Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 437807, 9 pages
doi:10.1155/2009/437807

Research Article

Low Delay Noise Reduction and Dereverberation for Hearing Aids

Heinrich W. Löllmann (EURASIP Member) and Peter Vary

Institute of Communication Systems and Data Processing, RWTH Aachen University, 52056 Aachen, Germany

Correspondence should be addressed to Heinrich W. Löllmann, loellmann@ind.rwth-aachen.de

Received 11 December 2008; Accepted 16 March 2009

Recommended by Heinz G. Goeckler

A new system for single-channel speech enhancement is proposed which achieves a joint suppression of late reverberant speech and background noise with a low signal delay and low computational complexity. It is based on a generalized spectral subtraction rule which depends on the variances of the late reverberant speech and background noise. The calculation of the spectral variances of the late reverberant speech requires an estimate of the reverberation time (RT) which is accomplished by a maximum likelihood (ML) approach. The enhancement with this blind RT estimation achieves almost the same speech quality as by using the actual RT. In comparison to commonly used post-filters in hearing aids which only perform a noise reduction, a significantly better objective and subjective speech quality is achieved. The proposed system performs time-domain filtering with coefficients adapted in the non-uniform (Bark-scaled) frequency-domain. This allows to achieve a high speech quality with low signal delay which is important for speech enhancement in hearing aids or related applications such as hands-free communication systems.

1. Introduction

Algorithms for the enhancement of acoustically disturbed speech signals have been the subject of intensive research over the last decades, cf., [1–3]. The wide-spread use of mobile communication devices and, not at least, the introduction of digital hearing aids have contributed significantly to the interest in this field. For hearing impaired people, it is especially difficult to communicate with other persons in noisy environments. Therefore, speech enhancement systems have become an integral component of modern hearing aids. However, despite significant progress, the development of speech enhancement systems for hearing aids is still a very challenging problem due to the demanding requirements regarding computational complexity, signal delay and speech quality.

A common approach is to use a beamformer with two or three closely spaced microphones followed by a post-filter, e.g., [4, 5]. An adaptive beamformer is often used, implemented by first- or second- order differential microphone arrays or a generalized sidelobe canceller (GSC), respectively, e.g., [5]. Due to the use of small microphone arrays, only a limited noise suppression can be achieved by this, especially for diffuse noise fields. Therefore, the output signal of the beamformer is further processed by a (Wiener) post-filter to achieve an improved noise suppression, e.g., [4–7]. A related approach is to use an extension of the GSC structure termed as speech distortion weighted multi-channel Wiener filter [8, 9]. This approach allows to balance the tradeoff between speech distortions and noise reduction and is more robust towards reverberation than a common GSC.

So far, such systems achieve only a very limited suppression of speech distortions due to room reverberation. Such impairments are caused by the multiple reflections and diffraction of the sound on walls and objects of a room. These multiple echoes add to the direct sound at the receiver and blur its temporal and spectral characteristics. As a consequence, reverberation and background noise reduce listening comfort and speech intelligibility, especially for hearing impaired persons [10, 11]. Therefore, algorithms for a joint suppression of background noise and reverberation effects are of special interest for speech enhancement in hearing instruments. However, many proposals are less suitable for this application.

For example, dereverberation algorithms based on linear prediction such as [12] achieve mainly a reduction of early reflections and do not consider additive noise,

while algorithms based on a time-averaging [13] exhibit a high signal delay. Coherence-based speech enhancement algorithms such as [14] or [15] can suppress background noise and reverberation, but they are rather ineffective if only two closely spaced microphones can be used. This problem can be alleviated to some extend by a noise classification and binaural processing [16] which, however, requires two hearing aid devices connected by a wireless data link. A single-channel algorithm for speech dereverberation and noise reduction has been proposed recently in [17]. However, this algorithm is less suitable for hearing aids due to its high computational complexity and signal delay as well as its strong speech distortions.

A more powerful approach for noise reduction and dereverberation is to use blind source separation (BSS), e.g., [18]. Such algorithms do not require *a priori* knowledge about the microphone positions or source locations. However, they depend on a full data link between the hearing aid devices and possess a high computational complexity. Therefore, further work remains to be done to integrate such algorithms into common hearing instruments [19].

In this contribution, a single-channel speech enhancement algorithm is proposed, which is more suitable for current hearing aid devices. It performs a suppression of background noise and late reverberant speech by means of a generalized spectral subtraction. The devised (post-)filter exhibits a low signal delay, which is important in hearing aids, e.g., to avoid comb filter effects. The calculation of the late reverberant speech energy requires (only) an estimate of the reverberation time (RT), which is accomplished by a maximum likelihood (ML) approach. Thus, no explicit speech modeling is involved in the dereverberation process as, e.g., in [20] such that an estimation of speech model parameters is not needed here.

The paper is organized as follows. In Section 2, the underlying signal model is introduced. The overall system for low delay speech enhancement is outlined in Section 3. The calculation of the spectral weights for noise reduction and dereverberation is treated in Section 4. An important issue is the determination of the spectral variances of the late reverberant speech, which in turn is based on an estimation of the RT. These issues are treated in Sections 4.2 and 4.3. The performance of the new system is analyzed in Section 5, and the main results are summarized in Section 6.

