Efficient Speech Dereverberation for Binaural Hearing Aids

Heinrich W. Löllmann and Peter Vary

Institute of Communication Systems and Data Processing, RWTH Aachen University, 52056 Aachen
Email: loellmann@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

An efficient two-stage algorithm for binaural cue preserving speech dereverberation by spectral weighting is presented. The first stage reduces late reverberant speech components where the second stage aims at a suppression of non-coherent early and remaining late reverberant speech components.

The proposed system is based on the binaural speech dereverberation algorithm of Jeub et al. [3]. In contrast to that contribution, the realization of the needed delay-and-sum beamformer and the blind estimation of reverberation time and time-difference-of-arrival with low complexity and low delay is tackled. In addition, the problem of a decreased signal energy for the enhanced signal due to spectral weighting is addressed by an efficient approach.

1 Introduction

The introduction of a binaural data link is a major advance in the development of digital hearing aids (HAs). A full wireless audio data link between left and right HA allows to execute binaural algorithms for signal classification or speech enhancement and leads to a better speech quality in comparison to a bilateral system where both HA devices work independently (apart from a possible exchange of control parameters).

Most binaural speech enhancement algorithms for HAs aim for a reduction of (background) noise with preservation of the binaural cues, e.g., [1, 2]. The binaural cues, i.e., the interaural time and level differences, are important for source localization and can be impaired by bilateral speech enhancement systems. More recently, a binaural cue preserving system for speech dereverberation has been presented [3].

However, the implementation of binaural speech enhancement algorithms in HAs is subject to challenging design constraints: First, speech enhancement algorithms for HAs operate mainly in the frequency-domain where the overall system delay must be rather low (at least below 10 ms) to avoid a detrimental comb-filter effect. As a consequence, low delay filter-banks for such purpose, as proposed, e.g., in [4, 5], can only operate with a very limited number of frequency-bands. A second major constraint is that algorithms for HAs must possess a low algorithmic complexity due to the limited battery size and capacity.

The goal of this paper is to present a binaural speech dereverberation algorithm which addresses the outlined needs for an implementation in HAs. The proposed system builds up on the binaural speech dereverberation algorithm of [3]. In contrast to that algorithm, the proposed system estimates all needed quantities blindly and has a lower complexity and signal delay. In addition, the problem of a decreased signal energy due to speech enhancement by spectral weighting is pointed out and tackled by a new approach.

2 Speech Dereverberation System

A block diagram of the proposed system for binaural speech dereverberation is shown in Fig. 1. The (reverberant) input signals of the left and right HA device, $x_\mathrm{l}(k)$ and $x_\mathrm{r}(k)$ with k denoting the discrete time index, are enhanced by spectral weighting using a DFT analysis-synthesis filter-bank (AS FB), which can be efficiently implemented by a polyphase network with the discrete Fourier transform (DFT) calculated by the fast Fourier transform (FFT) (e.g., [5]). The employed prototype lowpass filter for the analysis and synthesis filter-bank is given by

$$h_0(n) = \frac{\sqrt{R}}{L}\left(1-\sqrt{2}\cos\Big(\frac{\pi}{M}\,(n+0.5)\Big)\right) \qquad (1)$$

with $n \in \{0,1,\ldots,L-1\}$ and filter length $L = 2M$. This filter provides perfect reconstruction for the considered downsampling rate of $R = M/4$. Here, DFT filter-banks with $M = 128$ subbands are used, which results in an overall signal delay of $L = 256$ sample instants. An alternative is to use the low delay filter-bank proposed in [5], which causes a delay of only $L/2$ samples, i.e., a latency of 8 ms for a sampling frequency of $f_s = 16$ kHz.

