Multizone Speech Reinforcement

João B. Crespo and Richard C. Hendriks

***Abstract*—In this article, we address speech reinforcement (near-end listening enhancement) for a scenario where there are several playback zones. In such a framework, signals from one zone can leak into other zones (crosstalk), causing intelligibility and/or quality degradation. An optimization framework is built by exploring a signal model where effects of noise, reverberation and zone crosstalk are taken into account simultaneously. Through the symbolic usage of a general smooth distortion measure, necessary optimality conditions are derived in terms of distortion measure gradients and the signal model. Subsequently, as an illustrative example of the framework, the conditions are applied for the mean-square error (MSE) expected distortion under a hybrid stochastic-deterministic model for the corruptions. A crosstalk cancellation algorithm follows, which depends on diffuse reverberation and across zone direct path components. Simulations validate the optimality of the algorithm and show a clear benefit in multizone processing, as opposed to the iterated application of a single-zone algorithm. Also, comparisons with least-squares crosstalk cancellers in literature show the profit of using a hybrid model.**

***Index Terms*—Near-end listening enhancement, speech reinforcement, multizone, public address system.**

I. Introduction

RECENTLY, the field of near-end (source-based) listening enhancement, also termed *speech reinforcement*, has gained increasing interest in the research community. While traditional speech enhancement systems apply a time-frequency weighting to a *received* noisy speech signal to enhance speech components with respect to the noise [1], *source-based* systems (*e.g.*, [2]–[6]) apply the weighting at a clean speech source in the hope that, when played back in—and corrupted by some acoustic communication channel, degradation is minimized at the listener. Examples of applications which could benefit from speech reinforcement range from mobile telephones or hearing aids to conference or public address (PA) systems.

In general terms, speech reinforcement works as depicted in Fig. 1. A certain source speech signal (*e.g.*, public announcement) is pre-processed ("reinforced") before being played back in a corruptive acoustic channel, and is listened to by a receiver submerged in the environment. The channel can be modeled convolutive, *e.g.*, if the environment is reverberant, and/or noise-additive for the situation that there are noise sources in the vicinity of the listener. Also, the reinforcement algorithm can take advantage of knowledge about the channel through the measurement of its properties (*e.g.*, noise spectral densities or reverberation parameters).

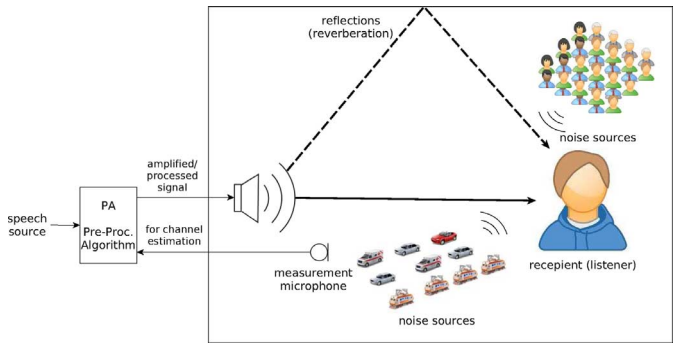

Fig. 1. Speech reinforcement concept.

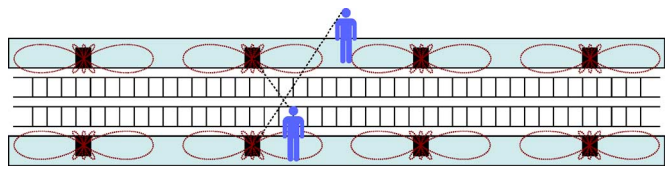

Fig. 2. Two-zone speech reinforcement system example (train station). Speech in one platform can leak into the other. Black boxes are loudspeakers, with corresponding directivity diagrams.

Current source-based systems only consider a single playback region where some kind of speech reinforcement is applied. However, many practical scenarios consist of multiple regions or zones, *e.g.*, consider public addressing in airports, train stations or shopping malls. In such a multizone situation, signals played back in one region can interfere with other regions, a phenomenon which we call acoustic leakage or *crosstalk*. To illustrate how this phenomenon can be a source of nuisance, consider a two-zone public address system installed on two platforms of a train station (Fig. 2), with some reciprocal crosstalk between zones. Consider also a conventional single-zone signal recovery strategy in each zone (*e.g.*, [2]). In this scenario, each zone working autonomously will potentially consider the speech coming from the other zone as noise, trying to amplify its own speech such as to mask the speech from the other zone. Due to the direct (linear) relation between playback level and leakage, a competition effect arises where each zone always tries to amplify even more than the other zone up to the level that the reinforced signals saturate. The instable positive feedback situation that arises hereby motivates the importance of the study of reinforcement taking multiple zones into account. Note that, although an ideal noise estimator would be designed not to detect the crosstalk speech as noise, a practical noise estimator will always provide for speech leakage into the noise estimate (*e.g.*, due to false negatives in voice

Manuscript received April 02, 2013; revised July 04, 2013, September 13, 2013; accepted September 16, 2013. Date of publication September 23, 2013; date of current version November 13, 2013. This work was supported in part by the Dutch Technology Foundation STW and Bosch Security Systems B.V., The Netherlands. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rongshan Yu.

The authors are with the Signal and Information Processing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: j.b.crespo@tudelft.nl; r.c.hendriks@tudelft.nl).

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

Digital Object Identifier 10.1109/TASL.2013.2283100

activity detection [1]), and this leakage would get aggravated in a multizone situation.

In contrast to the sub-optimal usage of independently working conventional single zone schemes with their potential issues as described above, we aim to pre-process all speech signals from the different zones jointly. As far as the authors' knowledge goes, multizone speech reinforcement as described in this paper has not been previously studied, and constitutes the main novelty of this paper.

Also, many speech reinforcement schemes are designed taking empirical and/or heuristic considerations into account without a formal quantitative methodology for pre-processing (*e.g.*, [2], [3], [7]–[9]). In contrast, in our approach, we work by quantifying the degradation between source and received speech at the listener by a general functional measure d, dubbed as the *distortion measure*. The measure could either model quality or intelligibility degradation by making use of an adequate model which, *e.g.*, could come from a psycho-acoustical model of distortion detectability [10]–[12] or from some speech intelligibility model [13]–[15]. See [16] for a comparative overview on some distortion and intelligibility measures.

Another observation about current speech reinforcement algorithms is that they either disregard explicit channel information [17], [8], [7], [9], or the dependence is only on noise, not taking any convolutive (reverberant) effects into account [18], [2]–[6]. Although some work has also been done on reverberant channels [19]–[21], the so-called "reverberation pre-processing" algorithms, that work does in turn not take the effect of noise into account. In a multiple source/receiver setting, crosstalk cancellation schemes (*e.g.*, [22]–[25]) also fail to take noise into account. For an overview on all of these algorithms, see Section II. In opposition to the described partial contributions, we synergize effects of reverberation, crosstalk and noise by developing a framework for multizone speech reinforcement with a signal model which considers the three effects simultaneously. On the basis of a suitable quantification of speech distortion (and/or intelligibility), we undergo a mathematical optimization problem, where we search for the optimal joint processing scheme for all zones which minimizes the expected global distortion, also measured jointly for all zones. We follow an abstract methodology, by splitting the solution of the minimization problem in two steps. In the first place, we derive general necessary conditions for the optimality of *any* smooth distortion measure. These conditions are given in terms of the acoustic channel from the proposed signal model, and of gradients of the distortion measure. Subsequently, we apply the derived conditions to a particular distortion measure. This two-step methodology has the advantage of reusability, *i.e.*, given the abstract conditions, several choices of distortion measures can be made, each one delivering a different algorithm optimized to the characteristics of its own distortion model. Note that algorithms based on optimization of a merit or cost figure (as here described) have been studied in the past [18], [5], [6], but these schemes do not use an abstract functional formulation (distortion measure d) and/or do only consider a single zone.

For the sake of simplicity, and to demonstrate the use of the general derivations, we choose the ℓ_2 error distortion measure (euclidean distance) for concretizing the general optimality conditions in our work. Thereby, we derive a pre-processing scheme using a hybrid deterministic-stochastic description of the acoustic channel. The scheme is an acoustic crosstalk cancellation scheme, and is given in terms of direct-path compensating terms which are perturbed by a stochastic component modeling late reverberation. In the limit case of low reverberation, the scheme boils down to pure direct path compensation as in conventional crosstalk cancellers [26]. Although the simple scheme derived takes a form similar to known crosstalk cancellers [26], [24], the problem setting devised here is structurally different. Indeed, in crosstalk cancellation, one is faced with a filter design problem, where an overall impulse response should be matched to a desired response [23], whereas in our work, we are minimizing distortion between source and reproduction signals by means of our multizone framework. Due to the observed relation between multizone speech reinforcement and crosstalk cancellation, we review work in the latter subject in Section II-C.

Also, the derived algorithm will be seen to be independent of noise statistics (see Section V), not making full use of the synergetic framework described above. Nevertheless, we would like to stress that this fact is dictated by the simple nature of the distortion measure chosen for application in the general framework, and that by choosing an adequate (more complex) distortion measure, noise-dependent algorithms can be derived (see Section VII for a discussion).

The paper is organized as follows. In Section II, we review previous work on speech reinforcement, reverberation pre-processing and on the related subject of acoustic crosstalk cancellation. Section III constructs the mathematical framework for multizone reinforcement that will be used throughout the paper. Section IV contains the core of the work; here, we derive necessary optimality conditions for the solution of the optimization problem motivated in the introduction and formally stated in Section III. Section V applies the derived conditions to the ℓ_2-error distortion measure with a hybrid deterministic-stochastic channel model, resulting in a crosstalk cancellation algorithm. In Section VI, we present simulations validating the algorithm of Section V. Section VII reflects upon the developed theory, points out some limitations thereof and states some challenges for future work. Finally, Section VIII concludes the article.