2. Signal Model

The distorted speech signal $x(k)$ is assumed to be given by a superposition of the reverberant speech signal $z(k)$ and additive noise $v(k)$ where k marks the discrete time index. The received signal $x(k)$ and the original (undisturbed) speech signal $s(k)$ are related by

$$\begin{aligned} x(k) &= z(k) + v(k) \\ &= \sum_{n=0}^{L_\mathrm{R}-1} s(k-n)h_\mathrm{r}(n,k) + v(k) \end{aligned} \tag{1}$$

with $h_\mathrm{r}(n,k)$ representing the time-varying *room impulse response* (RIR) of (possibly infinite) length L_R between source and receiver. The reverberant speech signal can be decomposed into its early and late reverberant components

$$z(k) = \underbrace{\sum_{n=0}^{L_\mathrm{e}-1} s(k-n)h_\mathrm{r}(n,k)}_{=z_\mathrm{e}(k)} + \underbrace{\sum_{n=L_\mathrm{e}}^{L_\mathrm{R}-1} s(k-n)h_\mathrm{r}(n,k)}_{=z_\mathrm{l}(k)}. \tag{2}$$

The late reverberation causes mainly overlap-masking effects which are usually more detrimental for the speech quality than the "coloration" effects of early reflections.

Here, the *early reverberant speech* $z_\mathrm{e}(k)$ (and not $s(k)$) constitutes the target signal of our speech enhancement algorithm. This allows to suppress the *late reverberant speech* $z_\mathrm{l}(k)$ and additive noise $v(k)$ by modeling them both as uncorrelated noise processes and to apply known speech enhancement techniques, such as Wiener filtering or spectral subtraction, respectively. This concept, which has been introduced by Lebart et al. [21] and further improved by Habets [22], forms the basis of our speech enhancement algorithm. It is more practical for hearing aids as it avoids the high computational complexity and/or signal delay required by algorithms which strive for an (almost) complete cancellation of background noise and reverberation as, e.g., BSS.

3. Low Delay Filtering

A common approach for (single-channel) speech enhancement is to perform spectral weighting in the short-term frequency-domain. The DFT coefficients of the disturbed speech $X(i,\lambda)$ are multiplied with spectral weights $W_i(\lambda)$ to obtain M enhanced speech coefficients

$$\hat{S}(i,\lambda) = X(i,\lambda) \cdot W_i(\lambda); \quad i \in \{0,1,\ldots,M-1\}, \tag{3}$$

where i denotes the frequency (channel) index and λ the subsampled time index $\lambda = \lfloor k/R \rfloor$. (The operation $\lfloor \cdot \rfloor$ returns the greatest integer value which is lower than or equal to the argument.) For block-wise processing, the downsampling rate $R \in \mathbb{N}$ corresponds to the frame shift and λ to the frame index.

An efficient and common method to realize the short-term spectral weighting of (3) is to use a polyphase network DFT *analysis-synthesis filter-bank* (AS FB) with subsampling which comprises the common overlap-add method as special case, [2, 23]. A drawback of this method is that subband filters of high filter degrees are needed to achieve a sufficient stopband attenuation in order to avoid aliasing distortions, which results in a high signal delay. For hearing aids, however, an overall processing delay of less than 10 milliseconds is desirable to avoid comb filter effects, cf., [24]. Such distortions are caused by the superposition of a processed, delayed signal with an unprocessed signal which bypasses the hearing aid, e.g., through the hearing aid vent. This is especially problematic for devices with an "open fitting." Therefore, the algorithmic signal delay of the AS FB should

be significantly below 10 ms. One approach to achieve a reduced delay is to design the prototype lowpass filter of the DFT filter-bank by numerical optimization with the design target to reduce the aliasing distortions with constrained signal delay, [25, 26].

A significantly lower signal delay can be achieved by the concept of the *filter-bank equalizer* proposed in [27, 28]. The adaptation of the coefficients is performed in the (uniform or non-uniform) short-term frequency-domain while the actual filtering is performed in the time-domain. A related approach has been presented independently in [29] for dynamic range compression in hearing aids. The concept of the filter-bank equalizer has been further improved and generalized in [30, 31]. This filter(-bank) approach is considered here as it avoids aliasing distortions for the processed signal. In addition, the use of the warped filter-bank equalizer causes a significantly lower computational complexity and signal delay than the use of a non-uniform (Bark-scaled) AS FB for speech enhancement as proposed, e.g., in [32–34].

A general representation of the proposed speech enhancement system is provided by Figure 1. The subband signals $X(i,\lambda)$ are calculated either by a uniform or warped DFT analysis filter-bank with downsampling by R, which can be efficiently implemented by a polyphase network. The choice of the downsampling rate R is here not governed by restrictions for aliasing cancellation as for AS FBs since the filtering is performed in the time-domain with coefficients adapted in the frequency-domain. The influence of aliasing effects for the calculation of the spectral weights is negligible for the considered application.

The frequency warped version is obtained by replacing the delay elements of the system by allpass filters of first order

$$z^{-1} \longrightarrow A(z) = \frac{1-\alpha z}{z-\alpha}; \quad \alpha \in \mathbb{R};\ |\alpha| < 1. \tag{4}$$

This allpass transformation allows to design a filter-bank whose frequency bands approximate the Bark frequency bands (which model the frequency resolution of the human auditory system) with great accuracy [35]. This can be exploited for speech enhancement to achieve a high (subjective) speech quality with a low number of frequency channels, cf., [30].