The left and right subband signals are denoted by $X_\mathrm{l}(\mu,\lambda)$ and $X_\mathrm{r}(\mu,\lambda)$ with subband index μ and decimated time index (frame index) λ. Only the subbands for $\mu = 0,1,\ldots,M/2$ are processed, since $X_\mathrm{l,r}(\mu,\lambda) = X^*_\mathrm{l,r}(M-\mu,\lambda)$ for M being even and real input signals, i.e., there are only $M_\mathrm{eff} = 65$ 'effective' subbands.

2.1 Spectral Weight Calculation

The first stage of the algorithm aims at a reduction of late reverberant speech components where the second stage suppresses non-coherent early and remaining late reverberant speech components. Identical weights are applied to the left and right subband signals to preserve the binaural cues. A detailed treatment of the weight calculation for both stages is provided by [3] and the cited references. Here, the weight calculation is only described as far as needed for the treatment of the new parts, which are marked by thick lines and green filling color in Fig. 1.

The weights of stage I are obtained by spectral magnitude subtraction

$$W_I(\mu,\lambda) = 1 - \frac{\hat{\sigma}_\mathrm{late}(\mu,\lambda)}{|\overline{X}(\mu,\lambda)|}\,. \qquad (2)$$

The spectral variance of the late reverberant speech is estimated by the relation[1]

$$\hat{\sigma}^2_\mathrm{late}(\mu,\lambda) = e^{-2\,T_l\,\rho(\lambda)}\cdot\hat{\sigma}^2_{\overline{x}}(\mu,\lambda-N_l) \qquad (3)$$

[1] This simple estimation rule leads to estimation errors in case of a high direct-to-reverberation-ratio (DRR). An approach to alleviate this problem is proposed in [6] which, however, is not considered here as the required estimation of the DRR is rather complex.

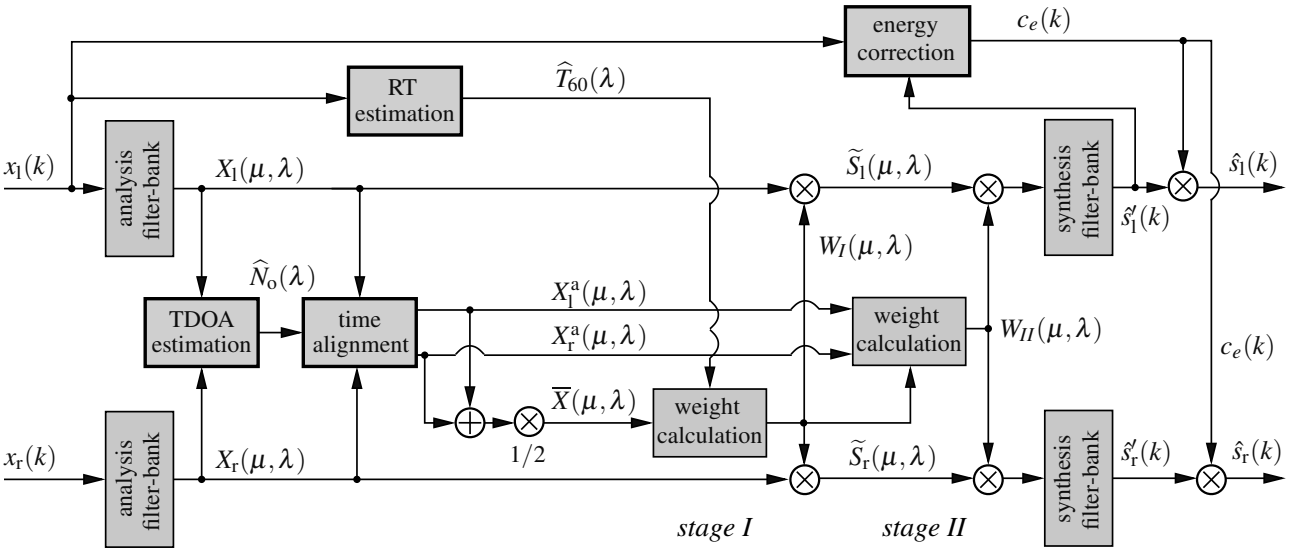

Figure 1: *Overall two-stage system for binaural speech dereverberation by spectral weighting.*

with $N_l = \mathrm{round}\{T_l\, f_s/R\}$. The parameter T_l of Eq. (3) corresponds to the time span after which the late reverberation presumably begins where a value of 50 ms is used here. The decay rate ρ is linked to the *reverberation time* (RT) T_{60} by the relation $\rho = 3\ln(10)/T_{60}$ where the time index is skipped to ease the notation.

