ON BLOCKING MATRIX-BASED DEREVERBERATION FOR AUTOMATIC SPEECH RECOGNITION

Andreas Schwarz, Klaus Reindl, Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7
91058 Erlangen, Germany
{schwarz, reindl, wk}@LNT.de

ABSTRACT

We investigate a two-channel reverberation suppression scheme comprising a blocking matrix for estimating late reverberation components by canceling the direct path and early reflections, and spectral enhancement filters for suppressing the late reverberation components. For idealized blocking matrices, we analyze the influence of the length of the blocking matrix filters and the impact of the coherence between the microphones on the resulting estimate. We show that blocking matrices that cancel more than the direct path of the desired signal can be of advantage for robust speech recognition in highly reverberant environments.

Index Terms— Dereverberation, reverberation suppression, signal enhancement, distant-talking speech recognition

1. INTRODUCTION

In applications like interactive television or gaming, the user usually wants to be untethered while controlling the appliances and hence, distant-talking microphones should be used for speech capture. Then, the acquired microphone signals are distorted by additive noise and reverberation, which leads to serious deterioration of speech intelligibility and speech recognition performance, if no countermeasures are taken. In this contribution, we consider the degradation of the speech signal caused by reverberation. While the early part of the room impulse response (RIR) mainly causes a coloration of the signals, late reverberation causes a temporal smearing of speech features that affects both speech intelligibility and the performance of automatic speech recognition (ASR) systems.

Aiming at increasing the robustness of ASR systems to reverberation, signal based dereverberation algorithms can be applied to the input signals, or reverberation can be addressed in the speech recognizer itself. Signal-based dereverberation schemes can again be differentiated into two categories: inverse filtering algorithms as, e.g., discussed in [1–3], and reverberation suppression approaches, see, e.g., [4].

In this paper, we focus on two-channel late reverberation suppression exploiting a blocking matrix for estimation of late reverberation components. In [5], a similar blocking matrix-based dereverberation scheme was presented, using a delay and subtract beamformer for cancellation of the direct path, and additional parameter estimation to obtain spectral enhancement filters from the beamformer output. The system that we analyze here was first presented in [6] and originally developed for noise and interference suppression [7, 8]. The system uses a blocking matrix (BM) to cancel not only the direct path, but also early reflections. Spectral enhancement filters are then directly derived from the blocking matrix output based on the assumption that late reverberation can be modeled as a diffuse noise field, thus avoiding the estimation of additional parameters.

The purpose of our evaluation is to gain a better understanding of the influence of the blocking matrix filter length on ASR accuracy. Using ideal blocking matrix filters of varying length, we analyze decisive parameters of the spectral enhancement scheme, correlation between early and late reverberation and coherence of the late reverberation, and ASR accuracy.

2. SIGNAL MODEL

The signal model for late reverberation suppression considered in this contribution with stereophonic audio capture is depicted in Fig. 1. Modeling the reverberation in the acoustic environment, the desired source signal s is filtered by a SIMO mixing system. The signal path between the desired source $s[k]$ and microphone $p,\ p \in \{1, 2\}$ is modeled by finite impulse response (FIR) filters of length M (denoted by $h_p[k],\ k = 0, \ldots, M-1$) leading to the sensor signals

$$x_p[k] = \sum_{\kappa=0}^{M-1} h_p[\kappa] s[k-\kappa], \quad p \in \{1, 2\}. \tag{1}$$

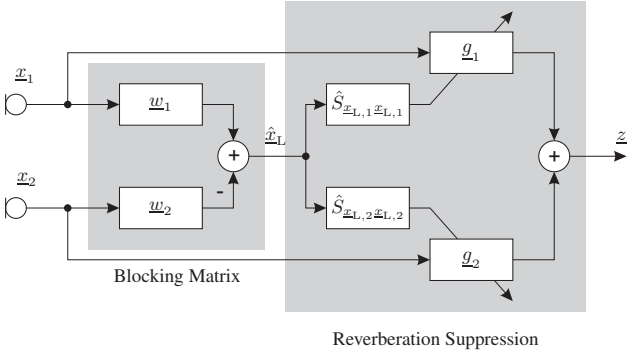

Fig. 1. Signal model for late reverberation suppression combining a blocking matrix and spectral enhancement filters

