SPEECH REINFORCEMENT WITH A GLOBALLY OPTIMIZED PERCEPTUAL DISTORTION MEASURE FOR NOISY REVERBERANT CHANNELS

João B. Crespo and Richard C. Hendriks

Delft University of Technology
Signal & Information Processing lab
{j.b.crespo, r.c.hendriks}@tudelft.nl

ABSTRACT

In this paper, a time-frequency weighting is proposed for speech reinforcement (near-end listening enhancement) in a noisy and reverberant environment, which optimizes a perceptual distortion measure globally for a number of time-frequency bins. Simulations confirm the optimality of the algorithm and a comparison is made to three reference methods using two additional instrumental measures.

Index Terms— Near-end listening enhancement, reverberant noisy channel, perceptual distortion measure

1. INTRODUCTION

Speech reinforcement (Fig. 1) aims to pre-process a clean speech signal, such that when it is played back and corrupted in an acoustic environment, it still comes across to the listener with good intelligibility and/or quality. The nature of the corruptions could be any, such as noise, coloring, reverberation, crosstalk, among others. Examples of applications where speech reinforcement is used are mobile telephony, conference systems and public addressing.

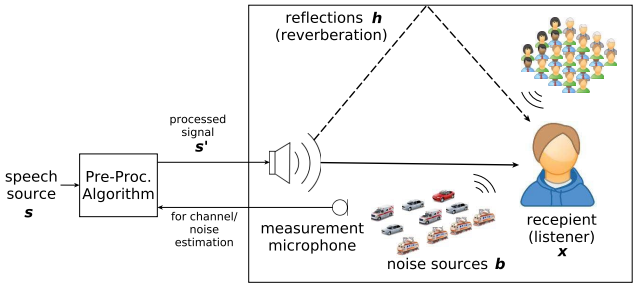

Fig. 1. Speech reinforcement concept.

Lately, the enhancement of intelligibility in such a framework under additive noise has received its due attention [1–3]. Despite its relevance in many practical scenarios, the reverberant case has been hardly looked into, possibly due to the complexity of its nature. Some relatively recent proposals include the enhancement of the modulation spectrum of the clean speech [4] and steady-state suppression for reducing overlap-masking [5]. In both cases, empirical observations concerning intelligibility in reverberant environments [6, 7] are used to motivate the algorithms. Approaches which are mathematically optimal with respect to some criterion concentrate on linear pre-filtering and simple cost functions based on ℓ_p-norms [8,9], so as to match the global impulse response to a reference response. Also, algorithms that do take reverberation into account do not consider additive noise in the environment.

Taal et al. [10] optimized a more complex distortion measure [11] based on a perceptual model of auditory periphery for the noise-only case. It was shown that, although the distortion measure was designed for assessing perceptual quality, it could also be used for intelligibility enhancement due to its short-time sensitivity. In [12], we extended the algorithm of [10] partially to include reverberant corruptions, where the optimization was performed locally for each time-frequency (T/F) bin independently. In this paper, we complete the extension of [10, 12] for the noisy reverberant case, by globally optimizing distortion summed over a number of T/F bins. We show that this global problem can be written down as a posynomial program, which can be elegantly solved in polynomial time by *e.g.*, interior point methods [13]. We compare the proposed algorithm to three reference methods and confirm its optimality via simulations. Finally, we discuss briefly on the algorithm behavior, showing with an illustrative example that it suppresses speech onsets.

2. PRELIMINARIES

As in [10], we work with a gamma-tone filterbank on top of a standard Discrete Fourier Transform (DFT)-based analysis/synthesis scheme of window size N and shift size $R < N$. Denote by $\boldsymbol{s}(f,t) \in \mathbb{R}^N$ the (band-pass) clean speech segment in the time-domain, at frame index $t \in \mathbb{Z}$ and auditory band $f \in \{0, 1, \ldots, M-1\}$, obtained by filtering the windowed short-time speech segment of frame t by the impulse

This research is partly supported by the Dutch Technology Foundation STW and Bosch Security Systems B.V., The Netherlands.