The spectral variance of the 'average' signal of left and right channel is calculated by recursive smoothing

$$\hat{\sigma}^2_{\overline{x}}(\mu,\lambda) = \alpha \cdot \hat{\sigma}_{\overline{x}}(\mu,\lambda-1) + (1-\alpha)\cdot\left|\overline{X}(\mu,\lambda)\right|^2 \quad (4)$$

with $\alpha = 0.99$. The averaged subband signals $\overline{X}(\mu,\lambda)$ are obtained by a *delay-and-sum beamformer* (DS BF) operating in the frequency-domain as depicted in Fig. 1.

The spectral weights of stage II are given by [3]

$$W_{II}(\mu,\lambda) = \frac{2\,\mathrm{Re}\left\{\widehat{\Phi}_{\tilde{s}_\mathrm{l}^\mathrm{a},\tilde{s}_\mathrm{r}^\mathrm{a}}(\mu,\lambda)\right\} - \Gamma(\mu)\cdot\left(\widehat{\Phi}_{\tilde{s}_\mathrm{l}^\mathrm{a},\tilde{s}_\mathrm{l}^\mathrm{a}}(\mu,\lambda) + \widehat{\Phi}_{\tilde{s}_\mathrm{r}^\mathrm{a},\tilde{s}_\mathrm{r}^\mathrm{a}}(\mu,\lambda)\right)}{\left(1-\Gamma(\mu)\right)\cdot\left(\widehat{\Phi}_{\tilde{s}_\mathrm{l}^\mathrm{a},\tilde{s}_\mathrm{l}^\mathrm{a}}(\mu,\lambda) + \widehat{\Phi}_{\tilde{s}_\mathrm{r}^\mathrm{a},\tilde{s}_\mathrm{r}^\mathrm{a}}(\mu,\lambda)\right)} \quad (5)$$

where $\Gamma(\mu)$ denotes a real-valued coherence function which takes the shadowing effect of the head into account. This spectral weight calculation requires the cross- and auto-power spectral densities (PSDs) of the *time-aligned* output signals of the first stage. These PSDs can be calculated by recursive averaging similar to Eq. (4) where the needed aligned subband signals $\widetilde{S}_\mathrm{l}^\mathrm{a}(\mu,\lambda)$ and $\widetilde{S}_\mathrm{l}^\mathrm{a}(\mu,\lambda)$ are obtained as follows

$$\widetilde{S}_\mathrm{l}^\mathrm{a}(\mu,\lambda) = X_\mathrm{l}^\mathrm{a}(\mu,\lambda)\cdot W_I(\mu,\lambda) \quad (6a)$$

$$\widetilde{S}_\mathrm{r}^\mathrm{a}(\mu,\lambda) = X_\mathrm{r}^\mathrm{a}(\mu,\lambda)\cdot W_I(\mu,\lambda)\,. \quad (6b)$$

The spectral weights of both stages are confined to lower thresholds, i.e., $W_I(\mu,\lambda) \geq \varepsilon_I$ and $W_{II}(\mu,\lambda) \geq \varepsilon_{II}$, to reduce musical noise artifacts where the threshold $\varepsilon_I = 0.25$ and $\varepsilon_{II} = 0.35$ are used for the later evaluations.

A smoothing of the spectral weights over frequency to avoid musical tones, as done in [3], is not performed to save computational complexity and since the effect of this smoothing is less dominant for a low number of subbands.