The FIR filter model with a typical order of several thousands captures reverberation of the acoustic environment. Distinguishing between early reflections and late reverberation, (1) can be reformulated as

$$x_p[k] = \sum_{\kappa=0}^{M_\mathrm{E}-1} h_p[\kappa]s[k-\kappa] + \sum_{\kappa=M_\mathrm{E}}^{M-1} h_p[\kappa]s[k-\kappa], \quad (2)$$

$$= \underbrace{h_{\mathrm{E},p}[k] * s[k]}_{x_{\mathrm{E},p}[k]} + \underbrace{h_{\mathrm{L},p}[k] * s[k]}_{x_{\mathrm{L},p}[k]}, \quad (3)$$

where f_s denotes the sampling frequency, and $M_\mathrm{E} = T_\mathrm{E} f_\mathrm{s}$ defines the time instant where the impulse response is split into a component representing the early reflections $h_{\mathrm{E},p}[k]$ and one representing the late reverberation $h_{\mathrm{L},p}[k]$. The undesired late reverberation component in the microphone signals is denoted by $x_{\mathrm{L},p}[k]$. Using the discrete-time Fourier transform (DTFT), the frequency-domain representation of the acquired microphone signals (3) is expressed as

$$\underline{x}_p(\mathrm{e}^{\mathrm{j}\Omega}) = \underline{h}_{\mathrm{E},p}(\mathrm{e}^{\mathrm{j}\Omega})\underline{s}(\mathrm{e}^{\mathrm{j}\Omega}) + \underline{x}_{\mathrm{L},p}(\mathrm{e}^{\mathrm{j}\Omega}), \quad p \in \{1,2\}, \quad (4)$$

where $\Omega = 2\pi f / f_\mathrm{s}$ is the normalized frequency. To simplify notation, the frequency-dependency $(\mathrm{e}^{\mathrm{j}\Omega})$ is omitted in the rest of the paper as long as ambiguities are precluded.

Using vector/matrix notation, with the superscripts $\{\cdot\}^*$, $\{\cdot\}^\mathrm{T}$, and $\{\cdot\}^\mathrm{H}$ denoting complex conjugation, transposition and conjugate transposition, respectively, the acoustic mixing (4) can be compactly written as

$$\underline{\mathbf{x}} = \underline{\mathbf{h}}_\mathrm{E}\underline{s} + \underline{\mathbf{x}}_\mathrm{L}, \quad (5)$$

where the DTFT-domain signal vectors are defined as

$$\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \underline{x}_2 \end{bmatrix}^\mathrm{T}, \quad (6)$$

$$\underline{\mathbf{x}}_\mathrm{L} = \begin{bmatrix} \underline{x}_{\mathrm{L},1} & \underline{x}_{\mathrm{L},2} \end{bmatrix}^\mathrm{T}, \quad (7)$$

and the DTFT-domain acoustic mixing vector $\underline{\mathbf{h}}_\mathrm{E}$ is given by

$$\underline{\mathbf{h}}_\mathrm{E} = \begin{bmatrix} \underline{h}_{\mathrm{E},1} & \underline{h}_{\mathrm{E},2} \end{bmatrix}^\mathrm{T}. \quad (8)$$

The output signal $\underline{z}$ of the MISO system $\underline{\mathbf{g}}$ for late reverberation suppression is obtained as

$$\underline{z} = \underline{z}_\mathrm{E} + \underline{z}_\mathrm{L} = \underline{\mathbf{g}}^\mathrm{H}\underline{\mathbf{h}}_\mathrm{E}\underline{s} + \underline{\mathbf{g}}^\mathrm{H}\underline{\mathbf{x}}_\mathrm{L}, \quad (9)$$

with the column vector $\underline{\mathbf{g}}$ defined as

$$\underline{\mathbf{g}} = \begin{bmatrix} \underline{g}_1 & \underline{g}_2 \end{bmatrix}^\mathrm{T}. \quad (10)$$

