SPEECH REINFORCEMENT IN NOISY REVERBERANT ENVIRONMENTS USING A PERCEPTUAL DISTORTION MEASURE

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 locally for each time-frequency bin. The algorithm acts as a dynamic range compressor, smearing out the energy of the clean speech along time. Simulations predict an intelligibility increase with respect to the unprocessed condition and two reference methods, for moderate smoothing windows, as measured by the optimized distortion measure and two objective intelligibility 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.

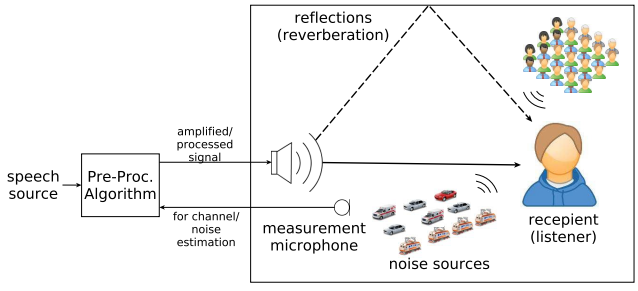

Fig. 1. Speech reinforcement concept.

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

A myriad of pre-processing algorithms have been proposed for the case where the corruption consists of additive noise (*e.g.*, [1–6]). Nevertheless, the reverberant case has been far less looked into. Some authors propose modulation filtering to pre-process speech in a reverberant environment [7, 8]. There, the rationale is that the decrease of the modulation spectrum of speech in reverberant environments [9] should be pre-compensated by filtering it by a fixed transfer function. Other authors work with steady state suppression [10]. In this case, the argumentation is that intelligibility degradation occurs when the reverberant tails of high energy steady-state components mask subsequent low-energy (transient) regions, the so-called "overlap-masking" effect.

The approaches targeting the reverberant case described above are of empirical nature, in the sense that they are not provably optimal with respect to any criterion, such as a model for speech quality or intelligibility. Existing optimization approaches in this setup focus more on pre-filtering the signal, such that the global impulse response gets matched to a desired response. Traditional approaches [11] use a deterministic inverse pre-filter, being inherently unrobust with respect to displacement of the listener. More recent approaches concentrate on shortening ("reshaping") the global impulse response, by maximizing an ℓ_p-norm based objective corresponding to a desired response segment [12, 13].

The described reverberation pre-processing approaches so far do not take noise into account. For the reverberant noisy case, only recently an approach relying on a reinforcement framework was proposed in [14], which optimizes the mean-square error (MSE) between clean and corrupted speech under a hybrid deterministic-stochastic channel assumption. Although the framework caters for noise, the algorithm is noise-independent.

In this paper, we aim to build further upon optimization driven speech reinforcement for reverberant noisy channels, by proposing a time-frequency (T/F) weighting scheme that optimizes the perceptual distortion measure introduced in [3, 15]. To keep the problem mathematically tractable, we optimize the distortion in each T/F bin independently, not considering the global signal distortion on all T/F bins altogether.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. PERCEPTUAL DISTORTION MEASURE

We choose the distortion measure of [3, 15] due to the fact that it is sensitive to short-time information, resulting its optimization in the case of a noise-only channel in a transient amplification algorithm [3]. As motivated in Sec. 1, this kind of effect is desirable in reverberant environments to diminish overlap-masking. In this section, we introduce the distortion measure and corresponding mathematical properties, which make it amenable for analytical optimization.

We work with a ubiquitous processing scheme, where speech is segmented in overlapping time frames of length N (shift size $R < N$), windowed, processed, and combined via weighted overlap-add (WOLA). Consider a (clean) speech frame $\boldsymbol{s}(t) \in \mathbb{R}^N$ in the time domain, where $t \in \mathbb{Z}$ denotes the frame number. The distortion measure proposed in [3] works on a time-frequency (T/F) representation $\boldsymbol{s}(f,t)$ induced by an auditory filterbank,

$$\boldsymbol{s}(f,t) = \boldsymbol{g}(f) * \boldsymbol{s}(t)\,, \tag{1}$$

where $f \in \{0, 1, \ldots, M-1\}$ denotes the frequency band index, $\boldsymbol{g}(f) \in \mathbb{R}^N$ the impulse response of the auditory filter of index f, and $*$ represents circular convolution of size N. Consider also a disturbance frame $\boldsymbol{\epsilon}(t) \in \mathbb{R}^N$ and its corresponding T/F representation $\boldsymbol{\epsilon}(f,t)$, defined similarly as in (1). We will use the disturbance signal in Sec. 3 to model additive noise and late reverberation. The distortion in one T/F bin is then quantified as

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

where $\mathbf{1}$ denotes the all-ones vector of length N, $\boldsymbol{h}_s \in \mathbb{R}^N$ is the impulse response of an exponential smoother, and vector squaring and division are defined as the point-wise 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 [15].