It should be noted that the weight calculation can be extended to perform a joint noise reduction and dereverberation, cf., [6, 7]. However, this approach is not considered here for the sake of clarity.

2.2 TODA Estimation and Beamforming

The treatment of Sec. 2.1 has shown that the estimation of the time-difference-of-arrival (TDOA) and the subsequent time-alignment of the subband signals are central elements of the algorithm as they are needed for the weight calculation of both stages. As shown in Fig. 1, these operations are here performed in the frequency-domain using the subband signals of the DFT analysis filter-banks to avoid additional spectral transforms (filter-banks). However, the use of the generalized cross-correlation (GCC) with phase transform (PHAT) weighting [8] for the TDOA estimation (as proposed in [3]) turns out to be less suitable for the used filter-bank with (only) $M_\mathrm{eff} = 65$ unique subbands. This problem is addressed in [9] where an efficient DS BF for binaural speech enhancement is proposed, which operates entirely in the subband-domain. The TDOA estimation for this beamformer is a modification of the GCC-PHAT algorithm to allow for a TDOA estimation with a limited number of frequency-bands and thus employed here.

The time-aligned subband signals $X_\mathrm{r}^\mathrm{a}(\mu,\lambda)$ and $X_\mathrm{l}^\mathrm{a}(\mu,\lambda)$ are obtained by multiplication with the factors

$$\Psi\big(\mu,\widehat{N}_\mathrm{o}(\lambda)\big) = \exp\left\{\frac{-j\,2\,\pi\,\mu\,|\widehat{N}_\mathrm{o}(\lambda)|}{M}\right\} \quad (7)$$

according to

$$X_\mathrm{r}^\mathrm{a}(\mu,\lambda) = X_\mathrm{r}(\mu,\lambda)\cdot\Psi\big(\mu,\widehat{N}_\mathrm{o}(\lambda)\big) \quad (8a)$$

$$X_\mathrm{l}^\mathrm{a}(\mu,\lambda) = X_\mathrm{l}(\mu,\lambda) \quad (8b)$$

for $\widehat{N}_\mathrm{o}(\lambda) \geq 0$ and

$$X_\mathrm{r}^\mathrm{a}(\mu,\lambda) = X_\mathrm{r}(\mu,\lambda) \quad (9a)$$

$$X_\mathrm{l}^\mathrm{a}(\mu,\lambda) = X_\mathrm{l}(\mu,\lambda)\cdot\Psi\big(\mu,\widehat{N}_\mathrm{o}(\lambda)\big) \quad (9b)$$

for $\widehat{N}_\mathrm{o}(\lambda) < 0$ where $\widehat{N}_\mathrm{o}(\lambda)$ denotes the estimated integer-valued TDOA. It is assumed that a source signal coming

from the right-hand side causes a positive TDOA and vice versa.

The presented TDOA estimation tends to a value of zero if it is not able to localize a source, e.g., due to strong noise or multiple speakers. This corresponds to the common default assumption that the desired source is usually in the look direction of the HA user. Thereby, the effects of an erroneous TDOA estimation are less severe than, e.g., for speech enhancement based on beamsteering, since the time-alignment is here employed for PSD calculations.

2.3 Blind RT Estimation

The blind estimation of the time-varying RT $T_{60}(\lambda)$ or decay rate $\rho(\lambda)$, respectively, is required for the calculation of the spectral variance of the late reverberant speech according to Eq. (3). For this purpose, the algorithm of [10] is used due to its low computational complexity in comparison to related approaches, cf., [11]. Two modifications are performed to further reduce the implementation cost for this approach.

One modification is that the RT estimation is performed on the reverberant and possibly noisy input signal $x_\mathrm{l}(k)$ to avoid a costly denoising prior to the RT estimation. This approach is justified by the fact that the employed RT estimation can tolerate a low amount of noise (see also [10]). For strong noise, the speech dereverberation is rather ineffective and should be switched off in such a case.