The short-term spectral coefficients of the disturbed speech $X(i,\lambda)$ are used to calculate the spectral weights for speech enhancement $W_i(\lambda)$ as well as the weights $\widetilde{W}_i(\lambda)$ for speech denoising prior to the RT estimation, see Figure 1. These spectral weights are converted to the time-domain filter coefficients $w_n(\lambda)$ and $\widetilde{w}_n(\lambda)$ by means of a generalized discrete Fourier transform (GDFT)

$$w_n(\lambda) = \frac{h(n)}{M} \sum_{i=0}^{M-1} W_i(\lambda) e^{-j(2\pi/M)i(n-n_0)}; \quad n, n_0 \in \{0, 1, \ldots L\}, \tag{5}$$

and accordingly for the weights $\widetilde{W}_i(\lambda)$. The sequence $h(n)$ denotes the real, finite impulse response (FIR) of the prototype lowpass filter of the analysis filter-bank. For the common case of a prototype filter with linear phase response and even filter degree L, (5) applies with $n_0 = L/2$. The GDFT of (5) can be efficiently calculated by the fast Fourier transform (FFT). It is also possible to approximate the (uniform or warped) time-domain filters by FIR or IIR filters of lower degree to further reduce the overall signal delay and complexity. A more comprehensive treatment can be found in [30, 31].

4. Spectral Weights for Noise Reduction and Dereverberation

Two essential components of Figure 1 are the calculation of the spectral weights and the RT estimation which are treated in this section.

4.1. Concept. The weights are calculated by the spectral subtraction rule

$$W_i^{(\mathrm{ss})}(\lambda) = 1 - \frac{1}{\sqrt{\hat{\gamma}(i,\lambda)}}; \quad i \in \{0, 1, \ldots, M-1\}. \tag{6}$$

This method achieves a good speech quality with low computational complexity, but other, more sophisticated estimators such as the spectral amplitude estimators of Ephraim and Malah [36] or even psychoacoustic weighting rules [37] can be employed as well, cf., [22].

The spectral weights of (6) depend on an estimation of the *a posteriori signal-to-interference ratio* (SIR)

$$\gamma(i,\lambda) = \frac{|X(i,\lambda)|^2}{\sigma_{z_l}^2(i,\lambda) + \sigma_v^2(i,\lambda)}. \tag{7}$$

The spectral variances of the late reverberant speech and noise are given by $\sigma_{z_l}^2(i,\lambda)$ and $\sigma_v^2(i,\lambda)$, cf., (1) and (2). Equation (6) can be seen as a *generalized* spectral subtraction rule. If no reverberation is present, that is, $z(k) = s(k)$, (7) reduces to the well-known *a posteriori signal-to-noise ratio* (SNR) and (6) to a "common" spectral magnitude subtraction for noise reduction.

The problem of musical tones can be alleviated by expressing the *a posteriori* SIR by the *a priori* SIR

$$\xi(i,\lambda) = \frac{E\left\{|Z_e(i,\lambda)|^2\right\}}{\sigma_{z_l}^2(i,\lambda) + \sigma_v^2(i,\lambda)} = \gamma(i,\lambda) - 1, \tag{8}$$

which can be estimated by the decision-directed approach of [36]

$$\hat{\xi}(i,\lambda) = \eta \cdot \frac{\left|\hat{Z}_e(i,\lambda-1)\right|^2}{\hat{\sigma}_{z_l}^2(i,\lambda-1) + \hat{\sigma}_v^2(i,\lambda-1)} + (1-\eta) \cdot \max\{\hat{\gamma}(i,\lambda) - 1, 0\} \tag{9}$$

with $0.8 < \eta < 1$. This recursive estimation of the *a priori* SIR causes a significant reduction of musical tones, cf., [38]. The spectral weights are finally confined by a lower threshold

$$W_i(\lambda) = \max\left\{W_i^{(\mathrm{ss})}(\lambda), \delta_{\mathrm{w}}(i,\lambda)\right\}. \tag{10}$$

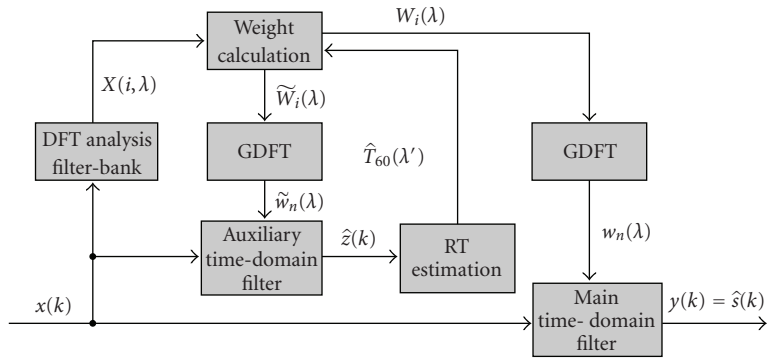

Figure 1: Overall system for low delay noise reduction and dereverberation. The frequency warped system is obtained by replacing the delay elements of the analysis filter-bank and both time-domain filters by allpass filters of first order.