The distortion measure has several desirable mathematical properties, which make it easy to manipulate. In the following, we discard T/F indexing (f,t) for a matter of visual comfort.

Disturbance additivity: Consider a clean speech T/F bin $\boldsymbol{s}$, modeled to be deterministic, and two uncorrelated stochastic disturbances $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$. The expected distortion of the combined disturbance is

$$\mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2)] = \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[(\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2)^2] * \boldsymbol{h}_s}{\boldsymbol{s}^2 * \boldsymbol{h}_s} \right) \tag{3}$$

$$= \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[\boldsymbol{\epsilon}_1^2] * \boldsymbol{h}_s}{\boldsymbol{s}^2 * \boldsymbol{h}_s} \right) + \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[\boldsymbol{\epsilon}_2^2] * \boldsymbol{h}_s}{\boldsymbol{s}^2 * \boldsymbol{h}_s} \right) \tag{4}$$

$$= \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_1)] + \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon}_2)]\,, \tag{5}$$

being $\mathrm{E}[\cdot]$ the expectation operator.

Scaling signal: If we scale the deterministic clean speech $\boldsymbol{s}$ by a positive real gain $\alpha > 0$, under an arbitrary stochastic disturbance $\boldsymbol{\epsilon}$, we get

$$\mathrm{E}[d(\alpha\boldsymbol{s}, \boldsymbol{\epsilon})] = \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[\boldsymbol{\epsilon}^2] * \boldsymbol{h}_s}{(\alpha\boldsymbol{s})^2 * \boldsymbol{h}_s} \right) = \frac{1}{\alpha^2} \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon})]\,. \tag{6}$$

Scaling disturbance: If we scale the stochastic disturbance $\boldsymbol{\epsilon}$ by a positive real gain $\beta > 0$, for an arbitrary deterministic signal $\boldsymbol{s}$, we get

$$\mathrm{E}[d(\boldsymbol{s}, \beta\boldsymbol{\epsilon})] = \mathbf{1}^{\mathrm{T}} \left(\frac{\mathrm{E}[(\beta\boldsymbol{\epsilon})^2] * \boldsymbol{h}_s}{\boldsymbol{s}^2 * \boldsymbol{h}_s} \right) = \beta^2 \mathrm{E}[d(\boldsymbol{s}, \boldsymbol{\epsilon})]\,. \tag{7}$$

3. OPTIMIZED ALGORITHM

As motivated in Sec. 1, we model the received (corrupted processed) speech $x[n]$ at the listener, at time sample n, by a convolutive and additive corruption, *i.e.*,

$$x[n] = \sum_{k=0}^{RT-1} h[k]\, s'[n-k] + b[n]\,, \tag{8}$$

where $h[n]$ is the impulse response of the reverberant room, $s'[n]$ is the processed signal, $b[n]$ is an additive noise term, and T reflects the length of $h[n]$ (in frames). It is easy to see that the model of (8) carries through to the T/F representation of (1) by

$$\boldsymbol{x}(f,t) = \sum_{\tau=0}^{T-1} \boldsymbol{h}(\tau) * \boldsymbol{s}'(f, t-\tau) + \boldsymbol{b}(f,t)\,, \tag{9}$$

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

$$\boldsymbol{h}(t) = [h[Rt], h[Rt+1], \ldots, h[Rt+R-1], 0, \ldots, 0]^{\mathrm{T}}\,. \tag{10}$$

To use the introduced signal model in the distortion measure of Sec. 2, we consider a processing gain function $\boldsymbol{s}'(f,t) = \alpha(f,t)\boldsymbol{s}(f,t)$, for positive T/F weights $\alpha(f,t)$, and rewrite (9) as

$$\boldsymbol{x}(f,t) = \alpha(f,t)\boldsymbol{s}_e(f,t) + \sum_{\tau=1}^{T-1} \alpha(f,t-\tau)\boldsymbol{s}_r(f,t,\tau) + \boldsymbol{b}(f,t) \tag{11}$$

where $\boldsymbol{s}_e(f,t) = \boldsymbol{h}(0) * \boldsymbol{s}(f,t)$ models early reverberant speech, and $\boldsymbol{s}_r(f,t,\tau) = \boldsymbol{h}(\tau) * \boldsymbol{s}(f,t-\tau)$ models late reverberant speech in frame t having as source the speech in frame $t-\tau$. We assume the early speech to be deterministic, whereas the noise and late reverberation terms are modeled as

stochastic processes. We also assume the latter to be mutually uncorrelated.