Another modification is that the approach of [10] to detect fast changing RTs is omitted to further reduce the complexity and memory consumption of the algorithm. Besides, a slow adaptation of the RT value is acceptable as a slight mismatch of the RT estimation has no noticeable influence on the speech quality, cf., Sec. 3.

2.4 Energy Correction

The speech enhancement by spectral weighting leads inevitable to a decreased signal energy, especially if a strong enhancement is performed. This effect is reasoned by the fact that $W_I(\mu,\lambda)W_{II}(\mu,\lambda) \leq 1 \forall \mu,\lambda$. The diminished signal energy can lead to output signals sounding quieter. This can lead to a decreased speech intelligibility, which is of course especially problematic for users of HAs.

This problem is tackled by an efficient approach termed as *energy correction*. As shown in Fig. 1, the output signals of the synthesis filter-banks, $\hat{s}'_\mathrm{l}(k)$ and $\hat{s}'_\mathrm{r}(k)$, are multiplied with a common scaling factor $c_\mathrm{e}(k)$ to obtain the final output signals $\hat{s}_\mathrm{l}(k)$ and $\hat{s}_\mathrm{r}(k)$. The time-varying scaling factor $c_\mathrm{e}(k)$ depends on the smoothed short-term energy ratio of reverberant input speech $x_\mathrm{l}(k)$ and enhanced output speech $\hat{s}'_\mathrm{l}(k)$. This scaling factor is obtained in a first step by

$$C_E(k) = \min\left\{ \frac{E_\mathrm{in}(k)}{E_\mathrm{out}(k)}, C_E^{(\max)} \right\} . \tag{10}$$

Nominator and denominator are calculated by recursive averaging

$$E_\mathrm{in}(k-1) = \beta_1 \cdot E_\mathrm{in}(k-1) + (1-\beta_1) \cdot |x_\mathrm{l}(k)| \tag{11}$$

$$E_\mathrm{out}(k-1) = \beta_1 \cdot E_\mathrm{out}(k-1) + (1-\beta_1) \cdot |\hat{s}'_\mathrm{l}(k)| \tag{12}$$

where a factor of $\beta_1 = 0.95$ is used here. The magnitude instead of the squared value is taken to save multiplications. In Eq. (10), the energy ratio is confined by an upper value

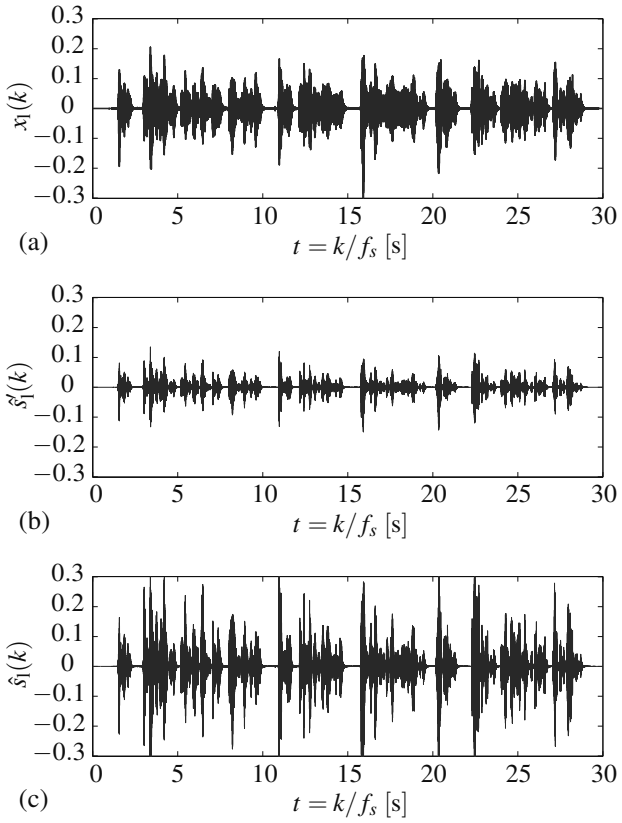

Figure 2: *Reverberated speech ($T_{60} \approx 0.7s$, $\theta = 60°$) of left channel (a) enhanced without energy correction (b) and with energy correction (c).*

$C_E^{(\max)}$ to avoid outliers caused by a denominator close to zero where a value of $C_E^{(\max)} = 25$ has been found suitable.