This allows to balance the tradeoff between the amount of interference suppression on the one hand, and musical tones and speech distortions on the other hand. Alternatively, it is also possible to bound the spectral weights implicitly by imposing a lower threshold to the estimated *a priori* SIR. The adaptation of the thresholds and other parameters can be done similar as for "common" noise reduction algorithms based on spectral weighting.

4.2. Interference Power Estimation. A crucial issue is the estimation of the variances of the interfering noise and late reverberant speech to determine the *a priori* SIR. The spectral noise variances $\sigma_v^2(i,\lambda)$ can be estimated by common techniques such as minimum statistics [39].

An estimator for the variances $\sigma_{z_l}^2(i,\lambda)$ of the late reverberant speech can be obtained by means of a simple statistical model for the RIR of (1) [21]

$$h_{\mathrm{m}}(k) = n(k)e^{-\rho k T_{\mathrm{s}}}\epsilon(k) \tag{11}$$

with $\epsilon(k)$ representing the unit step sequence. The parameter $T_{\mathrm{s}} = 1/f_{\mathrm{s}}$ denotes the sampling period and $n(k)$ is a sequence of i.i.d. random variables with zero mean and normal distribution.

The *reverberation time* (RT) is defined as the time span in which the energy of a steady-state sound field in a room decays 60 dB below its initial level after switching-off the excitation source, [40]. It is linked to the *decay rate* ρ of (11) by the relation

$$T_{60} = \frac{3}{\rho \log_{10}(e)} \approx \frac{6.908}{\rho}. \tag{12}$$

Due to this dependency, the terms decay rate and reverberation time are used interchangeably in the following. The RIR model of (11) is rather coarse, but allows to derive a simple relation between the spectral variances of late reverberant speech $\sigma_{z_l}^2(i,\lambda)$ and reverberant speech $\sigma_z^2(i,\lambda)$ according to [21]

$$\sigma_{z_l}^2(i,\lambda) = e^{-2\nu(i,\lambda)T_l} \cdot \sigma_z^2(i,\lambda - N_l). \tag{13}$$

The value $\nu(i,\lambda)$ denotes the frequency and time dependent decay rate of the RIR in the subband-domain whose blind estimation is treated in Section 4.3. The integer value $N_l = \lfloor T_l f_s/R \rfloor$ marks the number of frames corresponding to the chosen time span T_l where f_s denotes the sampling frequency. The value for T_l is typically in a range of 20 to 100 ms and is related to the time span after which the late reverberation (presumably) begins.

The variances of the reverberant speech can be estimated from the spectral coefficients $\hat{Z}(i,\lambda)$ by recursive averaging

$$\hat{\sigma}_z^2(i,\lambda) = \kappa \cdot \hat{\sigma}_z^2(i,\lambda-1) + (1-\kappa) \cdot \left|\hat{Z}(i,\lambda)\right|^2 \tag{14}$$

with $0 < \kappa < 1$. The spectral coefficients of the reverberant speech are obtained by spectral weighting

$$\hat{Z}(i,\lambda) = X(i,\lambda) \cdot \widetilde{W}_i(\lambda) \tag{15}$$

using, for instance, the spectral subtraction rule of (6) based on an estimation of the *a posteriori* SNR. It should be noted that the spectral weights $\widetilde{W}_i(\lambda)$ are also needed for the denoising prior the the RT estimation (see Figure 1).

A more sophisticated (and complex) estimation of the late reverberant speech energy is proposed in [22]. It takes model inaccuracies into account, if the source-receiver distance is lower than the critical distance and requires an estimation of the direct-to-reverberation ratio for this.

4.3. Decay Rate Estimation. The estimation of the frequency dependent decay rates $\nu(i,\lambda)$ of (13) requires non-subsampled subband signals, which causes a high computational complexity. To avoid this, we estimate the decay rate in the time-domain at decimated time instants $\lambda' = \lfloor k/R' \rfloor$ from the (partly) denoised, reverberant speech signal $\hat{z}(k)$ as sketched by Figure 1. The prime indicates that the update rate for this estimation R' is not necessarily identical to that for the spectral weights $W_i(\lambda)$ and $\widetilde{W}_i(\lambda)$. In general, the update intervals for the RT estimation can be longer than for the calculation of the spectral weights as the room acoustics changes usually rather slowly.

The filter coefficients $\widetilde{w}_n(\lambda)$ for the "auxiliary" time-domain filter which provides $\hat{z}(k)$ are obtained by a GDFT of the spectral weights $\widetilde{W}_i(\lambda)$ used in (15), see Figure 1. The frequency dependent decay rates $\nu(i,\lambda')$, needed to evaluate

(13), are obtained by the time-domain estimate of the decay rate $\hat{\rho}(\lambda')$ according to

$$\hat{\nu}(i,\lambda') \approx \hat{\rho}(\lambda') \quad \forall i \in \{0,1,\ldots,M-1\}. \tag{16}$$

This approximation is rather coarse, but it yields good results in practice with a low computational complexity.

A *blind estimation* of the decay rate (or RT) can be performed by a *maximum likelihood* (ML) approach first proposed in [41, 42]. A generalization of this approach to estimate the RT in noisy environments has been presented in [43]. The ML estimators are also based on the statistical RIR model of (11).

