtering.

When the range of frequencies associated to bin k is essentially outside subband i, the estimated contribution $\omega_{i,k}$ will be very small. In such cases, to avoid unnecessary computational complexity, $\omega_{i,k}$ can be rounded off to zero. We therefore set to zero values of $\omega_{i,k}$ that are below a certain threshold. As an example, we consider a sampling frequency of 20 kHz, using an STFT with a frame size of 128 samples and 50 percent overlap between the frames for the time-frequency representation, and one-third octave band filter-banks with $n = 18$ subbands (nominal midband frequencies ranging from 0.16 to 8 kHz) for the subband decomposition. The result is summarized in Table I, where for each subband $i = 1, ..., 18$, the contributing frequency bins k (for the proposed scheme with a threshold of 0.01) and k' (for the traditional partitioning method introduced at the beginning of this section) are listed. The weights $\omega_{i,k}$ are also shown in the table. However, since $\omega_{i,k'} = 1$ for all the contributing frequency bins k', we have not included them in Table I. As shown in the table, with the traditional partitioning method, no frequency bin is assigned to subbands 2,3 and 5. This means that these subbands will be ignored. With the proposed scheme, this issue is clearly addressed by assigning some of the frequency bins to multiple subbands. Moreover, it is seen that the distribution of the weights $\omega_{i,k}$ with the proposed scheme has a tapered form rather than the binary form resulting from the partitioning scheme. This leads to a more precise account of the contributions that each frequency bin makes to the given subband.

For the traditional partitioning, the proposed minimum processing beamformer in (35) leads to one weight vector per frequency bin. For the procedure described in Fig. 9, if a frequency bin is associated to more than one subband, one obtains different weight vectors for the same frequency bin. To produce a final beamformer output for each frequency bin, one needs to apply a combination formula such as the one in (37). The weights $\eta_{i,k}$ can be simply calculated using the first term on the right-hand side of (54). However, we use the square root of the variances, since the weights are to be applied to beamformer weight vectors, unlike $\omega_{i,k}$ in (54), which are applied to power spectra.

Appendix B

We prove that the μMWF beamformer in (12) can be written as the cascade of the MVDR beamformer and the generalized postfilter given in (13). Starting from (12) and substituting (5), we have:

$$\begin{aligned}\mathbf{w}_k^{\mu\text{MWF}} &= \left(\mathbf{C}_{x_k}^{(\mu)}\right)^{-1} \mathbf{C}_{s_k}\mathbf{e}_r \\ &= \left(\mathbf{C}_{s_k} + \mu\mathbf{C}_{v_k}\right)^{-1} \mathbf{C}_{s_k}\mathbf{e}_r \\ &= \left(I_M + \frac{1}{\mu}\mathbf{C}_{v_k}^{-1}\mathbf{C}_{s_k}\right)^{-1} \frac{1}{\mu}\mathbf{C}_{v_k}^{-1}\mathbf{C}_{s_k}\mathbf{e}_r\end{aligned}$$

[3] The white noise sequence, in effect, serves as stimulus for measuring the impulse responses of the subband filters. Alternatively, other types of noise such as speech-shaped noise can be used, or one can directly use the impulse responses if available.

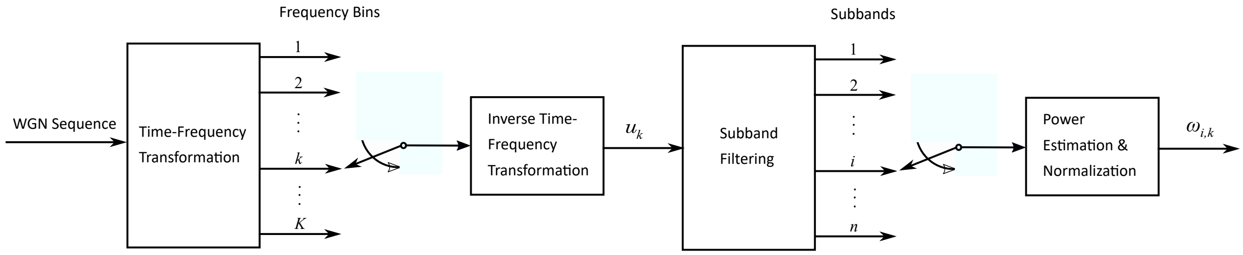
Frequency Bins
1
2
⋮
k
⋮
K
WGN Sequence
Time-Frequency Transformation
Inverse Time-Frequency Transformation
u_k
Subbands
1
2
⋮
i
⋮
n
Subband Filtering
Power Estimation & Normalization
$\omega_{i,k}$