The values of $C_E(k)$ show strong variations over time, which is counteracted by recursive smoothing

$$c'_\mathrm{e}(k) = \beta_2 \cdot c'_\mathrm{e}(k-1) + (1-\beta_2) \cdot C_E(k) . \tag{13}$$

A rather high smoothing factor of $\beta_2 = 0.999$ is employed as audible speech modulations due to a time-varying scaling factor are much more disturbing than a slight time mismatch of the energy correction.

The final scaling factor is obtained by the limitation

$$c_\mathrm{e}(k) = \begin{cases} 1 & \text{if } c'_\mathrm{e}(k) < 1 \\ c_\mathrm{e}^{\max} & \text{if } c'_\mathrm{e}(k) > c_\mathrm{e}^{\max} \\ c'_\mathrm{e}(k) & \text{otherwise.} \end{cases} \tag{14}$$

The limitation by a lower threshold of 1 precludes an unwanted attenuation of the signal. The upper limit of $c_\mathrm{e}^{\max}$ avoids an overly high amplification where a value of $c_\mathrm{e}^{\max} = 3$ is used here.

An example for the effects of the proposed energy correction is provided by Fig. 2. The reverberant speech was generated by convolving a speech signal with a binaural room impulse response (BRIR) measured in a foyer as explained in Sec. 3 in more detail. It can be seen how the dereverberation leads to a significant reduction of the signal energy such that the enhanced signal sounds much quieter than the input signal. The perceived 'loudness' of the enhanced signal with energy correction was found to be similar to that of the reverberant input signal.

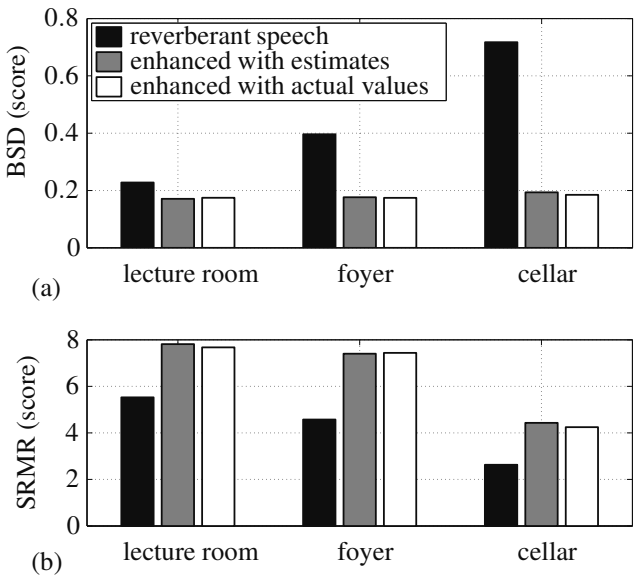

Figure 3: *Bark spectral distortion (BSD) (a) and speech to reverberation modulation energy ratio (SRMR) (b) for binaural speech dereverberation with different setups. The reverberant speech is either enhanced with the actual TDOA and RT values or blindly estimated values.*

3 Simulation Results

A comparison of the underlying binaural speech dereverberation algorithm with related approaches is provided in [3]. Here, the performance of the new efficient modules for beamforming, TDOA and RT estimation is at focus.