For a blind determination of the RT, an ML estimation for the decay rate ρ is performed at decimated time instants λ' on a frame with N samples $\hat{z}(\lambda' R' - N + 1)$, $\hat{z}(\lambda' R' - N + 2), \ldots, \hat{z}(\lambda' R')$ according to

$$\hat{\rho}(\lambda') = \arg\left\{ \max_{\rho} \{\mathcal{L}(\lambda')\} \right\} \tag{17}$$

with the log-likelihood function given by

$$\mathcal{L}(\lambda') = -\frac{N}{2}\left((N-1)\ln(a) + \ln\left(\frac{2\pi}{N} \sum_{i=0}^{N-1} a^{-2i}\, \hat{z}^2(\lambda' R' - N + 1 + i) \right) + 1 \right), \tag{18}$$

where $a = \exp\{-\rho T_s\}$, cf., [43]. The corresponding RT is obtained by (12).

A correct RT estimate can be expected, if the current frame captures a free decay period following the sharp offset of a speech sound. Otherwise, an incorrect RT is obtained, e.g., for segments with ongoing speech, speech onsets or gradually declining speech offsets. Such estimates can be expected to overestimate the RT, since the damping of sound cannot occur at a rate faster than the free decay. However, taking the minimum of the last K_1 ML estimates is likely to underestimate the RT, since the ML estimate constitutes also a random variable. This bias can be reduced by "order-statistics" as known from image processing [44]. In the process, the histogram of the K_1 most recent ML estimates is built and its first local maximum is taken as RT estimate $\hat{T}_{60}^{(\text{peak})}(\lambda')$ excluding maxima at the boundaries. The effects of "outliers" can be efficiently reduced by recursive smoothing

$$\hat{T}_{60}(\lambda') = \beta \cdot \hat{T}_{60}(\lambda' - 1) + (1-\beta) \cdot \hat{T}_{60}^{(\text{peak})}(\lambda') \tag{19}$$

with $0.9 < \beta < 1$. A strong smoothing can be applied as the RT changes usually rather slowly over time.

The devised RT estimation relies only on the fact that speech signals contain occasionally distinctive speech offsets, but it requires no explicit speech offset detection [21] or a calibration period [45]. Another important advantage of this RT estimation is that it is developed for noisy signals as the prior denoising can only achieve a *partial* noise suppression.

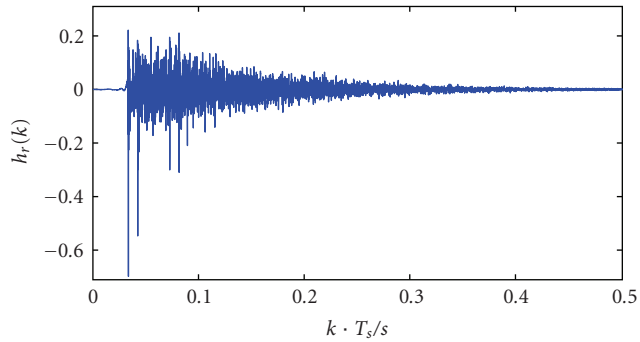

FIGURE 2: Measured RIR with $T_{60} = 0.79$ seconds.

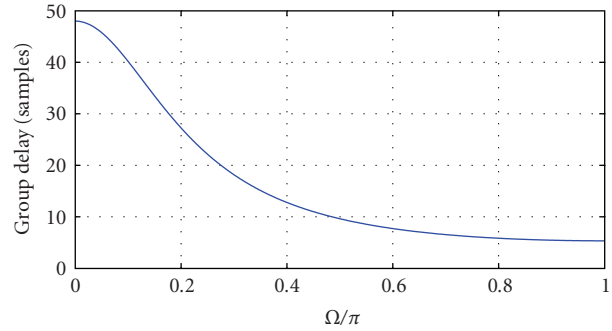

FIGURE 3: Group delay of the warped filter-bank equalizer with filter degree $L = 32$ and allpass coefficient $\alpha = 0.5$.

In principle, it is also conceivable to use other methods for the continuous RT estimation, such as the Schroeder method [46] or a non-linear regression approach [47]. However, the use of such estimators has lead to inferior results as the obtained histograms showed a higher spread and less distinctive local maxima. This resulted in a much higher error rate in comparison to the ML approach.

5. Evaluation

The new system has been evaluated by means of instrumental quality measures as well as informal listening tests. The distorted speech signals are generated according to (1) for a sampling frequency of $f_s = 16\,\text{kHz}$. A speech signal of 6 minutes duration is convolved with a RIR shown in Figure 2. The RIR has been measured in a highly reverberant room and possesses a RT of 0.79 s. (This value for T_{60} has been determined from the measured RIR by a modified Schroeder method as described in [43].) The reverberant speech signal $z(k)$ is distorted by additive babble noise from the NOISEX-92 database with varying global input SNRs for anechoic speech $s(k)$ and additive noise $v(k)$.

For the processing according to Figure 1, a warped filter-bank equalizer is used with allpass coefficient $\alpha = 0.5$, $M = 32$ frequency channels, a downsampling rate of $R = 32$ and a Hann prototype lowpass filter of degree $L = M$. This processing with non-uniform frequency resolution allows to achieve a good subjective speech quality with low signal delay, cf., [30]. The time-invariant group delay of

both warped time-domain filters is shown in Figure 3. The group delay varies only between 0.5 ms and 3.125 ms for $f_s = 16$ kHz. Such variations do not cause audible phase distortions so that a phase equalizer is not needed here. In contrast, the use of a corresponding warped AS FB yields not only a significantly higher signal delay but requires also a phase equalization, see [31].