Fig. 9. Procedure for estimating the contribution $\omega_{i,k}$ of the frequency bin k to subband i.

TABLE I
FREQUENCY BINS k (PROPOSED) AND k' (PARTITIONING) FOR THE 18 ONE-THIRD OCTAVE BANDS TOGETHER WITH THE WEIGHTS $\omega_{i,k}$

i	k'	k	$\omega_{i,k}$
1	2	1,2	0.01,0.25
2	-	2,3	0.23,0.05
3	-	2,3	0.10,0.24
4	3	2-4	0.01,0.47,0.02
5	-	3,4	0.22,0.32
6	4	4,5	0.60,0.13
7	5	4-6	0.05,0.80,0.10
8	6	5-7	0.07,0.87,0.21
9	7,8	6-9	0.03,0.79,0.69,0.02
10	9,10	8-10	0.31,0.98,0.55
11	11,12	10-13	0.45,0.99,0.85,0.05
12	13-15	12-16	0.15,0.95,0.99,0.81,0.04
13	16-18	15-19	0.19,0.96,1,0.99,0.54
14	19-23	19-24	0.46,0.99,1,1,0.97,0.22
15	24-29	23-30	0.03,0.78,1,1,1,1,0.94,0.14
16	30-36	29-37	0.07,0.86,1,1,1,1,1,0.99,0.49
17	37-46	36-47	0.01,0.52,0.99,1,1,1,1,1,1,1,0.79,0.04
18	47-58	46-58	0.25,0.93,1,1,1,1,1,1,1,1,1,0.55

$$=\left(I_M+\frac{1}{\mu}\sigma_{s_k}^2\mathbf{C}_{v_k}^{-1}\mathbf{d}_k\mathbf{d}_k^H\right)^{-1}\frac{1}{\mu}\sigma_{s_k}^2\mathbf{C}_{v_k}^{-1}\mathbf{d}_k \tag{55}$$

$$=\left(I_M-\frac{\frac{1}{\mu}\sigma_{s_k}^2}{1+\frac{1}{\mu}\sigma_{s_k}^2\mathbf{d}_k^H\mathbf{C}_{v_k}^{-1}\mathbf{d}_k}\mathbf{C}_{v_k}^{-1}\mathbf{d}_k\mathbf{d}_k^H\right)\frac{\sigma_{s_k}^2}{\mu}\mathbf{C}_{v_k}^{-1}\mathbf{d}_k \tag{56}$$

$$=\frac{\sigma_{s_k}^2}{\mu}\mathbf{C}_{v_k}^{-1}\mathbf{d}_k-\frac{\frac{\xi_k}{\mu}}{1+\frac{\xi_k}{\mu}}\frac{\sigma_{s_k}^2}{\mu}\mathbf{C}_{v_k}^{-1}\mathbf{d}_k \tag{57}$$

$$=g_k^{(\mu)}\mathbf{w}_k^{\text{MVDR}},$$

where (55) follows because $\mathbf{C}_{s_k}=\sigma_{s_k}^2\mathbf{d}_k\mathbf{d}_k^H$ and $\mathbf{d}_k^H\mathbf{e}_r=1$, (56) follows from the matrix inversion lemma [37], and (57) is because the SNR at the output of the MVDR beamformer is given by $\xi_k=\sigma_{s_k}^2\mathbf{d}_k^H\mathbf{C}_{v_k}^{-1}\mathbf{d}_k$.