3. REVERBERATION SUPPRESSION

Although the source signal $s[k]$ and the room impulse responses (RIRs) $h_p[k]$, and correspondingly, their DTFT representations, are unknown in practice, an estimate of the desired signal components $\underline{h}_{\mathrm{E},p}\underline{s}$ can be obtained using spectral enhancement techniques. In order to realize the spectral enhancement filter $\underline{\mathbf{g}}$, we propose to apply a BM to suppress the direct path and early reflections first. Defining the filter weights of the BM as $\underline{\mathbf{w}} = [\underline{w}_1\ \underline{w}_2]^\mathrm{T}$, the overall reference of the undesired late reverberation components is given by

$$\underline{\hat{x}}_\mathrm{L} = \underline{\mathbf{w}}^\mathrm{H}\underline{\mathbf{x}} = \underline{\mathbf{w}}^\mathrm{H}\underline{\mathbf{h}}_\mathrm{E}\underline{s} + \underline{\mathbf{w}}^\mathrm{H}\underline{\mathbf{x}}_\mathrm{L}, \quad (11)$$

Our aim is the ideal equalization of the desired components and thereby a cancellation of the desired signal and its early reflections:

$$\underline{\mathbf{h}}_\mathrm{E}^\mathrm{H}\underline{\mathbf{w}} = 0. \quad (12)$$

An ideal solution is given by

$$\underline{w}_1 = \underline{h}_{\mathrm{E},2}, \quad (13)$$

$$\underline{w}_2 = -\underline{h}_{\mathrm{E},1}. \quad (14)$$

Note that this is only one possible solution, as an identification of the room impulse responses is not a general requirement for equalization. In [6,9] we proposed an ICA-based concept for an approximate equalization of the target signal.

The power spectral density (PSD) estimate of the late reverberation reference reads

$$\hat{S}_{\underline{\hat{x}}_\mathrm{L}\underline{\hat{x}}_\mathrm{L}} = |\underline{\mathbf{w}}^\mathrm{H}\underline{\mathbf{h}}_\mathrm{E}|^2 \hat{S}_{\underline{s}\underline{s}} + \underline{\mathbf{w}}^\mathrm{H}\hat{\mathbf{S}}_{\underline{\mathbf{x}}_\mathrm{L}\underline{\mathbf{x}}_\mathrm{L}}\underline{\mathbf{w}}, \quad (15)$$

assuming that the desired components and the undesired late reverberation components are mutually orthogonal. Modeling late reverberation as a spherically isotropic diffuse noise field [10] with a coherence given as

$$\Gamma_{\underline{x}_{\mathrm{L},1}\underline{x}_{\mathrm{L},2}} = \mathrm{sinc}\left(\Omega\ f_\mathrm{s}\ \frac{d}{c}\right),\ \mathrm{sinc}(\cdot) = \frac{\sin(\cdot)}{\cdot}, \quad (16)$$

where c is the speed of sound and d is the microphone spacing, and assuming that the power of late reverberation is identical in both channels, i.e., $\hat{S}_{\underline{x}_{\mathrm{L},1}\underline{x}_{\mathrm{L},1}} = \hat{S}_{\underline{x}_{\mathrm{L},2}\underline{x}_{\mathrm{L},2}}$, the PSD of the channel-specific late reverberation components can be obtained from (15) as [8]

$$\hat{S}_{\underline{x}_{\mathrm{L},p}\underline{x}_{\mathrm{L},p}} = \frac{\hat{S}_{\underline{\hat{x}}_\mathrm{L}\underline{\hat{x}}_\mathrm{L}}}{|\underline{w}_1|^2 + |\underline{w}_2|^2 + 2\Re\{\underline{w}_1\underline{w}_2^*\Gamma_{\underline{x}_{\mathrm{L},1}\underline{x}_{\mathrm{L},2}}\}}. \quad (17)$$

Using (17), optimum Wiener filter weights $\underline{g}_p,\ p \in \{1,2\}$ for late reverberation suppression are derived as

$$\underline{g}_p = 1 - \frac{\hat{S}_{\underline{x}_{\mathrm{L},p}\underline{x}_{\mathrm{L},p}}}{\hat{S}_{\underline{x}_p\underline{x}_p}}, \quad p \in \{1,2\}. \quad (18)$$

In order to balance the suppression of late reverberation and the distortion of the desired early components, the spectral weights (18) are modified with real-valued constants representing a gain factor μ and the spectral floor $\underline{g}_\mathrm{min}$:

$$\underline{g}_p = \max\left[1 - \mu\frac{\hat{S}_{\underline{x}_{\mathrm{L},p}\underline{x}_{\mathrm{L},p}}}{\hat{S}_{\underline{x}_p\underline{x}_p}},\ \underline{g}_\mathrm{min}\right] \quad p \in \{1,2\}. \quad (19)$$

4. EXPERIMENTAL EVALUATION

4.1. Signals and Setup

We use clean speech utterances from the GRID corpus [11] as source signals. The GRID corpus consists of 34000 utterances with a fixed syntax, spoken by 34 different speakers. For the evaluation, we use a test set of 500 randomly selected utterances from the corpus.