978-1-4799-6808-4/14/$31.00 ©2014 IEEE

response of the gamma-tone filter of index f, $\boldsymbol{g}(f) \in \mathbb{R}^N$. We use circular convolutions ($*$) of size N for filtering. Consider also a disturbance segment $\boldsymbol{\epsilon}(f,t) \in \mathbb{R}^N$ defined similarly. We will use the disturbance signal to model additive noise and late reverberation. The distortion in T/F bin (f,t) is then quantified as

$$d(\boldsymbol{s}(f,t), \boldsymbol{\epsilon}(f,t)) = \frac{1}{N}\,\mathbf{1}^{\mathrm{T}} \left(\frac{\boldsymbol{\epsilon}^2(f,t) * \boldsymbol{h}_s}{\boldsymbol{s}^2(f,t) * \boldsymbol{h}_s} \right), \qquad (1)$$

where $\mathbf{1} \in \mathbb{R}^N$ denotes the all-ones vector, $\boldsymbol{h}_s \in \mathbb{R}^N$ is the impulse response of an exponential smoother (150 Hz cut-off frequency), and vector squaring and division are the pointwise squares and quotients, respectively. The distortion measure essentially measures the detectability of the disturbance under speech by constructing an internal auditory representation of the clean and corrupted speech signals, and comparing the internal representations using an ℓ_1 distance [11].

In this work, we will use several mathematical properties of (1) which were derived in [12]. For deterministic speech $\boldsymbol{s}$, uncorrelated stochastic disturbance segments $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_{1,2}$ and scaling factors α, β, within T/F bin (f,t), we have

$$\mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2)] = \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_1)] + \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_2)] \qquad (2)$$

$$\mathrm{E}[d(\alpha\boldsymbol{s}, \boldsymbol{\epsilon})] = \frac{1}{\alpha^2}\,\mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon})] \qquad (3)$$

$$\mathrm{E}[d(\boldsymbol{s}, \beta\boldsymbol{\epsilon})] = \beta^2\,\mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon})]\,. \qquad (4)$$

We follow the model of [12] for the received corrupted speech $\boldsymbol{x}$, which is written down as (see also Fig. 1)

$$\boldsymbol{x}(f,t) = \sum_{\tau=0}^{T-1} \boldsymbol{h}(\tau) * \boldsymbol{s}'(f, t-\tau) + \boldsymbol{b}(f,t)\,, \qquad (5)$$

where $\boldsymbol{s}'$ and $\boldsymbol{b}$ are the processed speech and noise, respectively and where $\boldsymbol{h}(t) \in \mathbb{R}^N$, $t \in \{0, 1, \ldots, T-1\}$, contains a zero-padded segment of the impulse response $h[n]$:

$$\boldsymbol{h}(t) = [h[Rt], h[Rt+1], \ldots, h[Rt+R-1], 0, \ldots, 0]^{\mathrm{T}}, \qquad (6)$$

being the number of zeros $N - R$. As to the processing function, we apply a T/F weighting in the gamma-tone domain, $\boldsymbol{s}'(f,t) = \alpha(f,t)\boldsymbol{s}(f,t)$, with weights $\alpha(f,t) > 0$. Rewriting (5), we have

$$\boldsymbol{x}(f,t) = \alpha(f,t)\boldsymbol{s}_e(f,t) + \sum_{\tau=1}^{T-1} \alpha(f,t-\tau)\boldsymbol{s}_r(f,t,\tau) + \boldsymbol{b}(f,t) \qquad (7)$$

where $\boldsymbol{s}_e(f,t) = \boldsymbol{h}(0) * \boldsymbol{s}(f,t)$ is considered to contain early reverberant speech, and $\boldsymbol{s}_r(f,t,\tau) = \boldsymbol{h}(\tau) * \boldsymbol{s}(f, t-\tau)$ is assigned to late reverberant speech in frame t having as source the speech in frame $t - \tau$. We assume the early speech to be deterministic, since it is available in a reinforcement scenario. Furthermore, we hypothesize that late reverberation behaves similarly as noise and model both noise and late reverberation as stochastic processes. Finally, we assume late speech and noise terms to be mutually uncorrelated, arguing that speech and noise can be considered to be independent and that inter-frame correlations of speech are limited.