APPENDIX C

The second constraint in (29) reflects the two boundary conditions in (23), making sure that the estimated band audibility remains in the interval $[0,1]$. While we get back to this constraint later, for the time being, we disregard it. The $\max(\cdot,\cdot)$ operator in (29) has to be tackled by a separate treatment of the two possible cases; i.e. whether the first argument is greater than the second or otherwise. We start with assuming that Λ_i is the smaller of the two arguments of the $\max(\cdot,\cdot)$ operator. Given these assumptions, the optimization problem in (29) reduces to the following:

$$\min_{\mathbf{w}_k\ \text{for all}\ k\in\mathbb{B}_i}\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{u}_k-\mathbf{w}_k\right)^H\mathbf{C}_{x_k}^{(\mu)}\left(\mathbf{u}_k-\mathbf{w}_k\right)\quad\text{s.t.}$$
$$\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}(\mathbf{e}_r-\mathbf{w}_k)^H\mathbf{C}_{s_k}(\mathbf{e}_r-\mathbf{w}_k)+\mu\omega_{i,k}\mathbf{w}_k^H\mathbf{C}_{v_k}\mathbf{w}_k\leq P'_{s_i}10^{-3\left(I'_i-\frac{1}{2}\right)}, \tag{58}$$

Writing (58) in Lagrangian form with parameter λ_i and setting the derivative with respect to $\mathbf{w}_k$ (for a certain $k\in\mathbb{B}_i$) equal to zero, we obtain:

$$\mathbf{C}_{x_k}^{(\mu)}\left(\mathbf{w}_k-\mathbf{u}_k\right)+\lambda_i\left\{\mathbf{C}_{s_k}(\mathbf{w}_k-\mathbf{e}_r)+\mu\mathbf{C}_{v_k}\mathbf{w}_k\right\}=0, \tag{59}$$

Rearranging the terms in (59), using (5) and (12), and solving for $\mathbf{w}_k$ yields:

$$\begin{aligned}\mathbf{w}_{k,i}^{\text{MP}}&=\frac{1}{1+\lambda_i}\mathbf{u}_k+\frac{\lambda_i}{1+\lambda_i}\mathbf{w}_k^{\mu\text{MWF}}\\&=\alpha_i\mathbf{u}_k+(1-\alpha_i)\mathbf{w}_k^{\mu\text{MWF}},\end{aligned} \tag{60}$$

where $0\leq\alpha_i\leq1$ is defined as:

$$\alpha_i\triangleq\frac{1}{1+\lambda_i}. \tag{61}$$

To calculate α_i, we first obtain the optimal processing penalty $\mathcal{D}^\star(\mathbf{y}_i^{\text{R}},\mathbf{y}_i)$ by substituting (60) in the cost function in (58):

$$\begin{aligned}\mathcal{D}^\star(\mathbf{y}_i^{\text{R}},\mathbf{y}_i)&=\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{u}_k-\mathbf{w}_{k,i}^{\text{MP}}\right)^H\mathbf{C}_{x_k}^{(\mu)}\left(\mathbf{u}_k-\mathbf{w}_{k,i}^{\text{MP}}\right)\\&=\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(1-\alpha_i\right)^2\left(\mathbf{u}_k-\mathbf{w}_k^{\mu\text{MWF}}\right)^H\mathbf{C}_{x_k}^{(\mu)}\left(\mathbf{u}_k-\mathbf{w}_k^{\mu\text{MWF}}\right)\\&=h_i\left(1-\alpha_i\right)^2,\end{aligned} \tag{62}$$

where (62) follows from (31). Next we substitute (60) in the constraint in (58) to calculate the optimal output error $N_i^\star$ as:

$$\begin{aligned}N_i^\star&=\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{w}_{k,i}^{\text{MP}}\right)^H\mu\mathbf{C}_{v_k}\mathbf{w}_{k,i}^{\text{MP}}\\&\quad+\omega_{i,k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{e}_r\right)^H\mathbf{C}_{s_k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{e}_r\right)\\&=\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k+\mathbf{u}_k\right)^H\mu\mathbf{C}_{v_k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k+\mathbf{u}_k\right)\\&\quad+\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k+\mathbf{u}_k-\mathbf{e}_r\right)^H\mathbf{C}_{s_k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k+\mathbf{u}_k-\mathbf{e}_r\right)\\&=\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k\right)^H\mathbf{C}_{x_k}^{(\mu)}\left(\mathbf{w}_{k,i}^{\text{MP}}-\mathbf{u}_k\right)\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{w}^{\mathrm{MP}}_{k,i}-\mathbf{u}_k\right)^H\mathbf{C}^{(\mu)}_{x_k}\left(\mathbf{u}_k-\mathbf{w}^{\mu\mathrm{MWF}}_k\right)\\
&+\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left(\mathbf{u}_k-\mathbf{w}^{\mu\mathrm{MWF}}_k\right)^H\mathbf{C}^{(\mu)}_{x_k}\left(\mathbf{w}^{\mathrm{MP}}_{k,i}-\mathbf{u}_k\right)\\
&+\frac{1}{\beta_i}\sum_{k\in\mathbb{B}_i}\omega_{i,k}\left\{\mu\mathbf{u}_k^H\mathbf{C}_{v_k}\mathbf{u}_k+(\mathbf{u}_k-\mathbf{e}_r)^H\mathbf{C}_{s_k}(\mathbf{u}_k-\mathbf{e}_r)\right\}\\
&=h_i\left(1-\alpha_i\right)^2-2(1-\alpha_i)h_i+N_i^{\mathrm{R}} \qquad (63)\\
&=N_i^{\mathrm{R}}-h_i\left(1-\alpha_i^2\right), \qquad (64)
\end{aligned}
$$

where (63) follows from (62), (30) and (31).

We now turn our attention to the second case where Λ_i is the greater of the two arguments in the $\max(\cdot,\cdot)$ operator in (29). In this case, the constraint becomes irrelevant, and the solution to the optimization problem is the sheer minimizer of the processing penalty; i.e.

$$
\begin{aligned}
\mathbf{w}^{\mathrm{MP}}_{k,i}&=\mathbf{u}_k \qquad (65)\\
N_i^{\star}&=N_i^{\mathrm{R}}. \qquad (66)
\end{aligned}
$$

It is easy to see that (65) and (66) are special cases of (60) and (64), respectively, for $\alpha_i=1$. In other words, when the error at the output of the reference beamformer is already below the threshold of hearing, there is no need to process its output, since in this case, the error will be inaudible to the listener.

The $\max(\cdot,\cdot)$ operator in (29) can be written as $\max(\Lambda_i,N_i^{\star})$. Noticing that $N_i^{\star}$ in general depends on α_i as seen in (64), one must have the knowledge of α_i in order to calculate $N_i^{\star}$. On the other hand, it appears that to calculate α_i, one has to compare Λ_i with $N_i^{\star}$ in order to evaluate $\max(\Lambda_i,N_i^{\star})$. Therefore, it may appear that these dependencies prevent an analytical solution. Below, we show that this is not the case. More precisely, we show that $\max(\Lambda_i,N_i^{\star})$ can be evaluated simply by comparing Λ_i and N_i^{R}.

Suppose that $\Lambda_i\geq N_i^{\star}$. From (66) it follows that $\Lambda_i\geq N_i^{\mathrm{R}}$. Now suppose that $\Lambda_i<N_i^{\star}$. We immediately conclude that $\Lambda_i<N_i^{\mathrm{R}}$, since $N_i^{\star}=N_i^{\mathrm{R}}-h_i\left(1-\alpha_i^2\right)$, and $h_i\geq 0$ and $\alpha_i\leq 1$. Combining these two statements, it follows that $\max(\Lambda_i,N_i^{\star})=\max(\Lambda_i,N_i^{\mathrm{R}})$.