The signals are convolved with measured impulse responses of two environments. Room A is a lecture hall ($7\,\mathrm{m} \times 11\,\mathrm{m} \times 3\,\mathrm{m}$) with a reverberation time $T_{60} \approx 900\,\mathrm{ms}$, room B is a large, empty foyer ($54\,\mathrm{m} \times 7\,\mathrm{m} \times 3\,\mathrm{m}$) which is connected to a second floor via an open staircase, and has a correspondingly long reverberation time of $T_{60} \approx 3500\,\mathrm{ms}$. The critical distance d_c is approx. $1\,\mathrm{m}$ in both rooms. The impulse responses are measured using maximum length sequences and truncated to 10000 samples at a sampling rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. We remove bulk delay in all impulse responses,

leaving 16 samples before the main peak to account for the non-causal taps caused by band limitation. The microphones are omnidirectional and spaced 8 cm apart, the source is located at 0 ° (broadside). We investigate source-microphone distances of 1 m, 2 m and 4 m. To evaluate early and late reverberation components separately, we split the impulse responses at the time T_{E} to create the two signal components according to (3).

We use ideal blocking matrices as in (13, 14). The spectral enhancement filter is implemented using a polyphase filterbank with a prototype filter length of 1024, 512 complex-valued subbands, and a downsampling rate of 128. The parameters for the Wiener filter are $\mu = 1.2$ and $\underline{g}_{\min} = 0.1$.

For an objective evaluation, we perform speech recognition experiments with a recognizer based on Pocketsphinx [12]. A finite-state language model is defined according to the structure of the GRID sentences. The recognizer uses triphone HMMs with 3 states per model, 8 Gaussian output densities per state, and a total number of 600 tied states, with features based on 13 mel-frequency cepstral coefficients (MFCCs) with velocity and acceleration. The frame length is 25.6 ms; cepstral mean normalization (CMN) is used to mitigate spectral coloring effects caused by early reverberation. The acoustic model is trained using 32000 clean speech utterances from the GRID corpus (the 500 utterances from the test set are not used for training). No speaker-specific adaptation is performed. For the recognition output, an accuracy score is computed in the same way as in the CHiME challenge [13], where only two words within the sentence (a letter and a digit) are evaluated. For clean speech, we obtain an accuracy of 92.1%.

4.2. Evaluation

First, we investigate how well the coherence of the late reverberation matches the assumption of diffuseness that we use in (17). In Fig. 2, we compare an ideal diffuse coherence function with coherence functions of the late reverberation in the room A at 4 m source-microphone distance for the cutoff times $T_{\mathrm{E}} = 1.2$ ms (corresponding to $M_{\mathrm{E}} = 20$, i.e., removing only the direct path) and $T_{\mathrm{E}} = 31.6$ ms (corresponding to $M_{\mathrm{E}} = 500$, i.e., removing the direct path and some early reflections). We can see that the coherence of the reverberation after 31.6 ms matches the diffuse noise field better than the coherence of reverberation after 1.2 ms. This is to be expected due to the typical distinct reflections in the early part of the impulse response, whereas the latter part mainly contains a mixture of higher-order reflections that better approximate a spherically isotropic and therefore diffuse noise field.

To obtain an indication of the mismatch between our assumption and the late reverberation coherence for varying cutoff time T_{E}, in Fig. 3 we show the mean squared error (MSE) over 0.1 . . . 4 kHz between the ideal diffuse coherence function (16) and the complex coherence function computed from the late part of the measured impulse responses for different cutoff times T_{E}. We observe that, for all investigated scenarios, the mismatch between the assumed model and the measurement reaches its minimum already after about 20 ms.