The spectral weights are calculated by the spectral subtraction rule of (6) using the thresholding of (10) with $\delta_{\mathrm{w}}(i,\lambda) \equiv 0.2$ for the weights $W_i(\lambda)$ and $\delta_{\mathrm{w}}(i,\lambda) \equiv 0.1$ for the weights $\widetilde{W}_i(\lambda)$. The spectral noise variances are estimated by minimum statistics [39] and the variances of the late reverberant speech by (13). For the blind estimation of the RT according to Section 4.3, a histogram size of $K_{\mathrm{l}} = 400$ values and an adaptation rate of $R' = 256$ are used. A smoothing factor of $\beta = 0.995$ is employed for (19).

The quality of the enhanced speech is evaluated in the time-domain by means of the *segmental signal-to-interference ratio* (SSIR) (cf., [48]). The difference between the anechoic speech signal of the direct path $s_{\mathrm{d}}(k)$ and the processed speech $y(k) = \hat{s}(k)$ (after group delay equalization) is expressed by

$$\frac{\mathrm{SSIR}}{\mathrm{dB}} = \frac{10}{\mathcal{C}(\mathbb{F}_s)} \sum_{l\in\mathbb{F}_s} \log_{10}\left(\frac{\sum_{n=0}^{N_{\mathrm{f}}-1} s_{\mathrm{d}}^2(l-n)}{\sum_{n=0}^{N_{\mathrm{f}}-1}\left(s_{\mathrm{d}}(l-n)-y(l-n)\right)^2}\right). \tag{20}$$

The set $\mathbb{F}_s$ contains all frame indices corresponding to frames with speech activity and $\mathcal{C}(\mathbb{F}_s)$ represents its total number of elements.

The speech quality is also evaluated in the frequency-domain by means of the mean *log-spectral distance* (LSD) between the anechoic speech of the direct path and the processed speech according to

$$\frac{\mathrm{LSD}}{\mathrm{dB}} = \frac{1}{\mathcal{C}(\mathbb{F}_s)} \sum_{l\in\mathbb{F}_s} \sqrt{\frac{1}{N_{\mathrm{f}}} \sum_{i=0}^{N_{\mathrm{f}}-1} \left| \mathcal{S}_{s_{\mathrm{d}}}(i,l) - \mathcal{S}_y(i,l) \right|^2} \tag{21}$$

with

$$\begin{aligned} \mathcal{S}_{s_d}(i,l) &= \max\left\{20\log_{10}(|S_{\mathrm{d}}(i,l)|),\ \delta_{\mathrm{LSD}}\right\}, \\ \mathcal{S}_y(i,l) &= \max\left\{20\log_{10}(|Y(i,l)|),\ \delta_{\mathrm{LSD}}\right\}, \end{aligned} \tag{22}$$

where $S_{\mathrm{d}}(i,l)$ and $S_y(i,l)$ denote the short-term DFT coefficients of anechoic and processed speech for frequency index i and frame l. The lower threshold δ_{LSD} confines the dynamic range of the log-spectrum and is set here to -50 dB. Half-overlapping frames with $N_{\mathrm{f}} = 256$ samples are used for the evaluations.

A perceptually motivated spectral distance measure is given by the *Bark spectral distortion* (BSD) [49]. The Bark spectrum is calculated by three main steps: critical band filtering, equal loudness pre-emphasis and a phone-to-sone conversion. The BSD is obtained by the mean difference between the Bark spectra of undistorted speech $\mathcal{B}_{s_{\mathrm{d}}}(i,l)$ and enhanced speech $\mathcal{B}_y(i,l)$ according to

$$\mathrm{BSD} = \frac{\sum_{l\in\mathbb{F}_s}\sum_{i=0}^{N_{\mathrm{f}}-1}\left(\mathcal{B}_{s_{\mathrm{d}}}(i,l) - \mathcal{B}_y(i,l)\right)^2}{\sum_{l\in\mathbb{F}_s}\sum_{i=0}^{N_{\mathrm{f}}-1}\mathcal{B}_{s_{\mathrm{d}}}(i,l)^2}. \tag{23}$$

A modification of this measure is given by the modified Bark spectral distortion (MBSD) which takes also into account the noise masking threshold of the human auditory system [50]. The (M)BSD has been originally proposed for the evaluation of speech codecs, but it can also be used as (additional) quality measure for speech enhancement systems, cf., [22].

The curves for the different measures are plotted in Figure 4. The joint suppression of late reverberant speech and noise yields a significantly better speech quality, in terms of a lower LSD and MBSD as well as a higher SSIR, in comparison to the noise reduction without dereverberation where $\sigma_{z_{\mathrm{l}}}(i,\lambda) = 0$ for (8) and (9), respectively. (Using the cepstral distance (CD) measure led to almost identical results as for the LSD measure.) For low SNRs, the dereverberation effect becomes less significant due to the high noise energy, cf., (8). This is a desirable effect as the impact of reverberation is (partially) masked by the noise in such cases. For high SNRs, the noise reduction alone still achieves a slight improvement as the noise power estimation does not yield zero values. The estimation errors of the blind RT estimation are small enough to avoid noteworthy impairments; the curves for speech enhancement with blind RT estimation are almost identical to those obtained by using the actual RT. (Using other RIRs and noise sequences led to the same results.) Therefore, the new speech enhancement system achieves a speech quality as the comparable approach of [22] which, however, assumes that a reliable estimate of the RT is given (and considers a common DFT AS FB).