II. Related Work

A. Speech Reinforcement for the Noisy Channel

Pioneering research on source-based speech processing with the aim of bridging intelligibility degradations concentrated on simple degradations and pre-processing. The algorithms were based on empirical studies that concluded that higher frequencies and the frequently thereto associated consonant components of speech are more important for intelligibility than low frequencies and vowels, respectively [27]–[29]. Within the noisy channel context (without reverberation), Thomas and Niederjohn [30] introduced a speech reinforcement system consisting of a high pass filter followed by an infinite clipper. Griffiths [18] derived a linear filter which was optimal with

respect to the Articulation Index [27], [31] objective intelligibility measure under power constraints. The filter had a high-pass characteristic which, independently of noise statistics, whitened processed speech. Also in a power constrained setting, Niederjohn and Groteluesch en [17] introduced dynamic range compression for intelligibility enhancement. Indeed, it can be noted that these schemes were designed to shift the energy of speech upwards in frequency (through high-pass filtering), and to reinforce consonants (through the additional usage of compressive mechanisms).

More recently, Hall and Flanagan [8] revisited the problem, applying comparable high-pass filtering techniques for speech in babble noise. Also, Chanda and Park [32] proposed the usage of a time-variant high-pass filter which provided for an initial consonant boost on a vowel-consonant transition. Skowronski and Harris [7] applied consonant-vowel energy ratio boosting by means of a voicing detector, in turn based on spectral flatness thresholding. The boosting ratio was chosen using empirical considerations.

The empirical observation that transient components in speech are more important than stationary ones [33], [34] led to algorithms enhancing transients. Yoo *et al.* [9] used a high-pass filter followed by a signal splitter in transient and quasi-steady state (QSS) components, boosting the former components with respect to the latter by a heuristically determined amount.

In parallel, still within the noisy channel context, researchers focus also on time-frequency (TF) weighting algorithms. In [2], Sauert and Vary studied reinforcement in a power unconstrained setting. They proposed heuristically motivated weightings which perform SNR recovery, by setting the enhanced speech power spectrum at a certain fixed log distance from the noise power spectrum. Shin *et al.* [35] proposed a similar idea, where instead of the SNR, they recover the perceptual loudness in each frequency. In [3], Sauert and Vary extended their previous work into the power constrained setting by proposing a frame-by-frame normalization procedure. With this method, redistribution of energy across frequency is possible. Also, a new heuristic weighting was introduced where, in opposition to the SNR recovery case, frequency bands corresponding to high SNR are reinforced in exchange for low SNR bands. The new approach was motivated by modeling cognitive processing of speech in the human brain through the usage of a Wiener filter as a pre-processing step. The proposed TF weightings were simulated and validated using the Speech Intelligibility Index (SII) [36].

A less empirical/heuristic and more mathematically driven methodology has been employed in the work hereafter. In [4] and [5], Sauert and Vary optimized the SII in a TF weighting framework in the power unconstrained and constrained scenarios, respectively. The authors showed that the new algorithms outperform the empirical algorithms previously proposed by them. Finally, Taal *et al.* [6] developed a TF weighting working in an internal auditory domain which redistributes energy in time and frequency, by optimizing a perceptual distortion measure [11] analytically. The algorithm inherits the short-time sensitivity of the distortion measure it is optimized for, resulting in amplified transients.

B. Reverberation Pre-Processing

All schemes described above concern the case where speech reinforcement is applied to increase robustness with respect to corruption in a noisy channel without reverberation effects. Although not as extensively as in the noisy channel case, some work has also been done for the reverberant channel. As mentioned in Section I, this work does not take noise corruptions into account. Langhans and Strube [37] introduced modulation filtering to pre-process reverberant speech. There, the speech signal is filtered in such a way that the signal power in critical bands gets enhanced by a fixed transfer function. However, no modulation depth increase could be observed after reverberation. Kusumoto *et al.* [19] applied heuristically designed and data-derived modulation filters to assess whether intelligibility could be increased in reverberant conditions.

Besides modulation filtering, steady-state suppression has been proposed for source-based reverberation processing. The rationale behind this kind of processing is that intelligibility degradation occurs in reverberated speech when the reverberant tails of high energy steady-state components mask subsequent low-energy regions. By decreasing the energy of these steady-state regions, masking can be commensurably decreased and intelligibility is recovered. Hodoshima *et al.* [20] explored this idea using a steady state suppressor, which thresholds a measure based on the sum of squares of linear regression slopes of multiband speech envelopes.

Finally, reverberation pre-processing schemes have been proposed on the basis of optimization procedures. In [38], the authors shaped the global system response (reverberant channel composed with a pre-processing filter) to a desired response by minimizing their ℓ_2 distance. In [21], an ℓ_p-norm based approach was undertaken, with an objective function given by a (log) ratio of undesired to desired global impulse response segments.

C. Acoustic Crosstalk Cancellation

In the following, we overview work on the subject of acoustic crossstalk cancellation, which is a topic related to multizone reinforcement. In crosstalk cancellation, the aim is to pre-filter and mix a set of multiple sources, such that when the sources are corrupted by a convolutive mixing channel (similarly to the multizone case, see Section III), each source gets through to its corresponding destination without crosstalk (signals from other sources). The idea was introduced for the two-source/receiver case by Atal and Schroeder [39], where two loudspeakers should deliver two signals to the ears of a listener separately to create a virtual 3D sound impression. Basic schemes work with the pure deterministic inverse of the channel transfer, which is frequently modeled as a direct-path delay [26]. Nevertheless, this kind of schemes is inherently unrobust with respect to loudspeaker-listener positioning, only working in a small listening region ("sweet spot"). To bridge this difficulty, in [22], the authors derive optimal loudspeaker positions which minimize the condition number of the channel transfer matrix.

Subsequent work concentrates on robust filter design techniques which minimize some distance measure between desired

(multichannel) filter responses and the global response of the system. In [23], a least-squares approach was proposed for this purpose. Optimum filters were derived for a single head position and taking multiple positions into account by spatial averaging. Kallinger and Mertins [24] model the channel by including a stochastic perturbation term, deriving thereby optimal filter coefficients in the squared sense. In addition to traditional ℓ_2-norm based distances, other cost criteria have also been used in designing the filter coefficients. Examples are approaches using regularized least-squares [40], approaches based on the ℓ_∞-norm [41] (minimax approaches) and based on the more general ℓ_p-norm [42]. Also, this last work combines crosstalk cancellation with the reverberation pre-processing strategies of [21] (described in Section II-B), and it was extended in [25] to include stochastic perturbations and a regularized objective function.

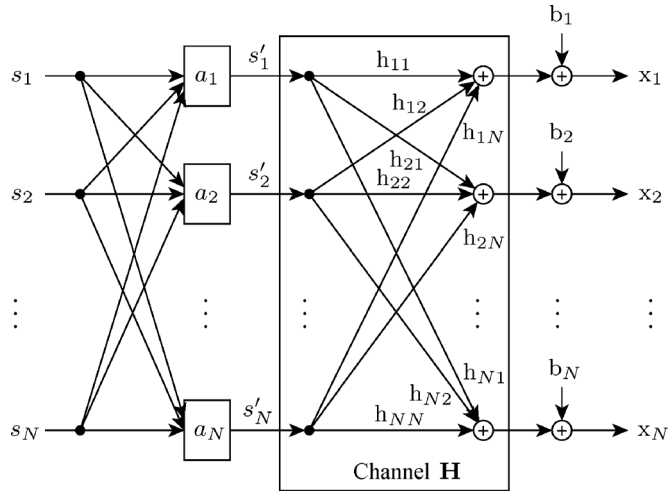

Fig. 3. Multizone speech reinforcement model for fixed frequency f (index omitted). The channel $\mathbf{H}(f)$ is a mixing matrix of filters $\mathrm{h}_{ij}(f)$, $i, j \in \{1, 2, \ldots, N\}$.

III. Preliminaries

In this section, we build an optimization framework for speech reinforcement based on an affine signal model and an unconstrained optimization problem, where the objective function is given by the expected value of a general real-valued smooth (continuously differentiable) distortion measure.

Beginning with notation, $\stackrel{\text{def}}{=}$ will be used for defining functions and operators. For scalars we will use lowercase regular font, whereas for vectors, we use lowercase bold and for matrices uppercase bold letters. A vector of size U containing ones in all entries will be denoted by $\mathbf{1}_U$. The ℓ_2 norm will be denoted by $\|\cdot\|$. Furthermore, we denote vector/matrix transposition, conjugation and conjugate transposition by $(\cdot)^{\mathrm{T}}$, $(\cdot)^*$ and $(\cdot)^{\mathrm{H}}$, respectively. Indexing vector and matrix expressions will be done using the notation $[\cdot]_i$ and $[\cdot]_{ij}$, respectively, where i (resp. ij) is the index. The operator $\mathrm{diag}(\cdot, \cdot, \ldots)$ builds a (block) diagonal matrix using the argument scalars/matrices as diagonal entries. For a vector input $\boldsymbol{v} = [v_1, v_2, \ldots, v_N]^{\mathrm{T}}$, $\mathrm{diag}\,\boldsymbol{v}$ is understood as $\mathrm{diag}(v_1, v_2, \ldots, v_N)$. Also, we notate random variables by upright letters (*e.g.*, v, $\mathbf{v}$) and their realizations by their slanted equivalent $(v, \boldsymbol{v})$. Deterministic variables are also notated slanted. The imaginary unit will be denoted by $\jmath$. As to differential calculus, we will use the notation $\partial/\partial\boldsymbol{v}$ to denote the transposed Jacobian matrix of a multivariate function with respect to vector argument $\boldsymbol{v}$, using thus the Hessian formulation for differentiation (gradients as column vectors).