3. GLOBALLY OPTIMAL DISTORTION

3.1. Algorithm derivation

Our aim is to pre-process the signal, so as to minimize the detectability of noise and late reverberation under early speech, jointly for a number of T/F bins. This can be motivated by the fact that early speech is known to contribute positively to intelligibility [14], whereas late reverberation and noise do the opposite [6]. We also include an energy constraint on the output for technical reasons and, *e.g.*, for loudspeaker or hearing damage protection. We are thus concerned with the problem

$$\min_{\substack{\alpha(f,t) \\ (f,t)\in\mathcal{L}}} \sum_{(f,t)\in\mathcal{L}} \mathrm{E}\left[d\Big(\alpha(f,t)\boldsymbol{s}_e(f,t)\,, \sum_{\tau=1}^{T-1} \alpha(f,t-\tau)\boldsymbol{s}_r(f,t,\tau) + \boldsymbol{b}(f,t) \Big) \right]$$

$$\text{s.t.} \sum_{(f,t)\in\mathcal{L}} \alpha^2(f,t)\|\boldsymbol{s}(f,t)\|^2 \le cR^2\,, \qquad (8)$$

where $\mathcal{L}$ is the set of T/F bins we are interested in, $R^2 = \sum_{(f,t)\in\mathcal{L}} \|\boldsymbol{s}(f,t)\|^2$ is the energy constraint on the output speech, set to the energy of the input signal, and $c \ge 1$ is a relaxation factor. Note also that $\alpha(f, t-\tau)$ is a constant for $(f, t-\tau) \notin \mathcal{L}$ instead of an optimization variable, and has therefore to be fixed *a priori*.

Using the properties in (2)–(4), we can rewrite (8) as

$$\min_{\substack{A(f,t) \\ (f,t)\in\mathcal{L}}} \sum_{(f,t)\in\mathcal{L}} D(f,t)$$

$$\text{s.t.} \sum_{(f,t)\in\mathcal{L}} A(f,t)\|\boldsymbol{s}(f,t)\|^2 \le cR^2\,,$$

$$D(f,t) = \frac{1}{A(f,t)} \left[\sum_{\tau=1}^{T-1} A(f,t-\tau)D_r(f,t,\tau) + D_b(f,t) \right], \qquad (9)$$

where $D(f,t)$ is the local distortion in a single T/F bin [12], where we substitute $A(f,t) = \alpha^2(f,t)$ with the implicit constraint $A(f,t) > 0$, and where we define

$$D_r(f,t,\tau) = \mathrm{E}[d(\boldsymbol{s}_e(f,t), \boldsymbol{s}_r(f,t,\tau))] \qquad (10)$$

$$D_b(f,t) = \mathrm{E}[d(\boldsymbol{s}_e(f,t), \boldsymbol{b}(f,t))] \qquad (11)$$

to be the partial distortions due to late reverberation and to noise, respectively. Problem (9) is a posynomial program [13]

which, in this form, is non-convex and difficult to solve. Nevertheless, using the log-transform $A(f,t) = \mathrm{e}^{\tilde{A}(f,t)}$, we can re-write it as a convex problem of the new variables $\tilde{A}(f,t)$, which is readily solved using efficient numerical algorithms [13].

3.2. Practical implementation

For the practical usability of the derived algorithm, we construct estimates of the expected values of (10) and (11) by assuming models for the stochastic quantities. For (10), we assume channel and late speech to be independent, we neglect early reflections, and take the late channel response to be white. The first assumption is based on the distinct nature of the channel and speech, whereas the last two assumptions are simplifications. Under these assumptions, we have

$$D_r(f,t,\tau) = \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}\left[\left(\boldsymbol{h}(\tau) * \boldsymbol{s}(f,t-\tau) \right)^2 \right] * \boldsymbol{h}_s}{\boldsymbol{s}^2(f,t) * \boldsymbol{h}_s} \right) \quad (12)$$

$$= \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}\left[\boldsymbol{h}^2(\tau) \right] * \mathrm{E}[\boldsymbol{s}^2(f,t-\tau)] * \boldsymbol{h}_s}{\boldsymbol{s}^2(f,t) * \boldsymbol{h}_s} \right) . \quad (13)$$