The results of the instrumental measurements comply with our informal listening tests. The new speech enhancement system achieves a significant reduction of background noise and reverberation, but still preserves a natural sound impression. The speech signals enhanced with blind RT estimation and known RT have revealed no audible differences. The noise reduction alone achieves only a slightly audible reduction of reverberation.

6. Conclusions

A new speech enhancement algorithm for the joint suppression of late reverberant speech and background noise is proposed which addresses the special requirements of hearing aids. The enhancement is performed by a generalized spectral subtraction which depends on estimates for the spectral variances of background noise and late reverberant speech. The spectral variances of the late reverberant speech are calculated by a simple rule in dependence of the RT. The time-varying RT is estimated blindly (without dedicated excitation signals) from a noisy and reverberant speech signal by means of an ML estimation and order-statistics filtering.

In reverberant and noisy environments, the devised single-channel speech enhancement system achieves a significant reduction of interferences due to late reverberation

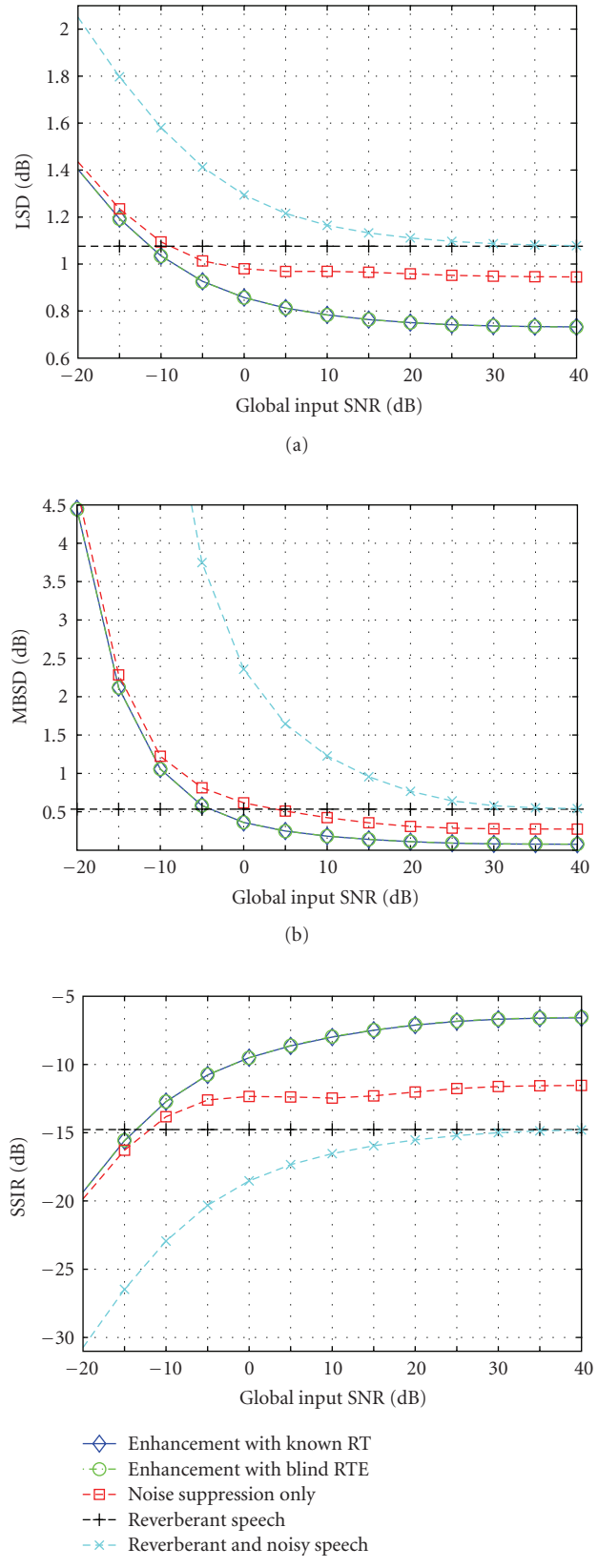

FIGURE 4: Log-spectral distance (LSD), modified Bark spectral distortion (MBSD) and segmental signal-to-interference ratio (SSIR) for varying global input SNRs and different signals.

and additive noise. The enhancement with the blind RT estimation achieves actually the same speech quality as by using the actual RT.

In contrast to existing algorithms for dereverberation and noise reduction, the proposed algorithm has a low signal delay, a reasonable computational complexity and it requires no (large) microphone array, which is of particular importance for speech enhancement in hearing aids. In comparison to commonly used post-filters in hearing aids which only perform noise reduction, a significantly better subjective and objective speech quality is achieved by the devised system.

Although the use for hearing instruments has been considered primarily here, the proposed algorithm is also suitable for other applications such as speech enhancement in hands-free devices, mobile phones or speech recognition systems.