As motivated in Section I, we consider a speech reinforcement scenario working across multiple zones, say $N \in \{1, 2, 3, \ldots\}$ zones. We consider frame-based signal representations in the discrete Fourier transform (DFT) domain. These could come, *e.g.*, in the context of a ubiquitous DFT-based speech processing scheme. The signal processing flow we consider is shown in Fig. 3. We depart from the clean speech signal. Denote the f-th DFT frequency bin of the clean speech of zone i by $s_i(f) \in \mathbb{C}$, $f \in \{0, 1, \ldots, L-1\}$, $i \in \{1, 2, \ldots, N\}$, where L is the DFT size. For reasons of notational convenience in writing down the signal model of Fig. 3, we consider a vector of clean speech stacked up for all zones for each fixed frequency f, denoted by $\boldsymbol{s}(f) \in \mathbb{C}^N$, and given by

$$\boldsymbol{s}(f) = [s_1(f), s_2(f), \ldots, s_N(f)]^{\mathrm{T}}. \tag{1}$$

For a convenient compact notation, we will further pack the "per-frequency" vectors $\boldsymbol{s}(f)$ together for all frequencies f, defining the joint clean speech vector for all zones and frequency bins $\boldsymbol{s} \in \mathbb{C}^{LN}$ by

$$\boldsymbol{s} = [\boldsymbol{s}(0)^{\mathrm{T}}, \boldsymbol{s}(1)^{\mathrm{T}}, \ldots, \boldsymbol{s}(L-1)^{\mathrm{T}}]^{\mathrm{T}}. \tag{2}$$

We also note that this model supports a source signal which is the same for all zones (single source broadcast), in which case we have $\boldsymbol{s}(f) = s(f)\,\mathbf{1}_N$ for some single-zone source speech DFT coefficient $s(f)$.

The N sources packed in $\boldsymbol{s}$ get jointly processed by N pre-processing functions (functions $a_1, a_2, \ldots, a_N$ in Fig. 3), thereby producing N pre-processed signals, denoted by $\boldsymbol{s}'(f)$ and $\boldsymbol{s}'$, in analogy to (1) and (2), respectively. In general, the processed speech $\boldsymbol{s}'$ is a function of the clean speech signal $\boldsymbol{s}$ and of noise and channel statistics. The source and processed speech will be modeled to be deterministic along our analysis, due to the fact that in source-based speech processing, the source realizations are directly available, without the necessity for an estimation process.

The received signal in each zone is then modeled as a combination of appropriately weighted signals from all (pre-processed) zone sources, plus a local noise term, as can be seen in Fig. 3. More formally, as we did for the pre-processed speech, we define the received speech $\mathbf{x}(f)$, $\mathbf{x}$ and additive noise term $\mathbf{b}(f)$, $\mathbf{b}$ in analogy to (1) resp. (2). These are modeled as stochastic processes, where we assume a zero-mean behavior of the noise process $\mathbf{b}(f)$ for all frequencies f. As to the acoustic channel transfer between pre-processed and received speech, also modeled as a stochastic process, we denote the f-th DFT frequency bin of the transfer function between zone j and zone i by $\mathrm{h}_{ij}(f)$, where j is the zone where the speech is played back and i is the target (reception) zone. In analogy to (1), we work on

a per-frequency notation by fixing frequency component f and collecting all frequency response values of the transfers $\mathrm{h}_{ij}(f)$, $i,j \in \{1,2,\ldots,N\}$, in an N-by-N matrix $\mathbf{H}(f)$:

$$\mathbf{H}(f) = \begin{bmatrix} \mathrm{h}_{11}(f) & \mathrm{h}_{12}(f) & \ldots & \mathrm{h}_{1N}(f) \\ \mathrm{h}_{21}(f) & \mathrm{h}_{22}(f) & \ldots & \mathrm{h}_{2N}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{h}_{N1}(f) & \mathrm{h}_{N2}(f) & \ldots & \mathrm{h}_{NN}(f) \end{bmatrix}. \quad (3)$$

As presented in Section I, effects of reverberation, crosstalk and noise can be included simultaneously in our framework. This is done by adequately modeling the components $\mathrm{h}_{ii}(f)$ for reverberation, $\mathrm{h}_{ij}(f), j \neq i$ for crosstalk, and $\mathrm{b}_i(f)$ for additive noise effects. In addition, for compacting the channel information $\mathbf{H}(f)$ for all f, we define the joint channel matrix $\mathbf{H}$ by the block-diagonal matrix

$$\mathbf{H} = \mathrm{diag}[\mathbf{H}(0), \mathbf{H}(1), \ldots, \mathbf{H}(L-1)]. \quad (4)$$

Using the symbols introduced, $\mathbf{x}(f)$ is formally modeled by (see Fig. 3)

$$\mathbf{x}(f) = \mathbf{H}(f)\,\boldsymbol{s}'(f) + \mathbf{b}(f), \quad (5)$$

for all frequencies $f \in \{0,1,\ldots,L-1\}$. In this affine model, the matrix product

$$[\mathbf{H}(f)\,\boldsymbol{s}'(f)]_i = \sum_{j=1}^{N} \mathrm{h}_{ij}(f)\, s'_j(f) \quad (6)$$

models the mixing operation occurring from all zones j into a specific zone i, and each term in the matrix product $\mathrm{h}_{ij}(f)\, s'_j(f)$ models convolution, which got mapped into a product in the frequency domain. The noise term $\mathbf{b}(f)$ on the right hand side of (5) reflects the local additive noise model which was postulated. A more compact notation of (5) reads

$$\mathbf{x} = \mathbf{g}(\boldsymbol{s}') \stackrel{\mathrm{def}}{=} \mathbf{H}\,\boldsymbol{s}' + \mathbf{b}, \quad (7)$$

where the definitions of (2) (and extensions) and (4) were used.

The aim of our work is then, given a mathematical description of an overall distortion measure $d(\boldsymbol{s},\boldsymbol{x})$, to find the optimally preprocessed signal $\boldsymbol{s}'$ which minimizes the expected distortion. We assume the distortion measure to be real-valued and continuously differentiable (in class C^1) when viewed as a (real) function of the $2LN$ real variables unraveled by $\boldsymbol{x}$, obtained by taking its real and imaginary parts $\boldsymbol{x}_R$ and $\boldsymbol{x}_I$, respectively. See Section VII for a discussion on the choice of this category of distortion measures. In a mathematical formulation, we want to find the minimizer of the optimization problem

$$\min_{\boldsymbol{s}' \in \mathbb{C}^{LN}} \mathrm{E}[d(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}'))], \quad (8)$$

where the affine function $\mathbf{g}(\,\cdot\,)$ is the one of (7).

IV. Optimality Conditions

In this section, we derive necessary conditions for the processed speech $\boldsymbol{s}'$ to solve the problem expressed in (8). We arrive at expressions given in terms of expectations of gradients of the distortion measure and of the acoustical channel.

To do so, we first consider the function $D : \mathbb{R}^{LN} \times \mathbb{R}^{LN} \mapsto \mathbb{R}$ defined by $D(\boldsymbol{s}'_R, \boldsymbol{s}'_I) \stackrel{\mathrm{def}}{=} \mathrm{E}[d(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}'_R + \jmath\boldsymbol{s}'_I))]$, *i.e.*, the objective function of (8) taken as a function of the real and imaginary components of $\boldsymbol{s}'$. We note that if $(\boldsymbol{x}_R, \boldsymbol{x}_I) \mapsto d(\boldsymbol{s}, \boldsymbol{x}_R + \jmath\boldsymbol{x}_I)$ is in $C^1(\mathbb{R}^{LN} \times \mathbb{R}^{LN})$, then so is D, since it is an integral of C^1 functions of $(\boldsymbol{x}_R, \boldsymbol{x}_I)$, corresponding to the expectation operator $\mathrm{E}[\cdot]$ worked out, composed with a C^1 signal model $(\boldsymbol{s}'_R, \boldsymbol{s}'_I) \mapsto (\boldsymbol{x}_R, \boldsymbol{x}_I)$, corresponding to the function $\mathbf{g}$ in (7). Furthermore, it is also known from calculus, that if $(\boldsymbol{s}'_R, \boldsymbol{s}'_I)$ is a locally optimal point of D (*i.e.*, if it maximizes or minimizes D in an epsilon neighborhood), then its gradient should vanish:

$$\begin{aligned} \frac{\partial D}{\partial \boldsymbol{s}'_R}(\boldsymbol{s}'_R, \boldsymbol{s}'_I) &= \frac{\partial}{\partial \boldsymbol{s}'_R} \mathrm{E}\left[d\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}'_R + \jmath\boldsymbol{s}'_I)\right)\right] = 0 \\ \frac{\partial D}{\partial \boldsymbol{s}'_I}(\boldsymbol{s}'_R, \boldsymbol{s}'_I) &= \frac{\partial}{\partial \boldsymbol{s}'_I} \mathrm{E}\left[d\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}'_R + \jmath\boldsymbol{s}'_I)\right)\right] = 0. \end{aligned} \quad (9)$$

By introducing the complex differential operators ([43], Ch. 13, Sec. 2)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{s}'} &\stackrel{\mathrm{def}}{=} \frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{s}'_R} + \frac{1}{\jmath}\frac{\partial}{\partial \boldsymbol{s}'_I}\right) \\ \frac{\partial}{\partial \boldsymbol{s}'^*} &\stackrel{\mathrm{def}}{=} \frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{s}'_R} - \frac{1}{\jmath}\frac{\partial}{\partial \boldsymbol{s}'_I}\right) \end{aligned} \quad (10)$$

and weighting and combining (9) accordingly, we arrive at the equivalent conditions

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{s}'} \mathrm{E}\left[d\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] &= 0 \\ \frac{\partial}{\partial \boldsymbol{s}'^*} \mathrm{E}\left[d\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] &= 0, \end{aligned} \quad (11)$$

where $\boldsymbol{s}' = \boldsymbol{s}'_R + \jmath\boldsymbol{s}'_I$. As the distortion measure $d(\boldsymbol{s},\boldsymbol{x})$ is a real-valued (complex argument) function, it follows from the property in ([43], Ch. 13, Sec. 2.3(b)) that the two branches in (11) are equivalent. In practice, this means that we only have to solve for one of the conditions of (11), since the other condition will be automatically satisfied.