A second requirement for successful spectral enhancement is orthogonality between desired and undesired signal components, i.e., low correlation between early and late reverberation. We can formulate this correlation based on the source signal autocorrelation (modelling the source signal as a wide-sense stationary process) and the impulse responses of early and late reverberation $h_{\mathrm{E},p}[k]$ and $h_{\mathrm{L},p}[k]$, respectively, as

$$R_{x_{\mathrm{E},p}x_{\mathrm{L},p}}[k] = R_{ss}[k] * h_{\mathrm{E},p}[k] * h_{\mathrm{L},p}[-k]. \tag{20}$$

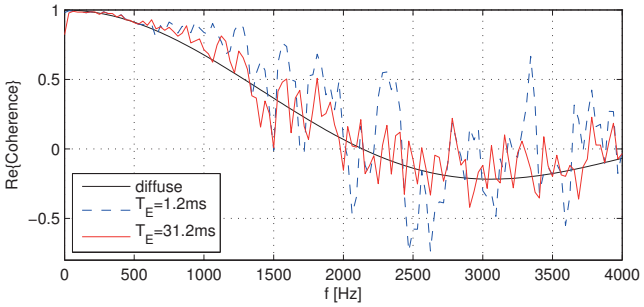

Fig. 2. Real part of coherence of diffuse noise field and late reverberation for different cutoff times T_{E}

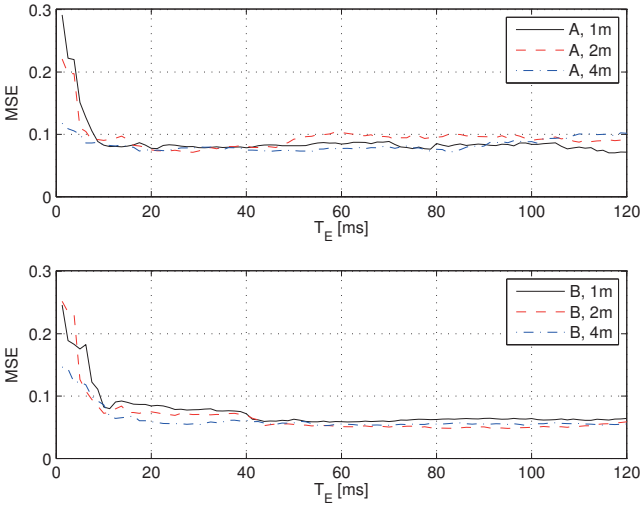

Fig. 3. MSE between ideal and measured coherence of late reverberation for varying cutoff time T_{E} in rooms A and B

For an uncorrelated source signal, the correlation $R_{x_{\mathrm{E},p}x_{\mathrm{L},p}}[0]$ is always zero, since $h_{\mathrm{E},p}[k]$ and $h_{\mathrm{L},p}[k]$ do not overlap. Speech signals, however, have a wider autocorrelation function; therefore, if we consider a small value of T_{E}, where the beginning of the late impulse response still contains strong first-order reflections, the high source signal autocorrelation $R_{ss}[k]$ at k close to zero contributes to $R_{x_{\mathrm{E},p}x_{\mathrm{L},p}}[0]$. Fig. 4 shows the correlation coefficient between the early and late reverberation components, defined using the cross-correlation as well as the autocorrelation functions $R_{x_{\mathrm{L},p}x_{\mathrm{L},p}}[k]$ and $R_{x_{\mathrm{E},p}x_{\mathrm{E},p}}[k]$ as

$$r_p = \frac{R_{x_{\mathrm{E},p}x_{\mathrm{L},p}}[0]}{\sqrt{R_{x_{\mathrm{L},p}x_{\mathrm{L},p}}[0] R_{x_{\mathrm{E},p}x_{\mathrm{E},p}}[0]}}, \tag{21}$$

for varying cutoff time T_{E}, averaged over all utterances and both channels. We can see that the correlation for cutoff times after about 20 ms is significantly lower; therefore, using corresponding long blocking matrix filters allows us to better match the assumption of orthogonality.