Acknowledgments

The authors are grateful for the support of GN ReSound, Eindhoven, The Netherlands. They would also like to thank the reviewers for their helpful comments as well as the Institute of Technical Acoustics of RWTH Aachen University for providing the measured RIRs.

References

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, Berlin, Germany, 2005.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons, Chichester, UK, 2006.

[3] E. Hänsler and G. Schmidt, Eds., *Speech and Audio Processing in Adverse Environments*, Springer, Berlin, Germany, 2008.

[4] R. A. J. de Vries and B. de Vries, "Towards SNR-loss restoration in digital hearing aids," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. 4004–4007, Orlando, Fla, USA, May 2002.

[5] V. Harnacher, J. Chalupper, J. Eggers, et al., "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2915–2929, 2005.

[6] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds., chapter 3, pp. 39–60, Springer, Berlin, Germany, 2001.

[7] H. W. Löllmann and P. Vary, "Post-filter design for superdirective beamformers with closely spaced microphones," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '07)*, pp. 291–294, New Paltz, NY, USA, October 2007.

[8] S. Doclo and M. Moonen, "GSVD-based optimal filtering for multi-microphone speech enhancement," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds., chapter 6, pp. 111–132, Springer, Berlin, Germany, 2001.

[9] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids," *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 911–925, 2005.

[10] A. K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *Journal of Speech and Hearing Research*, vol. 24, no. 3, pp. 375–383, 1981.

[11] A. K. Nábělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[12] N. D. Gaubitch, P. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 99–102, Kyoto, Japan, September 2003.

[13] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 92–95, Hong kong, April 2003.

[14] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[15] R. Martin, "Small microphone arrays with postfilters for noise and acoustic echo reduction," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds., chapter 12, pp. 255–279, Springer, Berlin, Germany, 2001.

[16] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Communication*, vol. 39, no. 1-2, pp. 111–138, 2003.

[17] T. Yoshioka, T. Nakatani, T. Hikichi, and M. Miyoshi, "Maximum likelihood approach to speech enhancement for noisy reverberant signals," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4585–4588, Las Vegas, Nev, USA, April 2008.

[18] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[19] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: technologies and algorithms," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds., chapter 14, pp. 401–429, John Wiley & Sons, Chichester, UK, 2008.

[20] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 6, pp. 3613–3616, Seattle, Wash, USA, May 1998.

[21] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[22] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. dissertation, Eindhoven University, Eindhoven, The Netherlands, June 2007.

[23] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[24] M. A. Stone and B. C. J. Moore, "Tolerable hearing aid delays—II: estimation of limits imposed during speech production," *Ear and Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[25] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of oversampled uniform DFT filter banks with delay specification using quadratic optimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '01)*, vol. 6, pp. 3633–3636, Salt Lake City, Utah, USA, May 2001.

[26] R. W. Bäuml and W. Sörgel, "Uniform polyphase filter banks for use in hearing aids: design and constraints," in *Proceedings of the 16th European Signal Processing Conference (EUSIPCO '08)*, Lausanne, Switzerland, August 2008.

[27] H. W. Löllmann and P. Vary, "Efficient non-uniform filterbank equalizer," in *Proceedings of European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[28] P. Vary, "An adaptive filter-bank equalizer for speech enhancement," *Signal Processing*, vol. 86, no. 6, pp. 1206–1214, 2006.

[29] J. M. Kates and K. H. Arehart, "Multichannel dynamic-range compression using digital frequency warping," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 3003–3014, 2005.

[30] H. W. Löllmann and P. Vary, "Uniform and warped low delay filter-banks for speech enhancement," *Speech Communication*, vol. 49, no. 7-8, pp. 574–587, 2007.

[31] H. W. Löllmann and P. Vary, "Low delay filter-banks for speech and audio processing," in *Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds., chapter 2, pp. 13–61, Springer, Berlin, Germany, 2008.

[32] T. Gülzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, pp. 5–19, 1998.

[33] I. Cohen, "Enhancement of speech using Bark-scaled wavelet packet decomposition," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 1933–1936, Aalborg, Denmark, September 2001.

[34] T. Fillon and J. Prado, "Evaluation of an ERB frequency scale noise reduction for hearing aids: a comparative study," *Speech Communication*, vol. 39, no. 1-2, pp. 23–32, 2003.

[35] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[36] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[37] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[38] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[39] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[40] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, UK, 4th edition, 2000.

[41] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[42] R. Ratnam, D. L. Jones, and W. D. O'Brien Jr., "Fast algorithms for blind estimation of reverberation time," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 537–540, 2004.

[43] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '08)*, pp. 1–4, Seattle, Wash, USA, September 2008.

[44] I. Pitas and A. N. Venetsanopoulos, "Order statistics in digital image processing," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1893–1921, 1992.

[45] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '08)*, pp. 329–332, Las Vegas, Nev, USA, March-April 2008.

[46] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[47] N. Xiang, "Evaluation of reverberation times using a nonlinear regression approach," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 2112–2121, 1995.

[48] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '05)*, pp. 89–92, Eindhoven, The Netherlands, September 2005.

[49] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[50] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified Bark spectral distortion as an objective speech quality measure," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 1, pp. 541–544, Seattle, Wash, USA, May 1998.