We now work out the left-hand side of (11) (lower branch). We get the succession of equalities

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{s}'^*} \mathrm{E}\left[d\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] &= \mathrm{E}\left[\frac{\partial d}{\partial \boldsymbol{s}'^*}\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] && (12) \\ &= \mathrm{E}\left[\frac{\partial \mathbf{g}}{\partial \boldsymbol{s}'^*}(\boldsymbol{s}') \cdot \frac{\partial d}{\partial \boldsymbol{x}}\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right. \\ &\quad \left. + \frac{\partial \mathbf{g}^*}{\partial \boldsymbol{s}'^*}(\boldsymbol{s}') \cdot \frac{\partial d}{\partial \boldsymbol{x}^*}\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] && (13) \\ &= \mathrm{E}\left[\left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{s}'}(\boldsymbol{s}')\right)^* \frac{\partial d}{\partial \boldsymbol{x}^*}\left(\boldsymbol{s}, \mathbf{g}(\boldsymbol{s}')\right)\right] && (14) \\ &= \mathrm{E}\left[\mathbf{H}^{\mathrm{H}} \frac{\partial d}{\partial \boldsymbol{x}^*}\left(\boldsymbol{s}, \mathbf{x}\right)\right]. && (15) \end{aligned}$$

In (12), we assume that the tail of the probability density function (PDF) of $\mathbf{g}$ converges to zero fast enough so that we can exchange differentiation and integration (expectation) order; in (13) we use the chain rule for the complex differential operators in question ([43], Ch. 13, Sec. 2.2); in (14), we use the fact that $\mathbf{g}(\boldsymbol{s}')$ is a complex analytical function to state ([43], Ch. 13, Sec. 2.1) $\partial \mathbf{g}/\partial \boldsymbol{s}'^* = 0$, letting thereby the first summand of (13)

vanish, and we use further properties of the complex differential operators ([43], Ch. 13, Sec. 2.3(b)); (15) uses (7) and computes the (transposed) Jacobian matrix $\partial \mathbf{g}/\partial \boldsymbol{s}'$.

If we compare (15) with (11) (lower branch), we find out that the necessary conditions for a pre-processed speech vector $\boldsymbol{s}'$ to optimize the expected value of the distortion measure $d(\boldsymbol{s}, \boldsymbol{x})$ are

$$\mathrm{E}\left[\mathbf{H}^{\mathrm{H}} \frac{\partial d}{\partial \boldsymbol{x}^*}(\boldsymbol{s}, \mathbf{x})\right] = 0. \tag{16}$$

By taking the block diagonal structure of $\mathbf{H}$ in (4) into account, we can write (16) using the per-frequency gradient entries as

$$\mathrm{E}\left[\mathbf{H}(f)^{\mathrm{H}} \frac{\partial d}{\partial \boldsymbol{x}(f)^*}(\boldsymbol{s}, \mathbf{x})\right] = 0, \quad \forall f \in \{0, 1, \ldots, L-1\}. \tag{17}$$

In words, optimality of the pre-processed speech $\boldsymbol{s}'$ is achieved when, for each frequency bin f, the complex gradient vector of the distortion measure with respect to the received DFT bins in all zones, $\partial d/\partial \boldsymbol{x}(f)^*$, is orthogonal to all columns of the channel matrix $\mathbf{H}(f)$.

We can simplify (17) if we make further assumptions with respect to the distortion measure. Indeed, if we assume the distortion measure to be additive over frequency, *i.e.*, if we take a distortion measure of the form

$$d(\boldsymbol{s}, \boldsymbol{x}) = \sum_{\nu=0}^{L-1} d'(\boldsymbol{s}(\nu), \boldsymbol{x}(\nu)) \tag{18}$$

for some intermediate distortion measure d' operating on the per-frequency sub-variables, all terms but the one with $\nu = f$ vanish while differentiating in (17). The condition simplifies to

$$\mathrm{E}\left[\mathbf{H}(f)^{\mathrm{H}} \frac{\partial d'}{\partial \boldsymbol{x}(f)^*}(\boldsymbol{s}(f), \mathbf{x}(f))\right] = 0. \tag{19}$$

V. Application Example

We now apply the optimality condition derived in Section IV to a specific distortion measure, namely, to the ℓ_2 error (euclidean distance) between the clean and received complex speech DFT coefficients. After working out the optimality condition, we model the channel present in the resulting equation as a hybrid deterministic-stochastic process. The deterministic component models the direct path in each zone transfer, whereas the stochastic component models the production of a late diffuse sound field.

Consider the distortion measure given by

$$d(\boldsymbol{s}, \boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{s}\|^2 = \sum_{f=0}^{L-1} \sum_{i=1}^{N} |x_i(f) - s_i(f)|^2, \tag{20}$$

where $\boldsymbol{s}$ and $\boldsymbol{x}$ collect all the frequency and zones of the clean and received speech signals as in (2). It is easy to see that the measure of (20) is additive over frequency; it is of the form of (18) with

$$d'(\boldsymbol{s}(f), \boldsymbol{x}(f)) = \|\boldsymbol{x}(f) - \boldsymbol{s}(f)\|^2. \tag{21}$$

If we calculate its complex gradient vector with respect to $\boldsymbol{x}(f)$ and apply (19), we get the raw form of the optimality condition, namely

$$\mathrm{E}\left[\mathbf{H}(f)^{\mathrm{H}}(\mathbf{x}(f) - \boldsymbol{s}(f))\right] = 0. \tag{22}$$

We can now work out the optimality condition of (22) by using the signal model of (5). We get

$$\mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)]\, \boldsymbol{s}'(f) + \mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{b}(f)] = \mathrm{E}[\mathbf{H}(f)]^{\mathrm{H}}\, \boldsymbol{s}(f). \tag{23}$$

Finally, we make the assumption that all zone transfers are uncorrelated with the noise DFT bins for all zones, and argue that for a non-degenerate distribution of $\mathbf{H}(f)$ for all f, the correlation matrix $\mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)]$ is positive definite. With these assumptions, the correlation matrix is invertible and the processed signal can be given as

$$\boldsymbol{s}'(f) = \mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)]^{-1}\mathrm{E}[\mathbf{H}(f)]^{\mathrm{H}}\, \boldsymbol{s}(f). \tag{24}$$

As can be seen in (24), the MSE optimally pre-processed speech signal is a stochastic pseudoinverse-like solution, given in terms of first and second order moments of the convolutive channel $\mathbf{H}(f)$. Also, due to the assumption that the noise and channel terms are uncorrelated, the solution does not depend on noise statistics.

The MSE optimal algorithm of (24) is abstract in the sense that no assumptions have been made what respects to the form of the channel $\mathbf{H}(f)$. In the following, we deduce a concrete reinforcement algorithm by incorporating a parametric hybrid deterministic-stochastic model in the channel, and subsequently giving the abstract algorithm of (24) in terms of the parameters.

Specifically, we model the impulse response of the zone transfer from zone j to zone i as

$$\check{\mathrm{h}}_{ij}[n] = \Delta_{ij}\delta[n - n_{\Delta_{ij}}] + a_{ij}^{\,n-n_{\mathrm{w}_{ij}}}\, u[n - n_{\mathrm{w}_{ij}}]\, \mathrm{w}_{ij}[n - n_{\mathrm{w}_{ij}}], \tag{25}$$

where $\Delta_{ij} \in \mathbb{R}$ is the direct path amplitude, $a_{ij} \in [0, 1)$ is a decaying exponential parameter, $u[\cdot]$ is the discrete Heaviside step function, $n_{\Delta_{ij}}, n_{\mathrm{w}_{ij}} \in \mathbb{N}_0$ are delays satisfying $n_{\Delta_{ij}} < n_{\mathrm{w}_{ij}}$, and w_{ij} is a real white stationary random process with first and second order moments given by

$$\mathrm{E}\{\mathrm{w}_{ij}[n]\} = 0, \quad \mathrm{E}\{\mathrm{w}_{ij}^2[n]\} = \sigma_{ij}^2, \quad \mathrm{E}\{\mathrm{w}_{ij}[n]\mathrm{w}_{kl}[m]\} = 0 \tag{26}$$

for all $i, j, k, l \in \{1, 2, \ldots, N\}$, $n, m \in \mathbb{Z}$, and $(i, j, n) \neq (k, l, m)$. In words, we model each impulse response between two zones as a sum of a deterministic direct path response with a stochastic response corresponding to a diffuse sound field. We thus simplify the early impulse response by neglecting early reflections and the late response by assuming that all reflections are stochastically described by an idealized response, similar to the Polack model [44]. Furthermore, we also assume this idealized response to be uncorrelated in time and across zones. See also Fig. 4 for a schematic picture of the impulse response.

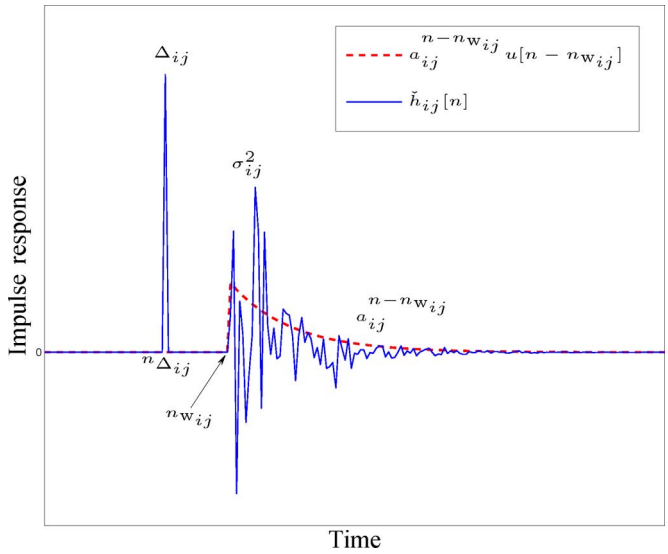

Fig. 4. Impulse response between zones (example).