From the argument above, we also derive a lower limit on α_i, when $\Lambda_i<N_i^{\star}$. From $\Lambda_i<N_i^{\mathrm{R}}-h_i\left(1-\alpha_i^2\right)$ it follows that:

$$
\alpha_i^2>1-\frac{N_i^{\mathrm{R}}-\Lambda_i}{h_i}.
$$

This together with $\alpha_i\geq 0$ give the lower limit (34) on α_i.

Substituting (60) in the third condition in (23), we also calculate the optimal band audibility:

$$
\begin{aligned}
\Psi^{\star}(\zeta_i)&=\frac{10\log\zeta_i+15}{30} \qquad (67)\\
&=\frac{1}{2}+\frac{1}{3}\log\frac{P'_{s_i}}{\max\left(\Lambda_i,N_i^{\mathrm{R}}-h_i\left(1-\alpha_i^2\right)\right)}. \qquad (68)
\end{aligned}
$$

Applying the first two conditions in (23) to limit the range of $\Psi^{\star}(\zeta_i)$ to $[0,1]$ in (68), we obtain:

$$
\Psi^{\star}(\zeta_i)=
$$

$$
\min\left(1,\frac{1}{2}+\frac{1}{3}\max\left(-\frac{3}{2},\log\frac{P'_{s_i}}{\max\left(\Lambda_i,N_i^{\mathrm{R}}-h_i\left(1-\alpha_i^2\right)\right)}\right)\right).
$$

Setting $\alpha_i=1$, we obtain the formula for $I_i^{\min}$ in (32). Similarly, setting $\alpha_i=\alpha_i^{\min}$ yields (33). Notice that $I_i^{\min}\geq 0$ and $I_i^{\max}\leq 1$. Recall that the role of the second constraint in (29) is two ensure that the estimated audibility will be limited to the interval $[0,1]$. Since the solution obtained above enforces a stronger constraint, i.e. $I_i^{\min}\leq\Psi^{\star}(\zeta_i)\leq I_i^{\max}$, the original constraint is already satisfied and there is no need for further check on it. Finally, to complete the proof, we substitute (60) in the constraint in (58) to obtain the third condition in (36). This yields:

$$
\alpha_i^2=1-\frac{N_i^{\mathrm{R}}-P'_{s_i}10^{-3\left(I_i-\frac{1}{2}\right)}}{h_i}. \qquad (69)
$$

Limiting the range of α_i as before, we obtain the statement in the third condition in (36). This completes the proof.

References

[1] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*. PhD dissertation, KU Leuven, 2003.

[2] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," in *Speech enhancement*. Springer, 2005, pp. 199–228.

[3] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[4] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Am. (JASA)*, vol. 124, no. 1, pp. 484–497, 2008.

[5] ——, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Am. (JASA)*, vol. 125, no. 1, pp. 360–371, 2009.

[6] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction," *J. Acoust. Soc. Am. (JASA)*, vol. 131, no. 6, pp. 4743–4755, 2012.

[7] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2162–2176, 2015.

[8] ——, "Perceptually motivated coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 2014.

[9] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2384–2397, 2015.

[10] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Extensions of the binaural mwf with interference reduction preserving the binaural cues of the interfering source," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 2016.

[11] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[12] ANSI-S3-22-1997, "Methods for calculation of the speech intelligibility index," *American National Standard Institute*, 1997.

[13] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Robust beamforming by linear programming," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1834–1849, 2014.

[14] M. O'Connor, W. B. Kleijn, and T. Abhayapala, "Distributed sparse mvdr beamforming using the bi-alternating direction method of multipliers," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*. IEEE, 2016, pp. 106–110.

[15] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, 2017.

[16] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 825–836, 2015.

[17] X. Xiao *et al.*, "A study of learning based beamforming methods for speech recognition," in *Proc. CHiME workshop*, 2016.

[18] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[19] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[20] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. 2001 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2001, pp. 749–752.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, 2011.

[23] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, 2016.

[24] J. M. Kates and K. H. Arehart, “The hearing-aid speech perception index (haspi),” *Speech Communication*, vol. 65, pp. 75–93, 2014.