Accordingly, we use a model of the form

$$h[n] = \delta[n] + a^{n-n_0} u[n-n_0] r[n-n_0] \quad (14)$$

for the impulse response, where $\delta[n]$ is the Dirac delta function, $u[n]$ the unit step, $0 < a < 1$ a damping factor, related to the reverberation time T_{60} by $a = 10^{-3/(\mathrm{T}_{60} f_s)}$, $n_0 = 3R$ the starting sample for late reverberation (assumed, for simplicity, to be a multiple of the shift size R), and $r[n]$ a zero-mean stationary white stochastic process of variance σ_r^2, related to the direct-to-reverberation ratio (DRR) by $\sigma_r^2 = (1-a^2)\mathrm{DRR}$. In essence, we hereby adopt a simplified version of the Polack model [15] for the late impulse response. We then have

$$\mathrm{E}[\boldsymbol{h}^2(\tau)] = a^{2(\tau-\tau_r)R} \sigma_r^2 \, \boldsymbol{a} \, , \quad (15)$$

where $\boldsymbol{a} = [1, a^2, a^4, \ldots, a^{2(R-1)}, 0, \ldots, 0]^{\mathrm{T}}$ and $\tau_r = n_0/R = 3$. For the late speech, we use the statistical model of [10] given by

$$\mathrm{E}[\boldsymbol{s}^2(f,t-\tau)] = \sigma_s^2(f,t-\tau)(\boldsymbol{g}^2(f) * \boldsymbol{w}^2) \, , \quad (16)$$

where $\sigma_s^2(f,t)$ is the average power spectral density (PSD) of the speech in band f, time frame t, $\boldsymbol{g}(f)$ the impulse response of the f^{th} auditory filter, and $\boldsymbol{w}$ the analysis window.

Following a similar procedure as above for D_b, we have

$$D_b(f,t) = \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[\boldsymbol{b}^2(f,t)] * \boldsymbol{h}_s}{\boldsymbol{s}^2(f,t) * \boldsymbol{h}_s} \right) , \quad (17)$$

where the model of [10] delivers

$$\mathrm{E}[\boldsymbol{b}^2(f,t)] = \sigma_b^2(f,t)(\boldsymbol{g}^2(f) * \boldsymbol{w}^2) \quad (18)$$

and where $\sigma_b^2(f,t)$ is the average PSD of the noise. We compute the PSDs using oracle estimators, which recursively apply an exponential smoothing to the periodograms (poles 0.96 and 0.996 for noise and speech, respectively). Naturally, for a real-world application, oracle estimators are not available. In that case, a standard noise PSD estimation approach from literature can be used [16]. Likewise, we use oracle information for the reverberation parameters T_{60} and DRR.

Finally, regarding the selection of T/F bins, we take speech blocks of dimension $M = 40$ bands by $B = 64$ frames, and include only voice active bins in $\mathcal{L}$ by thresholding $\|\boldsymbol{s}(f,t)\|^2$ as prescribed in [10]. The inactive bins are assigned a gain $\alpha(f,t) = 1$. The optimization of (9) is carried out using the `MOSEK` package [17], after which the block of processed speech bins, $\{\boldsymbol{s}'(f,t)\}_{f,t}$, is synthesized using overlap-add at the block level with a size B Hann window (along time), followed by a gamma-tone filterbank synthesis step and weighted overlap-add (WOLA). The block position is subsequently shifted by $B/2$ frames and the process is repeated. Regarding the gamma-tone analysis filters, we use bands with linearly spaced center frequencies in ERB scale ranging from 150 Hz to 8 kHz. In turn, the gamma-tone synthesis process consits of summating the bandpass analysis signals scaled according to the result of (9). As to the WOLA step, the used parameters were 32 ms frame and DFT sizes, and 50% overlap square-root Hann analysis and synthesis windows.

4. EVALUATION

We evaluate the proposed approach of Sec. 3 using 60 seconds of speech sampled at $f_s = 16$ kHz, obtained by concatenating sentences randomly chosen out of the TIMIT database [18]. For each sentence, leading and trailing silences were trimmed and a minimal duration of 2 s was used. The speech signal is processed and corrupted by reverberation and additive noise, where we use realizations of (14) for the former and of speech-shaped noise (SSN) for the latter.