To perform the computation of the channel matrix components $[\mathbf{H}(f)]_{ij}$ for all zones $i,j \in \{1,2,\ldots,N\}$ and frequencies $f \in \{0,1,\ldots,L-1\}$, we transform (25) into the frequency domain using the continuous Discrete-Time Fourier Transform (DTFT) $\mathrm{h}_{\mathrm{DTFT},ij}(\omega)$, and sample the DTFT at an evenly spaced grid $\omega = 2\pi f/L$. The vector dimension L will be considered large enough so that time-domain aliasing has negligible effect on the model. The result is given by

$$[\mathbf{H}(f)]_{ij} = \mathrm{h}_{ij}(f) = \Delta_{ij} e^{-\jmath \frac{2\pi f}{L} n_{\Delta_{ij}}} + e^{-\jmath \frac{2\pi f}{L} n_{\mathrm{w}_{ij}}} \left[\sum_{n=0}^{\infty} a_{ij}^n \, \mathrm{w}_{ij}[n] e^{-\jmath \frac{2\pi f}{L} n} \right]. \quad (27)$$

We are now ready to compute the first and second order moments $\mathrm{E}[\mathbf{H}(f)]$ and $\mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)]$, respectively, by simple stochastic manipulation rules. As $\mathrm{w}_{ij}[n]$ is zero-mean, (26), we obtain

$$\mathrm{E}[\mathbf{H}(f)]_{ij} = \Delta_{ij} e^{-\jmath \frac{2\pi f}{L} n_{\Delta_{ij}}}. \quad (28)$$

For the second order moments, we have

$$\mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)]_{ij} = \sum_{k=1}^{N} \mathrm{E}\left\{ [\mathbf{H}(f)]_{ki}^{*} \, [\mathbf{H}(f)]_{kj} \right\} \quad (29)$$

$$= \sum_{k=1}^{N} \Delta_{ki}\Delta_{kj} e^{-\jmath \frac{2\pi f}{L} \left(n_{\Delta_{kj}} - n_{\Delta_{ki}}\right)} + \delta_{ij} \sum_{k=1}^{N} \sum_{n=0}^{\infty} a_{ki}^{2n} \, \sigma_{ki}^2 \quad (30)$$

$$= \sum_{k=1}^{N} \Delta_{ki}\Delta_{kj} e^{-\jmath \frac{2\pi f}{L} \left(n_{\Delta_{kj}} - n_{\Delta_{ki}}\right)} + \delta_{ij} \sum_{k=1}^{N} \frac{\sigma_{ki}^2}{1-a_{ki}^2}. \quad (31)$$

In (30), we use (27), linearity of the expectation operator, the moments of (26) and the Kronecker delta notation δ_{ij} equal to one when $i = j$ and equal to zero when $i \neq j$; in (31), we compute the sum of the geometric series of base a_{ki}^2. In matrix notation, the first and second moments are given by

$$\mathrm{E}[\mathbf{H}(f)] = \boldsymbol{D}(f) \quad (32)$$

$$\mathrm{E}[\mathbf{H}(f)^{\mathrm{H}}\mathbf{H}(f)] = \boldsymbol{D}(f)^{\mathrm{H}}\boldsymbol{D}(f) + \mathrm{diag}\,\boldsymbol{r}, \quad (33)$$

where $\boldsymbol{D}(f)$ is the direct path matrix, with each component equal to the direct path transfer function of the corresponding cross-channel (zone pair), *i.e.*,

$$[\boldsymbol{D}(f)]_{ij} = \Delta_{ij} e^{-\jmath \frac{2\pi f}{L} n_{\Delta_{ij}}}, \quad (34)$$

and the diagonal matrix $\mathrm{diag}\,\boldsymbol{r}$ is built from the (real-valued) reverberation vector $\boldsymbol{r} = [r_1, r_2, \ldots, r_N]^{\mathrm{T}}$, with entries given by

$$r_i = \sum_{k=1}^{N} \frac{\sigma_{ki}^2}{1-a_{ki}^2}, \quad (35)$$

for $i \in \{1,2,\ldots,N\}$. We can relate the reverberation vector $\boldsymbol{r}$ directly to the diffuse component of the impulse response by noting that, from (25), the diffuse response energy of cross-channel (i,j) is equal to

$$\rho_{ij}^2 = \sum_{n=n_{\mathrm{w}_{ij}}}^{\infty} \mathrm{E}\{\check{\mathrm{h}}_{ij}^2[n]\} = \sum_{n=0}^{\infty} \mathrm{E}\left\{ a_{ij}^{2n} \, \mathrm{w}_{ij}^2[n] \right\} = \frac{\sigma_{ij}^2}{1-a_{ij}^2}. \quad (36)$$

Indeed, if we compare (35) and (36), we find out that for each component i, the reverberation vector is expressed as a sum of the diffuse response energies of the channels from that zone i to all other zones, *i.e.*, we have

$$r_i = \sum_{k=1}^{N} \rho_{ki}^2. \quad (37)$$

Summarizing, by incorporating the newly acquired information in (24), we state that the optimal processing scheme for the ℓ_2 distortion measure with a channel matrix following the model of (25) is given by

$$\boldsymbol{s}'(f) = (\boldsymbol{D}(f)^{\mathrm{H}}\boldsymbol{D}(f) + \mathrm{diag}\,\boldsymbol{r})^{-1}\boldsymbol{D}(f)^{\mathrm{H}}\,\boldsymbol{s}(f), \quad (38)$$

where $\boldsymbol{D}(f)$ and $\boldsymbol{r}$ are given componentwise in (34) and (35), respectively.

VI. Simulations

A. Room Model and Simulation Parameters

To assess the performance of the proposed MSE optimal algorithm of Section V, we apply it in a multizone speech reinforcement context, with $N = 4$ zones. The multizone channel

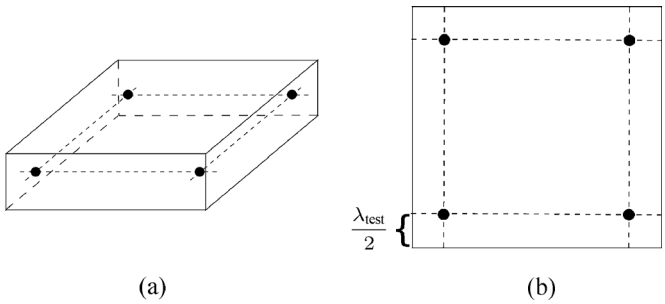

Fig. 5. Rectangular box room model (schematic). Black dots denote zone placement. (a) 3D sideview, (b) Top view.

$\mathbf{H}$ is used to model a room with a rectangular box form of dimensions $20 \times 20 \times 5$ m^3 (L × W × H), where the four zones are put at the corners (Fig. 5) at a height $h = 1.5$ m. This room setup can be seen as an idealization of the practical scenario of, *e.g.*, a museum room, where there are multiple exhibited paintings, and around each painting, some interpretive speech segment is played back continuously. A guard distance of half a wavelength is used with respect to the walls, since the approximation of a diffuse late impulse response as in (25) is only valid starting from that distance [45]. For this, we used a reference wavelength $\lambda_{\text{test}} = c/f_{\text{test}} = 69$ cm, with $c = 344$ m/s equal to the speed of sound, corresponding to the frequency $f_{\text{test}} = 500$ Hz. The distance between loudspeaker (source) and microphone (receiver) in different zones is approximated by the point distance between zones as depicted in Fig. 5. As to loudspeaker-microphone pairs in the same zone, we set their distance to 10% of the smallest point distance between zones, which results in $d_{\text{loud,mic}} = 1.931$ m. The criterion to choose this distance is that it is much smaller than the distance between zones, so that the approximation of using the point distance for loudspeaker-microphone separation in different zones remains accurate.

With this shoebox model, the parameters of the stochastic model in Section V are determined as follows. The direct path magnitudes Δ_{ij} in (34) are modeled using a standard free-field law,

$$\Delta_{ij} = \frac{1}{d_{ij}}, \tag{39}$$

where d_{ij} is the distance between zone i and j. The direct path phase delays $n_{\Delta_{ij}}$ are computed from these distances using the distance-time relation of a travelling wave. As to the reverberation vector components r_i of (37), we first model the direct-to-reverberation ratio, $\text{DRR}_{ij} = \Delta_{ij}^2/\rho_{ij}^2$ (by definition), using [45]

$$\text{DRR}_{ij} = \frac{\left(10^{\frac{24V}{c\,S\,\text{T}_{60}}} - 1\right) S}{16\pi d_{ij}^2}, \tag{40}$$

where we simplified the sources (loudspeakers) to be omnidirectional, T_{60} is the 60 dB reverberation time, and S and V are the room surface and volume, respectively. Using (39) and (40), we come to diffuse reverberation response energies given by

$$\rho^2 = \rho_{ij}^2 = \frac{\Delta_{ij}^2}{\text{DRR}_{ij}} = \frac{16\pi}{\left(10^{\frac{24V}{c\,S\,\text{T}_{60}}} - 1\right) S}. \tag{41}$$

We note that these energies are independent of the zone pair (index) ij, since all zones are placed in the same room and the diffuse sound field is homogeneous. Consequently, we also have reverberation components r_i which are independent of zone i, and which are given by

$$r_i = N\rho^2 = \frac{16\pi N}{\left(10^{\frac{24V}{c\,S\,\text{T}_{60}}} - 1\right) S}, \tag{42}$$

where we used (37) and (41), and noted the independence of ρ^2 on the zone pair.