[25] ——, “The hearing-aid speech quality index (hasqi),” *J. Audio Eng. Soc. (JAES)*, vol. 58, no. 5, pp. 363–381, 2010.

[26] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise,” *J. Acoust. Soc. Am. (JASA)*, vol. 120, no. 6, 2006.

[27] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone arrays*. Springer, 2001, pp. 39–60.

[28] T. Piechowiak, J. Udesen, K. Moeller, F. Gran, and A. Dittberner, “Promoting off-axis listening and preserving spatial cues with binaural directionality ii,” in *Proc. Int. Symp. Auditory, Audiological Research*, 2015.

[29] M. A. Akeroyd and W. M. Whitmer, “Spatial hearing and hearing aids,” in *Hearing aids*. Springer, 2016, pp. 181–215.

[30] J. Sherman and W. J. Morrison, “Adjustment of an inverse matrix corresponding to a change in one element of a given matrix,” *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.

[31] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, “Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,” *IEEE Trans. Signal Process.*, vol. 55, no. 4, 2007.

[32] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[33] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, “Personalized signal-independent beamforming for binaural hearing aids,” *J. Acoust. Soc. Am. (JASA)*, vol. 145, no. 5, 2019.

[34] U. Kjems and J. Jensen, “Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement,” in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, 2012, pp. 295–299.

[35] J. Jensen and M. S. Pedersen, “Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2015.

[36] K. Wagener, J. L. Josvassen, and R. Ardenkjær, “Design, optimization and evaluation of a danish sentence test in noise,” *International journal of audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[37] D. S. Bernstein, *Matrix mathematics: theory, facts, and formulas*. Princeton university press, 2009.

Adel Zahedi received the M.Sc. degree from Iran University of Science and Technology, Iran in 2011, and the Ph.D. degree from Aalborg University, Denmark in 2016. From 2016 to 2018, he was a postdoctoral researcher with the Department of Electronic Systems, Aalborg University. In 2018, he joined Oticon A/S, Denmark, where he is an Industrial Postdoc. Adel’s research areas include Statistical Signal Processing with focus on Audio and Speech Processing.

Michael S. Pedersen

Jan Østergaard (S’98–M’99–SM’11) received the M.Sc.E.E. degree from Aalborg University, Aalborg, Denmark, in 1999 and the Ph.D. degree (*cum laude*) from Delft University of Technology, Delft, The Netherlands, in 2007. He was an R&D Engineer with ETI Inc., VA, USA. Between September 2007 and June 2008, he was a Postdoctoral Researcher at The University of Newcastle, NSW, Australia. Dr. Østergaard is currently a Professor in Information Theory and Signal Processing, Head of the Section on AI & Sound, and Head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University.

Thomas U. Christiansen received the M.Sc. degree in computer science as a major and linguistics as minor from the University of Copenhagen, Copenhagen, Denmark, in 1999 and the Ph.D. degree in electrical engineering from the Technical University of Denmark, Lyngby, in 2004. From 2004 to 2012, he was Post-doc, Assistant Professor, Associate Professor and Senior Scientist with the Centre for Applied Hearing Research at the Technical University of Denmark. In 2013 he was employed as Senior Research and Development Engineer with Oticon A/S. At present he is still employed at Oticon. His major topics of interests include models of the normal and impaired auditory periphery and its relation to phonetics and speech perception.

Lars Bramsløw is a Senior Scientist within the Augmented Hearing Group at the Eriksholm Research Centre, part of Oticon A/S, Denmark. He holds an M.Sc. and a Ph.D. degree, both from Technical University of Denmark. The Ph.D. was carried out at Eriksholm. Lars has 30 years of extensive experience in acoustics, hearing science and hearing aid research and development, including employment at House Ear Institute in Los Angeles, Eriksholm and Oticon headquarters in Smørum, Denmark. He currently works on the applications of deep learning algorithms in hearing health care.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Principal Scientist with Oticon A/S, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, at Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.