Our aim is then to process the signal, such as to minimize the detectability of noise and late reverberation under early speech. This can be motivated by the fact that early speech is known to contribute positively to intelligibility [16], whereas late reverberation and noise do the opposite [9]. Also, for bridging the technical difficulty of having a non-achievable solution, we add an energy constraint on the output. The optimization problem we are concerned with, is thus

$$\min_{\substack{\alpha(f,t-\tau),\\ \tau=0,\dots,T-1}} \mathrm{E}\left[d\Bigg(\alpha(f,t)\boldsymbol{s}_e(f,t)\,, \sum_{\tau=1}^{T-1}\alpha(f,t-\tau)\boldsymbol{s}_r(f,t,\tau)+\boldsymbol{b}(f,t)\Bigg)\right]$$
$$\text{s.t.}\ \sum_{\tau=0}^{T-1}\alpha^2(f,t-\tau)\|\boldsymbol{s}(f,t-\tau)\|^2 = R^2\,, \qquad (12)$$

where $R^2 = \sum_{\tau=0}^{T-1}\|\boldsymbol{s}(f,t-\tau)\|^2$ is the input signal energy. It should be noted that the problem stated has only local view on the perceptual distortion caused in one single T/F bin. Also, since we get a set of T gains per T/F bin (f,t) as a result of the optimization problem, namely, $\alpha(f,t-T+1),\alpha(f,t-T+2),\dots,\alpha(f,t)$, we somehow have to combine overlapping gains for different t. In this work, we choose to average overlapping gains.

Using Properties (5)–(7), we can rewrite Problem (12) as

$$\min_{\substack{A(f,t-\tau),\\ \tau=0,\dots,T-1}} \frac{1}{A(f,t)}\left[\sum_{\tau=1}^{T-1}A(f,t-\tau)D_r(f,t,\tau)+D_b(f,t)\right]$$
$$\text{s.t.}\ \begin{cases}\sum_{\tau=0}^{T-1}A(f,t-\tau)\|\boldsymbol{s}(f,t-\tau)\|^2 = R^2\\ A(f,t-\tau)\geq 0,\ \tau=0,\dots,T-1\,,\end{cases} \qquad (13)$$

where we substituted $A(f,t) = \alpha^2(f,t)$, and where we define $D_r(f,t,\tau) = \mathrm{E}[d(\boldsymbol{s}_e(f,t),\boldsymbol{s}_r(f,t,\tau))]$ to be the partial distortion due to late reverberation and $D_b(f,t) = \mathrm{E}[d(\boldsymbol{s}_e(f,t),\boldsymbol{b}(f,t))]$ to be the one due to noise.

Problem (13) is a linear fractional program, which can be cast into a linear program via the Charnes-Cooper variable transformation [17, 18]. The resulting program has an analytical solution which, when substituted back into Problem (13), takes the form

$$A(f,t-\tau) = \begin{cases}\frac{R^2}{\|\boldsymbol{s}(f,t)\|^2} & ,\ \tau=0\\ 0 & ,\ \tau=1,2,\dots,T-1\,.\end{cases} \qquad (14)$$

This result is intuitive, since, if we only have a local view on the distortion, we would like to suppress speech of past frames $t-\tau$, $\tau>0$, to avoid reverberation in the current frame, represented by the distortion terms $D_r(f,t,\tau)$, and amplify the current frame as much as possible (within the energy constraint) to mask the noise, represented by the distortion term $D_b(f,t)$. By using the averaging strategy to combine overlapping gains, the resulting overall gains are given by

$$\alpha^2(f,t) = \frac{\frac{1}{T}\sum_{\tau=0}^{T-1}\|\boldsymbol{s}(f,t-\tau)\|^2}{\|\boldsymbol{s}(f,t)\|^2}\,, \qquad (15)$$

where we also used the definition of R^2. Note that the effect of (15) on the speech signal is that its energy gets smeared out in time, *i.e.*, (multiband) dynamic range compression occurs. Indeed, the energy after processing, $\|\boldsymbol{s}'(f,t)\|^2 = \alpha^2(f,t)\|\boldsymbol{s}(f,t)\|^2$ simply gets to be the average energy over the past T time frames. Such compression techniques have been shown to benefit intelligibility in the noisy channel case [6, 19, 20], and also for the reverberant case, they act in a comparable way as steady state suppressors, seen that stationary components frequently belong to high energy speech portions and transients to lower energy portions. Also, the algorithm acts independently of noise statistics and of specific details of reverberation, taking only a smoothing constant T as a parameter.