Finally, the reverberation parameters a_{ij} and σ_{ij} for the generation of the channel in (25) and (26) are computed using the definition of 60 dB reverberation time and (36). Also here it holds that they are index independent, and they are given by

$$a = a_{ij} = 10^{-\frac{3}{\text{T}_{60}\,f_s}}, \quad \sigma^2 = \sigma_{ij}^2 = (1 - a^2)\rho^2, \tag{43}$$

where f_s is the sampling frequency used. Last but not least, we would like to mention that we set the diffuse response delay $n_{\text{w}_{ij}}$ equal to $n_{\Delta_{ij}}$ and use pseudorandom white Gaussian noise to produce the stochastic component of (25).

B. Signal-to-Distortion Ratio (SDR)

Per zone, one hundred sentences are used, randomly chosen out of the TIMIT database [46], each sentence having a duration of at least two seconds. The sentences are silence-trimmed at the extremities, processed and passed through the signal model according to (5). In total, 19.1 minutes of speech are used. We evaluate four noise types for $\mathbf{b}(f)$, namely white Gaussian noise, speech-shaped Gaussian noise (SSN), recorded babble noise and a noise sample of trains passing by. The first two types are rigorously wide-sense stationary, whereas the third type is a real-life approximation to stationary noise, and the fourth type is highly non-stationary. We vary the reverberation time T_{60} from 10 ms to 10 s exponentially in steps of $10^{0.2}$ and the Signal-to-Noise ratio (SNR) from -20 to 60 dB in steps of 10 dB. We use three pre-processing cases for comparison, namely the original proposed algorithm of (38), a single zone variant, where we run the proposed algorithm independently for each zone (adapting the equations for the case $N = 1$), and a reference unprocessed situation. Finally, besides evaluating the algorithm for a multi-source broadcast as described, we evaluate the single source broadcast case, $\boldsymbol{s}(f) = s(f)\,\mathbf{1}_N$, by setting the single source speech $s(f)$ as the speech of zone one of the multi-source case. A total number of $4 \times 16 \times 9 \times 3 \times 2 = 3456$ conditions are thus evaluated. We compute the signal-to-distortion ratio (SDR)

$$\text{SDR} = 10 \log_{10} \frac{\langle \|\boldsymbol{s}\|^2 \rangle}{\langle \|\boldsymbol{x} - \boldsymbol{s}\|^2 \rangle}, \tag{44}$$

TABLE I
SIMULATION PARAMETERS USED IN SECTIONS VI-B AND VI-C

N	dim. [m^3]	V [m^3]	S [m^2]	λ_{test} [m]	$d_{\text{loud,mic}}$ [m]
4	$20 \times 20 \times 5$	2000	1200	0.69	1.931

Δ_{ij} [a]	$n_{\Delta_{ij}}$ [ms] [a]	ρ^2	T_{60} [s]	SNR [dB]
0.5178	5.61	9.8669×10^{-14}	0.01 ... 10	-20 ... 60
0.0518	56.14			
0.0366	79.39			

[a] values for (in sequence): $j = i$, straight, cross direct path; see Fig. 5.

f_s [kHz]	N_{fr}	L	processing windows	overlap
16	16384	16384	square-root Hann	50%

where $\langle \cdot \rangle$ denotes averaging across frames, for each condition.

All signals are sampled at a sample frequency $f_s = 16$ kHz. For the reconstruction of $\boldsymbol{s}'$, we use a weighted overlap-add (WOLA) mechanism, with frame and DFT size $N_{\text{fr}} = L = 16384$, corresponding to 1.024 s speech segments. This unusually long frame size is used to accommodate the long reverberation tails of the impulse responses between zones for high reverberation times. Also, square-root Hann analysis and synthesis windowing is used with 50% overlap. Regarding the filtering operation in (6) (product on the right hand side of the equation), a standard (non-weighted) overlap-add method is used. Finally, the computation of the SDR in (44) is done with resource to a short-time DFT (STDFT) analysis. The frame and DFT size are again set to 16384, and a Hann analysis window with 50% overlap is used. All parameters used for the simulation are summarized in Table I.

Due to the large dimension of the results, only a representative selection will be shown here. In Fig. 6, we plot the difference between SDR in the processed and the unprocessed cases, for the SSN noise type and multi-source input. The predicted error of this difference, measured by 95% confidence intervals (CI's), was calculated under the assumption of independent Gaussian distortion samples $\|\boldsymbol{x}_l - \boldsymbol{s}_l\|^2$, where l is the frame number. The magnitude of this error lies within 0.33 dB for all SNR and T_{60} values. Furthermore, we have found out that the noise type has no influence on SDR results and, as such, we perform the analysis for this exemplifying noise type only. We observe that the proposed algorithm performs best for high SNR and low reverberation times (T_{60}), corresponding in the limit to a non-noisy and non-reverberant direct path component in the channel of (25) and (5). In fact, it can be seen from (38) that for this limiting case, the proposed algorithm boils down to

$$\boldsymbol{s}'(f) = \boldsymbol{D}(f)^{-1}\, \boldsymbol{s}(f), \tag{45}$$

as for low T_{60} the diffuse component $\boldsymbol{r}$ vanishes. In words, in the limit of low reverberation times, the proposed algorithm boils down to a direct path compensation, as in early crosstalk cancellation works [26]. The single-zone variant cannot achieve the same performance due to the fact that only the intra-zone path is compensated, being the direct paths of other zones not taken into account.

We also observe that a decrease in SNR degrades performance gradually, down to the level that no benefit is obtained by the processing scheme. This is motivated by the additive noise model and the noise independency of the proposed algorithm.

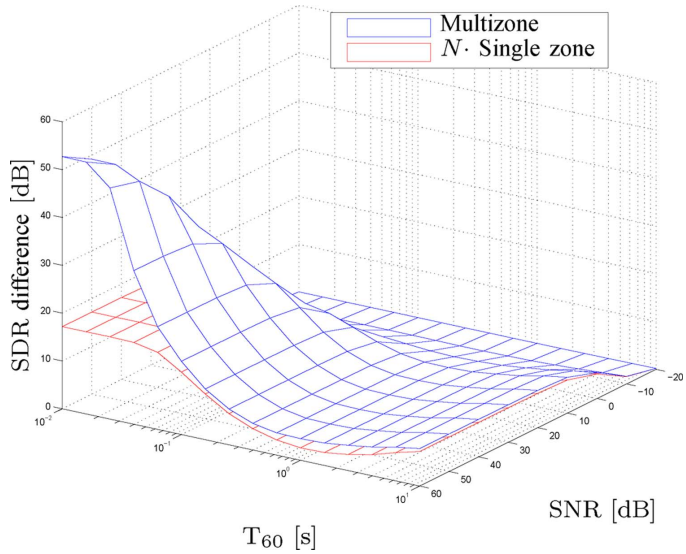

Fig. 6. SDR benefit upon processing compared to the unprocessed case, as a function of T_{60} and SNR. Multi-source broadcast scenario.

Furthermore, when varying T_{60} in the high SNR region, minimum performance is achieved in the order of one second, and for higher T_{60} again improvements are obtained, although they are much lower than improvements for low reverberation. We note that the characteristic T_{60} which provides this dip in performance is unrelated to the chosen frame size N_{fr}, as was confirmed by simulations with a larger frame size (results not shown).

Also, for high T_{60} values, both multi- and single zone approaches perform approximately equally well, with only a slight gain for the multizone case. The motivation for this can be found in the fact that the diffuse response energy gets larger than the direct path energy for high T_{60} (*i.e.*, the DRR gets smaller than one). Indeed, when this happens, the reverberant component $\boldsymbol{r}$ of the algorithm dominates and, seen that this component is zone-independent for the proposed room scenario, multizone specific processing fades.

The results described above concern the case where multiple source signals are pre-processed in our multi-zone context. For the case that the same source signal is used for all zones (single source broadcast), a similar distortion figure as in Fig. 6 is obtained (not shown). A difference of about 1 to 2.5 dB is observed in the optimal operating region of low T_{60} and high SNR (the multi-source case performing the best), whereas in the other regions no difference is observed. Also, the iterated single zone algorithm shows the largest differences.

C. Objective Intelligibility Measures

The assessment in Section VI-B was further extended, under the same test conditions, to two objective intelligibility measures, namely to the running speech variant of the speech transmission index (STI) ([14], Section IIA) and to the short-time objective intelligibility (STOI) [15] measures. The STI is a good estimator of speech intelligibility under reverberant corruption [47] and the STOI measure was designed for time-frequency weighted (noisy) speech [15]. Both measures are able to estimate speech intelligibility degradation caused by additive noise correctly. Although no unified intelligibility

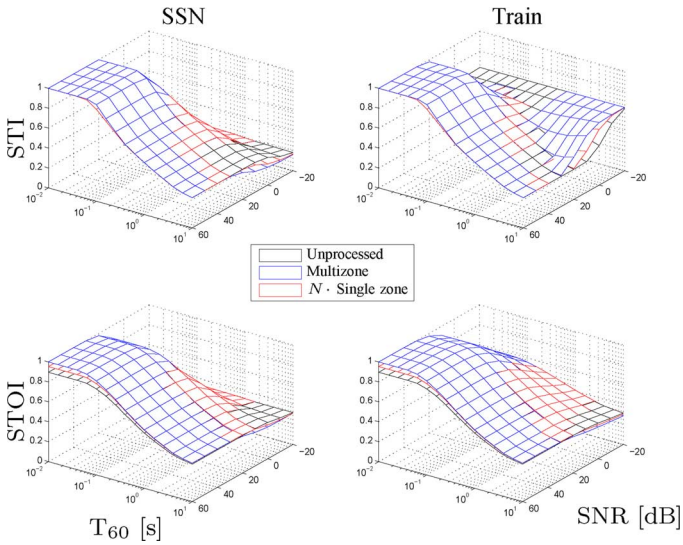

Fig. 7. Simulated STI and STOI as a function of T_{60} and SNR.

measure exists which was designed to predict corruptions caused by source-based time-frequency weighting *and* reverberant corruption *and* additive noise simultaneously, we still attempt to gain some insight on speech intelligibility with the analysis of these measures applied in our framework. We also note that we should not expect mandatory improvements *a priori* of the proposed algorithm under these measures, since the algorithm was optimized for the quadratic measure of (20) and not for these more complex measures.