We vary the signal-to-noise ratio (SNR) of the SSN in the range from -20 to 20 dB in steps of 5 dB while keeping $\mathrm{T}_{60} = 0.5$ s constant, and subsequently, we vary T_{60} from 100 ms to 1.58 s in exponential steps of $10^{0.2}$ while keeping SNR $= 0$ dB constant. The DRR and relaxation factor c are fixed to 1 (0 dB) and 10 (10 dB), respectively. For each combination of parameters, we compute the signal-to-distortion ratio $\mathrm{SDR} = -10\log_{10}\langle D(f,t)\rangle_{f,t}$, where $D(f,t)$ is the local distortion of (9) and $\langle\cdot\rangle_{f,t}$ denotes averaging in time and frequency (across all available speech blocks). We furthermore compute the short-time objective intelligibility (STOI) index [19] and the speech transmission index (STI) for running speech [20, Sec. II-A]. The STOI is a good predictor of intelligibility of speech in non-stationary noise [19], whereas the STI is used for speech in noise and/or reverberation [6]. Under the hypothesis that the disturbance in our

scenario (late reverberation plus noise) can be considered to behave (to some extent) as non-stationary noise, these two measures should give us some insight about subjective results.

We assess the algorithm of Sec. 3 (Proposed), the algorithm by Taal et al. [10] (Taal), which corresponds to $D_r \equiv 0$, the steady state suppressor by Hodoshima et al. [5] (Hodoshima), the normalized SNR recovery approach by Sauert et al. [21] (Sauert), and a reference signal, where the processing consists of a simple frequency independent gain c (Flatgain). For the Taal and Sauert algorithms, we use the SSN as input to the oracle noise PSD estimator. Also, we do not compute the SDR for Sauert and Hodoshima, since they operate within other signal processing frameworks, so that we do not have access to $A(f,t)$ in (9). In total, we assess $(9{+}7){\cdot}5 = 80$ conditions using two to three merit figures.

The results are summarized in Fig. 2, where we plot the merit figures as a function of the SNR (left) and T_{60} (right). We observe from the SDR figures that, as expected, the proposed algorithm optimizes problem (9). Also, for high SNR values, the proposed algorithm is most beneficial with respect to the Taal algorithm, where reverberation is neglected; for low SNR values, the benefit vanishes. As to the independent objective measures, we see that STOI results are in line with the SDR results, predicting an intelligibility increase of the proposed algorithm for high SNR values and for a wide range of reverberation times. In contrast, the STI measure predicts a degradation of the proposed algorithm with respect to Taal. The biggest degradation is observed for high SNR values, where the proposed algorithm seems to be outperformed by simple flat-gain processing. Regarding the other algorithms, both the proposed algorithm and Taal outperform Hodoshima and Sauert under most conditions. Since STOI and STI show contradictory results, the question remains on whether the measures can compare intelligibility differences

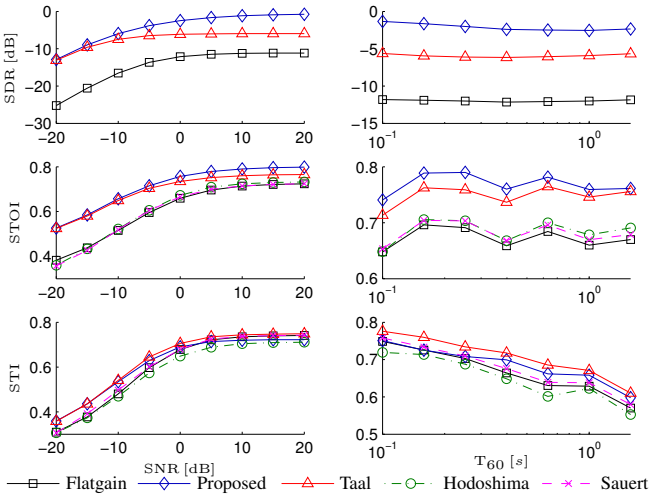

Fig. 2. Evaluated SDR (top), STOI (mid) and STI (bottom) as a function of SNR (left) and T_{60} (right).

of the algorithms correctly, and whether they are applicable in this reinforcement scenario, where we are comparing intelligibility of *non-linearly* processed speech under noise and reverberation simultaneously.