4. EVALUATION

To evaluate the proposed approach of (15), we use one hundred sentences randomly chosen out of the TIMIT database [21], sampled at $f_s = 16$ kHz, where each sentence has a duration of at least two seconds. The sentences are silence-trimmed at the extremities and concatenated. In total, 284 seconds of speech are used. The resulting speech signal is processed and corrupted according to (8). The used WOLA parameters for the processing step were 32 ms frame and DFT sizes, and 50% overlap square-root Hann analysis and synthesis windows. Therein, the short-time processed speech was obtained by applying a filterbank with $M = 40$ gammatone filters [3] to the short-time input speech, with linearly spaced center frequencies in ERB scale ranging from 150 Hz to 8 kHz, and subsequently adding the resulting bandpass signals scaled according to (15). Regarding the corruptions, we use speech-shaped noise (SSN) for the additive noise term, and for the reverberant corruption, we produce realizations of

$$h[n] = \delta[n] + a^{n-n_0}u[n-n_0]w[n-n_0]\,, \qquad (16)$$

where $\delta[n]$ is the Dirac delta function, $u[n]$ the unit step function, $0<a<1$ a damping factor, related to the reverberation time T_{60} by $a = 10^{-3/(\mathrm{T}_{60}f_s)}$, n_0 the starting sample for late reverberation, set to $50\ \mathrm{ms}\cdot f_s$, and $w[n]$ a zero-mean stationary white stochastic process of variance σ^2, set such that the direct-to-reverberation ratio, $\sigma^2/(1-a^2)$, is one. We hereby neglect early reflections in the reverberant channel and use the Polack model [22] for the late reverberation. Note that although it is irrealistic to neglect early reflections, the simulations thereby deliver worst-case performance, since early reflections benefit intelligibility [16]. A Gaussian pseudo-random number generator was used for the realizations of $b[n]$ and $w[n]$.

For assessing the resulting corrupted processed speech x, we define the signal-to-distortion ratio in dB, corresponding to the chosen distortion measure (2), as

$$\mathrm{SDR} = -10\log_{10}\left\langle \frac{1}{N}\mathbf{1}^{\mathrm{T}}\left(\frac{\boldsymbol{\epsilon}^2(f,t) * \boldsymbol{h}_s}{\boldsymbol{s}'^2(f,t) * \boldsymbol{h}_s}\right)\right\rangle_{f,t}, \quad (17)$$

where we took the disturbance $\boldsymbol{\epsilon} = \boldsymbol{x} - \boldsymbol{s}'$ to be equal to the noise plus late reverberation, and where $\langle\cdot\rangle_{f,t}$ denotes averaging through all frequency bands and time frames. A first-order smoothing filter $\boldsymbol{h}_s$ with a cutoff frequency $f_c = 125$ Hz was chosen, and the same filterbank as for the processing function was used, working on top of a short-time DFT (STDFT) analysis framework with DFT and frame size of 32 ms and a 50% overlap Hann analysis window.

For a fixed $\mathrm{T}_{60} = 0.5$ s, we vary the signal-to-noise ratio (SNR) of the SSN with respect to the source (input) speech from -20 to 20 dB in steps of 5 dB, and for a fixed $\mathrm{SNR} = 0$ dB, we vary T_{60} from 100 ms to $10^{0.2} = 1.58$ s in exponential steps of $10^{0.2}$. For each $(\mathrm{T}_{60}, \mathrm{SNR})$ combination, we compute the SDR of (17), the short-time speech intelligibility index (ST-SII) [23] and the speech transmission index (STI) for running speech [24, Sec. II-A]. The ST-SII is a good predictor of intelligibility of speech under non-stationary noise [23]. Assuming that we can model the disturbance in our scenario (late reverberation plus noise) as non-stationary noise, the ST-SII should correlate well with subjective results. Complementarily, the STI is a good predictor of intelligibility for speech under reverberation and noise [9]. We assess the proposed algorithm of (15) for $T = 4$, $T = 8$ and $T = 16$ (running averages of 64, 128 and 256 ms windows, respectively), the steady state suppressor of Hodoshima et al. [10], the normalized SNR recovery approach of Sauert et al. [1], and a reference unprocessed signal. For the Sauert algorithm, we use the combined SSN plus late reverberation as input to the oracle noise PSD estimator described in [1]. We assess a total number of $(9 + 7) \times 6 = 96$ conditions using three merit figures.