The reverberant binaural speech signals were generated by convolving an anechoic speech signal with a BRIR for a sampling frequency of 16 kHz. The BRIRs were measured with hearing aid dummies (Siemens Life 500) mounted on a dummy head using a sweep excitation signal. Setups with different microphone-loudspeaker distances d, directions-of-arrival θ and RTs T_{60} were used: a lecture room ($d = 1.74$ m, $\theta = 0°$, $T_{60} = 0.67$s), a foyer room ($d = 2.6$ m, $\theta = 60°$, $T_{60} = 0.70$s) and a cellar room ($d = 4.4$ m, $\theta = -45°$, $T_{60} = 0.91$s).[2]

The (perceptual) speech quality of reverberant and enhanced speech signals were measured by the speech to reverberation modulation energy ratio (SRMR) [12] and Bark spectral distortion (BSD) [13]. The reference signal for the BSD measure is the speech signal convolved with the direct path of the BRIR, which is determined by the maximum peak of the impulse response plus a time span of 2 ms. The energy correction was switched off for the instrumental evaluation as a scaling falsifies the BSD measure. Silence periods were removed prior to the evaluation.

The obtained results are shown in Fig. 3. The binaural speech dereverberation leads to a significantly improved speech quality in terms of a lower BSD and higher SRMR. The enhancement with the actual values for TDOA and RT achieves the same perceptual quality as the enhancement with estimated quantities. This can be explained by the fact that the TDOA estimation has mostly achieved an accuracy of ± 1 sample and the RT estimation an accuracy of about ± 0.1 ms. This accuracy is sufficient for the presented system and does not cause a perceivable impairment of the speech quality as confirmed by informal listening tests.

[2]The stated RT values are rounded average values obtained from the RTs of each BRIR being determined by the Schroeder method.

4 Conclusions

An efficient algorithm for binaural cue preserving speech dereverberation for hearing aids is presented, which is based on the speech dereverberation algorithm of [3]. In contrast to that, the realization of the needed blind RT and TDOA estimation as well as the time-alignment of the subband signals is now (explicitly) addressed. The proposed algorithm with its blind RT and TDOA estimation can achieve the same (perceptual) quality for the enhanced speech as the use of the actual RT and TDOA values. The new system has a low complexity and is designed to operate with a limited number of effective subbands ($M_{\text{eff}} \leq 65$) to enable subband processing with low delay.

It turns out that binaural speech dereverberation by spectral weighting leads to a significant signal attenuation. A new algorithm to solve this problem with low complexity is proposed, which relies on recursive energy calculations of the input and enhanced output signal.

References

[1] T. Lotter, *Single and Multimicrophone Speech Enhancement for Hearing Aids*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2004.

[2] T. J. Klasen, S. Doclo, T. van den Bogaert, M. Moonen, and J. Wouters, "Binaural Multi-Channel Wiener Filtering for Hearing Aids: Preserving Interaural Time and Level Differences," in *Proc. of Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, vol. 5, pp. 145–148.

[3] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-Based Dereverberation Preserving Binaural Cues," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, Sept. 2010.

[4] R. W. Bäuml and W. Sörgel, "Uniform Polyphase Filter Banks for Use in Hearing Aids: Design and Constraints," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[5] 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, Heidelberg, 2008.

[6] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Eindhoven University, Eindhoven, The Netherlands, 2007.

[7] H. W. Löllmann and P. Vary, "Low Delay Noise Reduction and Dereverberation for Hearing Aids," *EURASIP Journal on Applied Signal Processing*, vol. 2009, pp. 1–9, 2009.

[8] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[9] H. W. Löllmann and P. Vary, "Beamformer for Driving Binaural Speech Enhancement," in *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, Sept. 2012.

[10] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An Improved Algorithm for Blind Reverberation Time Estimation," in *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[11] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance Comparison of Algorithms for Blind Reverberation Time Estimation from Speech," in *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, 2012.

[12] T. Falk and W.-Y. Chan, "A Non-Intrusive Quality Measure of Dereverberated Speech," in *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle (Washington), USA, 2008.

[13] 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, June 1992.