Fig. 7 shows the computed STI and STOI values as a function of SNR and T_{60} for the SSN and train noise conditions, for the case of a multiple source broadcast. The white and babble noise conditions display behavior similar to the SSN condition. Also, for the STI, a CI analysis is more difficult to perform than for the SDR of Section VI-B, since the STI is the result of a whole chain of complex operations applied to a population average, which in turn comes from power spectral density estimates [14]. For the STOI, the analysis is straightforward, since the STOI is directly defined as a population average. Confidence errors smaller than 10^{-3} were obtained for the STOI, again under independent Gaussian assumptions.

As intuition tells, both intelligibility measures are monotonically increasing for increasing SNR and decreasing T_{60}. The only exception to this is the STI measure for the train noise; here, we find an unexpected increase of predicted intelligibility with decreasing SNR. This can be motivated by the fact that the STI operates on long-term power spectra of the signals being compared, in contrast to short-time measures as STOI. Consequently, the STI cannot incorporate temporal fluctuations of the noise signal and is inadequate for non-stationary noise.

Furthermore, we observe that for low T_{60} values, we can sort from highest to lowest intelligibility as follows: multizone, iterated single zone and unprocessed speech (not clearly visible for low SNR from the angle used in the figure). This order is intuitive since, as explained in Section VI-B, the multizone algorithm then applies perfect direct path cancellation, whereas the single zone variant applies partial compensation. In the other extreme of high T_{60}, we observe the inverse order: apart from the STI for train noise analysed above, there is some predicted intelligibility degradation, or at least no improvement, when applying the proposed algorithms (and more so for multizone than for iterated single zone processing). The authors cannot be sure if the predictions in this region are accurate, since neither STOI was designed for reverberant corruptions nor STI for general time-frequency weighted speech. Nevertheless, the figures do stress the fact that optimization of a simple distortion measure like the ℓ_2 error may not always guarantee an intelligibility improvement, thereby motivating the inclusion of more complex perceptual features in the distortion measure (*e.g.*, [48], [12], [10]). See Section VII for a discussion on this.

We have also evaluated the two objective intelligibility measures for the case that a same (single) source is provided as input (results not shown). We observe less predicted intelligibility in the multi-source case than for a single source. Equivalently, we can state that a single source is predicted to be less prone for intelligibility degradation than multiple sources. This is to be expected, since in the single source case, crosstalk has the same effect as if there were extra reverberation sources contributing to the received speech. Since the presence of reverberation (specially of the early components [45]) is less critical to intelligibility than the presence of "real" crosstalk, better intelligibility can be expected for a single source.

D. Least Squares Crosstalk Canceller

As the proposed algorithm in fact behaves as a crosstalk canceller, we compare it to the crosstalk canceller of Kallinger and Mertins [24]. This canceller adopts a deterministic least-squares (pseudoinverse) type of solution, but where an extra stochastic term is included to compensate for deviations from the known impulse response. The stochastic term is parameterized by the radius around which the system should remain robust upon physical displacement, which we will denote by R. Due to the costly memory consumption of the algorithm, derived from its usage of large convolution matrices representing the acoustic (reverberant) channel, we restrict ourselves to smaller impulse responses of length $L = 2048$ (128 ms). We also use a smaller room of dimensions $4.22 \times 3.10 \times 2.6\ \mathrm{m}^3$ (L × W × H) with the same guard distance to the walls $\lambda_{\mathrm{test}}/2$ as previously. We reduce the number of zones to $N = 2$, putting the zones at the edges of the longitudinal dimension. The microphones are placed at $1/3^{\mathrm{rd}}$ the distance between the two zones (closer to the corresponding zone). The proposed algorithm is run with the smaller frame and DFT sizes $N_{\mathrm{fr}} = L = 2048$ as well. The rest of the parameters is left essentially the same as in Section VI-B. Also, since our algorithm adopts a stochastic approach and the algorithm in [24] departs from a deterministically known channel, we need to make a fair comparison. For this, we generate 10 different realizations of the channel in (25) and average the performance (ordinate) while delivering the different realizations as an input to [24].

For a T_{60} running from 1 ms to 100 ms exponentially in steps of $10^{0.2}$, we assess the (averaged) SDR of (44), the STI and STOI for the pre-processings: unprocessed; proposed algorithm of (38); Kallinger [24] with no stochastic compensation ($R = 0$); Kallinger with the proposed radius in [24], $R = 0.02$ m; and Kallinger with a radius $R = 0.30$ m. For simplicity, we assume a noise-free scenario. The first usage case of Kallinger

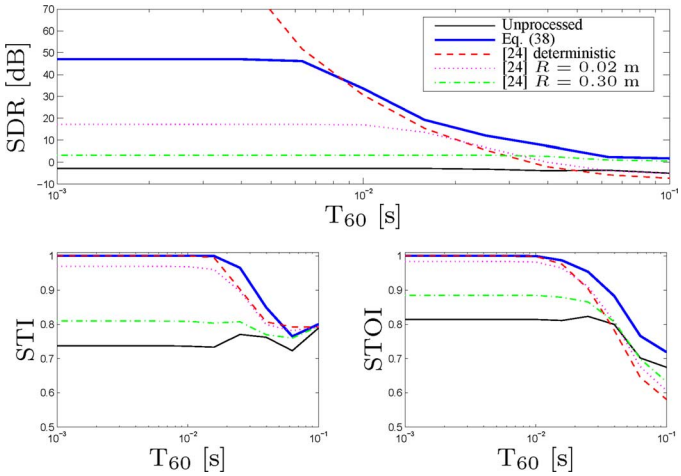

Fig. 8. SDR, STI and STOI of proposed algorithm and several least-squares crosstalk cancellers based on [24], as a function of T_{60}.

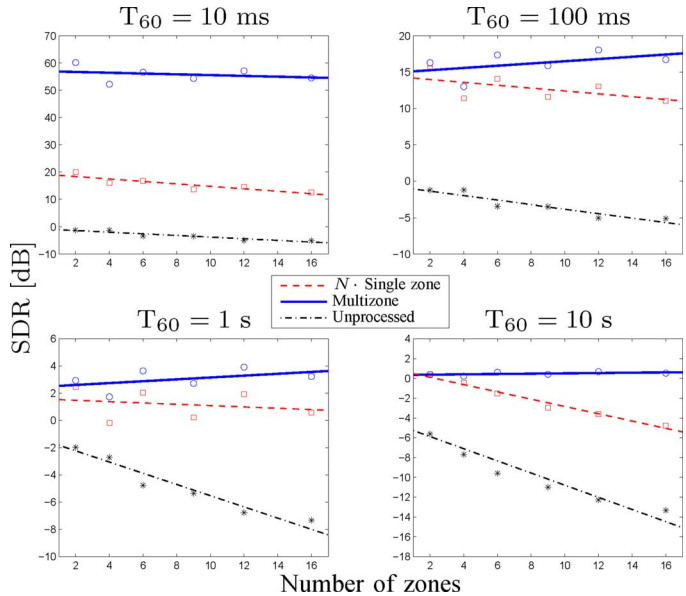

Fig. 9. Simulated SDR as a function of the number of zones N for selected reverberation times. Markers: simulated points; lines: least-squares fitted regression lines.

$(R = 0)$ corresponds to a fully deterministic traditional least-squares (pseudoinverse) solution, whereas the last case ($R = 0.30$ m) corresponds to having a displacement around half a wavelength, for which the impulse response decorrelates with respect to the original response. The results are displayed in Fig. 8.

We observe that for low T_{60}, where the direct paths are the predominant contribution for the channel, the purely deterministic approach performs the best. There, we have a plateau of about 175 dB SDR for the lowest reverberation times (not visible). Our approach also manages to perform good direct path compensation, which is perceptually as good as the deterministic approach (confirmed by informal listening and by the STI and STOI performance in the lower part of Fig. 8). When including a stochastic compensation term in [24], the higher the compensation radius R is, the more degraded crosstalk compensation performance is. The tradeoff of this performance decrease is directly seen in the high T_{60} region. High compensation radii lead to better (average) performance under stochastically described reverberant channels. There, the deterministic approach performs the worst, while the stochastic approaches manage to compensate for diffuse reverberation. The proposed approach is the best performing algorithm for $T_{60} \geq 10$ ms, and performance approaches the one for the highest compensation radius $R = 0.30$ m, for the highest reverberation times. All in all, we can thus conclude that our algorithm joins the best of deterministic direct path compensation and stochastic diffuse field compensation, by showing good tradeoff performance both for low as for high reverberation times.

E. Number of Zones

Finally, we assess the algorithm performance with an increasing number of zones. In a noise-free environment, we compute the SDR of (44), varying T_{60} from 10 ms to 10 s in decades, and the number of zones in the range $N \in \{2, 4, 6, 9, 12, 16\}$, corresponding to maximally spanned room constellations of 2×1, 2×2, 3×2, 3×3, 4×3 and 4×4 grids, respectively. The same guard distances to the walls and remaining settings were used as in Sections VI-A and VI-B.

Fig. 9 plots the resulting SDR as a function of the number of zones, in conjunction with least-squares lines fitted to the results. The general trend is observed that as the number of zones increases, SDR decreases (distortion increases) in the unprocessed and single-zone cases, while multizone processed SDR manages to stabilize. This effect is specially prominent for high reverberation times ($T_{60} \geq 100$ ms).