Finally, Fig. 5 shows the ASR word accuracy obtained with the proposed system using ideal blocking matrices of varying length. For reference, the baseline values obtained by applying the recognizer to a sum of both unprocessed microphone signals are also shown. We can observe that longer blocking matrix filters, which

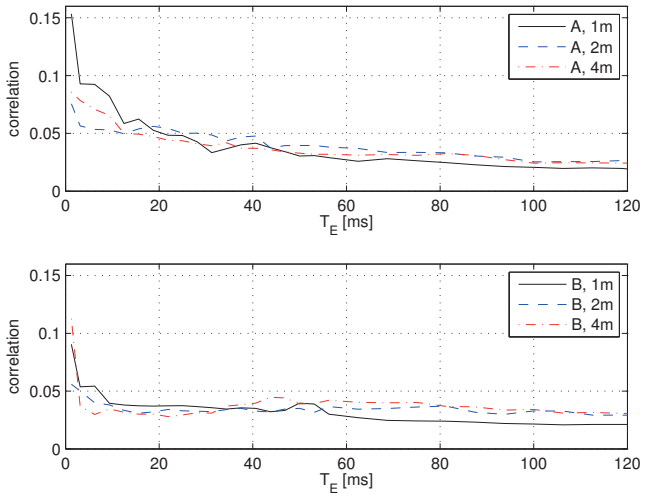

Fig. 4. Correlation coefficient after (21) between early and late reverberation for varying cutoff time T_{E} in rooms A and B

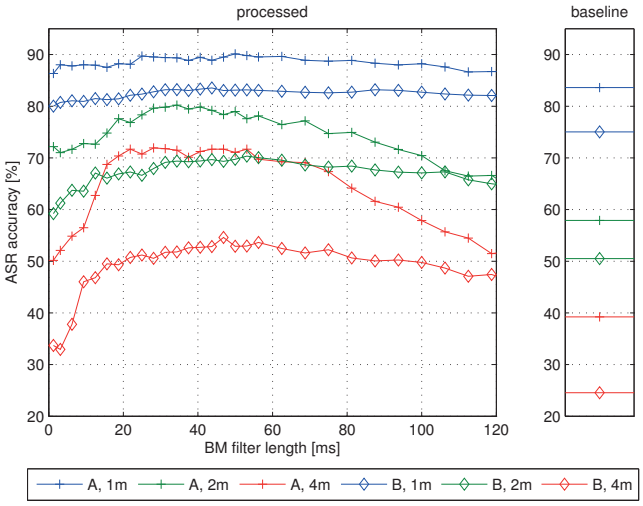

Fig. 5. ASR word accuracy using the proposed system with an ideal blocking matrix of varying length T_{E}, compared to using the unprocessed signal (sum of both microphones), in rooms A and B and for different source-microphone distances

cancel not only the direct path as a delay and subtract beamformer does, but also early reflections, lead to a significant increase in recognition accuracy. This is especially noticeable in the smaller room A due to the presence of more pronounced early reflections. The fact that the results are qualitatively similar for both rooms and all distances, and the width of the maximum, suggest that, as long as the blocking matrix cancels the main reflections of the signal, the exact length is not critical. This confirms what we observed in the analysis of the signal characteristics, where we found that for a cutoff time of about 20 ms, correlation and coherence mismatch from the assumed diffuse model have already largely decayed to the minimum.

5. CONCLUSIONS

In this contribution we presented a further analysis of a previously proposed blocking matrix-based late reverberation suppression scheme [6]. Using ideal blocking matrices, we have shown that the blocking matrix-based approach can lead to a significant increase in ASR accuracy in highly reverberant environments. This holds especially for blocking matrices which cancel not only the direct path of the target source, but also the first $20 \ldots 60$ ms of reflections. These results are in line with those presented in [14, 15], where it was shown in the context of other dereverberation algorithms that it is beneficial to attenuate only the reverberation tail after about 50 ms.

6. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[2] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[3] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Montreal, Canada, May 2004, vol. 3, pp. 889–892.

[4] E.A.P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[5] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Honolulu, USA, Apr. 2007.

[6] A. Schwarz, K. Reindl, and W. Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012.

[7] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *IEEE Int. Symposium Communications, Control, Signal Processing (ISCCSP)*, Limassol, Cyprus, Mar. 2010.

[8] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, Sep. 2011.

[9] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Int. Workshop on Comp. Advances in Multi-Sensor Adapt. Proc. (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009, pp. 253–256.

[10] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, 2000.

[11] M. Cooke, J. Barker, S. Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[12] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. ICASSP*, May 2006.

[13] H. Christensen, J. Barker, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. Interspeech*, 2010.

[14] A. Sehr, E. A. P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[15] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012.