The results are summarized in Fig. 2, where we plot the merit figures as a function of the SNR in the top row ($\mathrm{T}_{60} = 0.5$ s) and as a function of T_{60} in the bottom row ($\mathrm{SNR} = 0$ dB). We observe that the proposed algorithm outperforms steady state suppression and normalized SNR recovery. This is in line with the results of [6, 20], and extends them for reverberant corruptions. Although we did only test the proposed algorithm for stationary noise, [20] shows the effectiveness of this kind of techniques for non-stationary maskers. Also, the longer the smoothing window of the algorithm is (parameter T), *i.e.*, the slower the resulting dynamic range compression is, the better the observed performance. It was also observed via informal listening that unfortunately, this comes at the price of poorer signal quality — for long smoothing windows, high energy regions of transient phonemes (*e.g.*, fricative consonants) get smeared out to such a degree that they mask subsequent phonemes, *i.e.*, an overlap-masking effect occurs. This algorithmic overlap-masking effect results in reduced signal quality and speech intelligibility. It can be avoided by setting the smoothing parameter not longer than the average phoneme length (maximally $T \approx 16$). Since the merit figures operate taking the processed signal as reference, and since they assume that the reference signal is fully intelligibile (or has perfect quality), they cannot predict the decrease in intelligibility/quality due to this type of overlap-masking. Note also that we choose to take the processed signal as reference, since we are interested in knowing to what extent the processed signal gets through the additive and convolutive corruptions; if we would take the source signal as reference, we would unfairly penalize algorithms which modify the input speech severely and which, nevertheless, provide for a notable intelligibility increase (*i.e.*, most of the approaches presented in Sec. 1 and the proposed approach). Assessing in how far algorithmic quality degradation plays a role is left as future work.

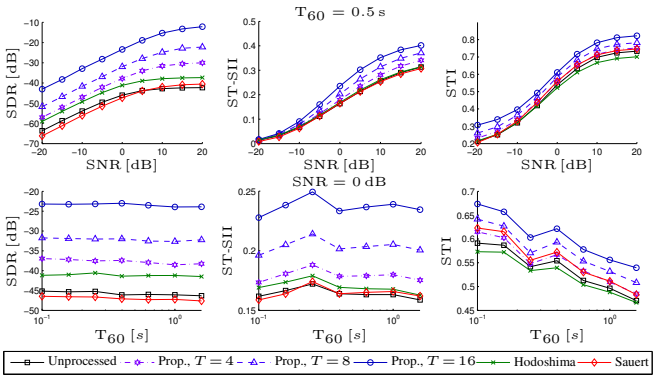

Fig. 2. Evaluated SDR, ST-SII and STI as a function of SNR (top) and T_{60} (bottom)

5. CONCLUSION

In this paper, we derived an algorithm for speech reinforcement under noise and reverberation, which optimizes a perceptual distortion measure locally for each time-frequency (T/F) bin. The algorithm acts as a multiband dynamic range compressor and ends up being independent of the local distortions. Simulations show that the algorithm outperforms two reference methods, and that slower dynamic range compression results in better performance, up to the point where overlap-masking occurs during processing. Example sound samples are available at `http://siplab.tudelft.nl/users/joao-crespo`.

6. REFERENCES

[1] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power con-

straining," in *Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006.

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

[3] 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.

[4] 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.

[5] S. D. Yoo et al., "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, Aug 2007.

[6] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, USA, 2012.

[7] T. Langhans and H. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, may 1982, vol. 7, pp. 156–159.

[8] 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.

[9] 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.

[10] 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.

[11] J. N. Mourjopoulos, "Digital equalization of room acoustics," *Audio Eng. Society*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

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

[13] 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.

[14] J. B. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 54–66, Jan. 2014.

[15] 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.

[16] 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.

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

[18] S. Schaible, "Parameter-free convex equivalent and dual programs of fractional programming problems," *Zeitschrift für Operations Research*, vol. 18, no. 5, pp. 187–196, 1974.

[19] 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. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 277–282, Aug 1976.

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

[21] 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.

[22] 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.

[23] K. S. Rhebergen and N. J. Versfeld, "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, April 2005.

[24] 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.