For a better insight on the energy distribution behavior of the processed signals, we plot the output level as a function of time for the proposed algorithm, Taal and Flatgain (SNR $=$ 10 dB, DRR $=$ -10 dB, T_{60} $=$ 0.5 s, and c $=$ 10 dB). The plotted energy profiles correspond to the words "rag like that", consisting of alternating transient and stationary segments. We observe that the proposed algorithm frequently

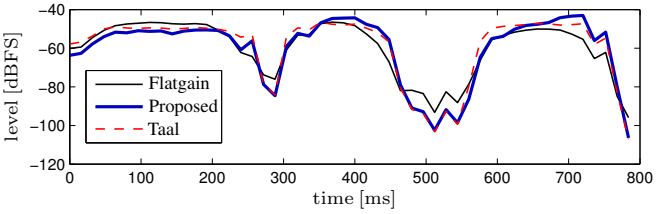

Fig. 3. Energy profile of processed signals in the time domain.

exhibits an onset suppression effect at stationary speech components: the level at an onset starts relatively lower and increases with time. This effect was experienced in informal listening as a perceptual boosting of vowel nuclei. Similarly, it was also observed by plotting the time-domain waveforms, that the transient peaks of the proposed algorithm are less pronounced compared to Taal. It is known that onsets are important for speech perception [22], so that it is questionable whether these effects benefit intelligibility. On the other hand, onset suppression has also the effect of reducing the effective time-span of stationary components, *i.e.*, it has a steady-state suppressing effect. Since steady-state suppressors are known to enhance intelligibility in reverberant environments by reducing overlap masking [5], the question remains on which of the opposing effects predominates and whether intelligibility enhancement is observed in practice. This issue can be further addressed via psychometric (subjective) testing, and is left as future work.

5. CONCLUSION

In this paper, we derived an algorithm for speech reinforcement under noise and reverberation, which optimizes a perceptual distortion measure globally for a number of time-frequency (T/F) bins. Simulations confirm the optimality of the algorithm and show mixed performance results on predicted intelligibility figures. It was seen that the algorithm suppresses speech onsets, and it was discussed whether this could produce a benefit in intelligibility. The mixed results and arguments motivate the need for more research in this category of algorithms, namely, in the domain of psychometric testing.

6. REFERENCES

[1] "The Listening Talker Project," `http://listening-talker.org`, 2013.

[2] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Elsevier Comput. Speech Language*, 2013.

[3] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge," in *ISCA Interspeech*, Lyon, France, 2013.

[4] 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," *Elsevier Speech Commun.*, vol. 45, no. 2, pp. 101–113, 2005.

[5] 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. Am.*, vol. 119, no. 6, pp. 4055–4064, June 2006.

[6] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, March 1985.

[7] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, vol. 21, no. 6, pp. 577–580, 1949.

[8] M. Kallinger and A. Mertins, "Room impulse response shortening by channel shortening concepts," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2005, pp. 898–902.

[9] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 249–259, 2010.

[10] 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 *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2012, pp. 4061–4064.

[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, Language Process.*, vol. 20, no. 5, pp. 1553–1564, July 2012.

[12] J. B. Crespo and R. C. Hendriks, "Speech reinforcement in noisy reverberant environments using a perceptual distortion measure," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2014.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[14] G. A. Soulodre, N. Popplewell, and J. S. Bradley, "Combined effects of early reflections and background noise on speech intelligibility," *Journal of Sound and Vibration*, vol. 135, no. 1, pp. 123–133, 1989.

[15] J. D. Polack, *La transmission de l'énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, Le mans, France, 1988, Thèse de doctorat d'etat.

[16] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement – A Survey of the State-of-the-Art*, Morgan & Claypool Publishers, 2013.

[17] MOSEK ApS, "MOSEK optimization software ver. 7.0.0.103," `http://www.mosek.com`, 1998–2014.

[18] 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, Linguistic Data Consortium, Philadelphia.

[19] 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, pp. 2125–2136, Sept. 2011.

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

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

[22] W. Strange, J. J. Jenkins, and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 695–705, 1983.