We compared the SDR slopes in Fig. 9 to each other by means of Analysis of Covariance (ANOCOVA) tests. The result is that for all assessed reverberation times with $T_{60} \geq 100$ ms, the processing condition has a significant effect on the slope ($F = 4.0013, p = 0.0466$; $F = 9.8580, p = 0.0029$; and $F = 40.0243, p < 0.0001$ for T_{60} equal to 100 ms, 1 s and 10 s, respectively). Multiple comparison tests indicate that the unprocessed condition has a significantly lower slope than the multizone processed condition (95% CI's used) for the three cases. From this result, we conclude that the proposed algorithm is able to significantly bridge SDR deterioration with an increasing number of zones, or at least so if T_{60} is large enough.

VII. Discussion

In this work, we used a smooth (continuously differentiable) distortion measure, which could either quantify quality or intelligibility degradation, to build an optimization framework. Although traditional intelligibility measures suffer from the fact that they do apply non-smooth techniques, such as hard clipping [36], [14], we would like to note that research has focused lately on building mathematically tractable intelligibility measures that are more amenable for optimization. For example, in [15], the authors present a measure which is based on a (smooth) Pearson's correlation coefficient of speech temporal envelopes. In [49], the hard clipping procedure of the Speech Intelligibility Index (SII) [36] is approximated by a smooth concave clipping function. Furthermore, the observation that speech reinforcement and enhancement algorithms which are designed having speech intelligibility in mind can degrade speech quality

[3], [6], [16], makes us believe that quality modeling should be somehow included in speech reinforcement, as is already done in the domains of audio and speech coding. These remarks and the mathematical simplicity of the approach motivate the choice of smooth distortion measures in this work.

We have also seen in Section V that, although the optimization framework we developed caters for noise, reverberation and crosstalk simultaneously, the application of the developed theory to the simple ℓ_2 error measure delivered us a crosstalk cancellation scheme which acts independently of the noise process. Nevertheless, the authors have recently shown that by applying the framework developed here to a spectral magnitude distortion measure, a noise dependent algorithm follows [50]. Using this insight, we conjecture that applying our framework to more complex distortion measures than the ℓ_2 error leads to more elaborate, complex and meaningful schemes than pure crosstalk cancelling. For example, one could think of bringing perceptual features into play in speech reinforcement by applying this work to analytically defined audibility measures (*e.g.*, [10], [11]) or to speech distortion measures such as the Log-Spectral Distance or the Itakura-Saito measure [51]. Also, the application of this work to intelligibility models, which frequently exhibit non-smooth behavior (see above), is a challenge in itself.

Concerning the hybrid stochastic-deterministic channel model which was chosen for the concretization of the derived MSE optimal algorithm, we remark that the impact of this choice on the algorithm is that it relies on a good estimation of the direct path (deterministic) components in the channel transfer matrix. Indeed, a small displacement of the listener in the room can introduce errors in the direct path estimates, thereby degrading algorithm performance. Furthermore, dynamic channel conditions also make channel estimation difficult, and non-updated or non-accurate estimates also contribute to a degraded algorithm performance. A practical scenario where dynamic channels could be an issue is a train station, where a train passing by changes space acoustics significantly. For bridging these difficulties, one could extend stochastic descriptions in the modeled channel, *e.g.*, by including a stochastically described perturbation (error) term on the direct path components.

VIII. Conclusion

In this work, we studied speech reinforcement (source-based listening enhancement) in a multiple zone scenario, where the goal is to optimize an overall expected distortion of multiple mutually interfering source-receiver constellations. The approach is abstract and works upon a functional quantification of degradation (distortion measure). After building an optimization framework including effects of noise, reverberation and crosstalk simultaneously, we derived analytical necessary conditions for optimality. Subsequently, we applied the conditions for the case of a simple distortion measure, namely the ℓ_2 error measure (euclidean distance).

Using the applied conditions and considering a hybrid deterministic-stochastic model for the acoustic channel, a Mean-Square Error (MSE) optimal algorithm was derived, which eventually boiled down to a crosstalk cancellation technique. Also, the general and abstract approach undertaken leaves space open for the development of more complex algorithms which feature more than plain crosstalk cancellation. The algorithm was thoroughly evaluated; a clear benefit of multizone processing could be established versus the iterated application of the corresponding single zone algorithm, and we showed that the algorithm combines the best of deterministic and stochastically compensated least-squares crosstalk cancellation approaches existing in literature.

To the best of the authors' knowledge, this work constitutes the first approach in solving speech reinforcement in a multizone scenario, *i.e.*, upon existence of adjacent channels with mutual crosstalk. Also, we stress the importance of the novel abstraction level undertaken, which provides conditions for reusability of the work.

References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Boca Raton, FL, USA: CRC, 2007.

[2] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2006, vol. I, pp. 493–496.

[3] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power constraining," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Sep. 2006.

[4] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2009, pp. 1844–1848.

[5] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," *ITG-Fachtagung Sprachkommunikation*, vol. Paper 8, Oct. 2010.

[6] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 4061–4064.

[7] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Commun.*, vol. 48, pp. 549–558, 2006.

[8] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, no. 1, pp. 280–285, Jan. 2010.

[9] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 1138–1149, Aug. 2007.

[10] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1292–1304, Jan. 2005.

[11] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1553–1564, Jul. 2012.

[12] T. Dau, D. Pschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, Jun. 1996.

[13] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility indexbased approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[14] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[15] 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, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Amer.*, vol. 130, no. 5, pp. 3013–3027, 2011.
[17] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 277–282, Aug. 1976.
[18] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Amer.*, vol. 43, no. 1, pp. 81–86, 1968.
[19] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Commun.*, vol. 45, no. 2, pp. 101–113, 2005.
[20] N. Hodoshima, T. Arai, A. Kusumoto, and K. Kinoshita, "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," *J. Acoust. Soc. Amer.*, vol. 119, no. 6, pp. 4055–4064, Jun. 2006.
[21] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.
[22] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 106–108, May 1999.
[23] D. B. Ward, "Joint least squares optimization for robust acoustic crosstalk cancellation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 211–215, Feb. 2000.
[24] M. Kallinger and A. Mertins, "A spatially robust least squares crosstalk canceller," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. I, pp. 177–180.
[25] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1829–1842, Aug. 2012.
[26] D. B. Ward and G. W. Elko, "Virtual sound using loudspeakers: robust acoustic crosstalk cancellation," in *Acoustic Signal Processing for Telecommunications*, S. L. Gay and J. Benesty, Eds. Boston, MA, USA: Kluwer, 2000, ch. 14.
[27] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.
[28] G. A. Miller, *Language and Communication*. New York, NY, USA: McGraw-Hill, 1963.
[29] I. B. Thomas, "The second formant and speech intelligibility," in *Proc. Nat. Electron. Conf.*, 1967, vol. 23, pp. 544–548.
[30] I. B. Thomas and R. J. Niederjohn, "The intelligibility of filtered-clipped speech in noise," *J. Aud. Eng. Soc.*, vol. 18, no. 3, pp. 299–303, Jun. 1970.
[31] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.
[32] P. S. Chanda and S. Park, "Speech intelligibility enhancement using tunable equalization filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2007, vol. 4, pp. IV-613–IV-616.
[33] J. J. J. W. Strange and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 695–705, 1983.
[34] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, 1986.
[35] J. W. Shin and N. S. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 887–890, Nov. 2007.
[36] ANSI S3.5-1997. Methods for Calculation of the Speech Intelligibility Index ANSI. New York, NY, USA, 1997.
[37] T. Langhans and H. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1982, vol. 7, pp. 156–159.
[38] M. Kallinger and A. Mertins, "Room impulse response shortening by channel shortening concepts," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2005, pp. 898–902.
[39] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," U.S. patent 3,236,949, Feb. 1966.
[40] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–195, Mar. 1998.
[41] H. I. K. Rao, V. J. Matthews, and Y.-C. Park, "A minimax approach for the joint design of acoustic crosstalk cancellation filters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2287–2298, Nov. 2007.
[42] J. O. Jungmann, R. Mazur, M. Kallinger, and A. Mertins, "Robust combined crosstalk cancellation and listening-room compensation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2011, pp. 9–12.
[43] J. B. Conway, *Functions of One Complex Variable II*, 1st ed. : Springer-Verlag, May 1995.
[44] J. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Univ. du Maine, Le Mans, France, 1988.
[45] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven Technical Univ. , Eindhoven, The Netherlands, 2007.
[46] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993. Philadelphia, PA, USA, Linguistic Data Consortium, Philadelphia.
[47] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.
[48] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin/Heidelberg, Germany: Springer, 2007.
[49] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.
[50] J. Crespo and R. Hendriks, "Multizone near-end speech enhancement under optimal second-order magnitude distortion," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2013.
[51] R. Gray, A. Buzo, J. Gray, A. , and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 367–376, Aug. 1980.

João B. Crespo received his M.Sc. in electrical engineering from the Technical University of Lisbon, Portugal, in 2009. During the last year of his M.Sc., he was an exchange student at the Signal and Information Processing Lab of the faculty of Electrical Engineering, Mathematics and Computer Science at the Technical University of Delft, The Netherlands. He worked at ExSilent B.V., The Netherlands, as a DSP developer. Currently, he is pursuing a Ph.D. at the Signal and Information Processing Lab. His current research is on source-based listening enhancement. Areas of interest include audio and speech processing, auditory perception and information theory.

Richard C. Hendriks obtained his M.Sc. and Ph.D. degrees (both cum laude) in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2003 and 2008, respectively. From 2003 till 2007 he was a Ph.D. researcher at Delft University of Technology, Delft, The Netherlands. From 2007 till 2010 he was a postdoctoral researcher at Delft University of Technology. Since 2010 he is an assistant professor in the Signal and Information Processing Lab of the faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 till March 2009 he was a visiting researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement and intelligibility improvement.