A STUDY ON JOINT BEAMFORMING AND SPECTRAL ENHANCEMENT FOR ROBUST SPEECH RECOGNITION IN REVERBERANT ENVIRONMENTS

Feifei Xiong[1], *Bernd T. Meyer*[2,3] *and Stefan Goetze*[1,3]

[1]Fraunhofer Institute for Digital Media Technology IDMT,
Project Group Hearing, Speech and Audio Technology (HSA), Oldenburg, Germany
[2]Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg, Germany
[3]Cluster of Excellence Hearing4All, Oldenburg, Germany
{feifei.xiong,s.goetze}@idmt.fraunhofer.de, bernd.meyer@uni-oldenburg.de

ABSTRACT

This work evaluates multi-microphone beamforming and single-microphone spectral enhancement strategies to alleviate the reverberation effect for robust automatic speech recognition (ASR) systems in different reverberant environments characterized by different reverberation times T_{60} and direct-to-reverberation ratios (DRRs). The systems consist of minimum variance distortionless response (MVDR) beamformers in combination with minimum mean square error (MMSE) estimators, and late reverberation spectral variance (LRSV) estimators, the latter employing a generalized model of the room impulse response (RIR). Various system architectures are analyzed with a focus on optimal speech recognition performance. The system combining an MVDR beamformer and a subsequent MMSE estimator was found to lead to the best results, with relative reductions of 27.7% compared to the baseline system. This is attributed to a more accurate LRSV estimate from spatial averaging and diffuse field refinement for the MMSE estimator.

Index Terms— Speech dereverberation, minimum variance distortionless response beamformer, minimum mean square error estimator, late reverberation spectral variance, speech recognition

1. INTRODUCTION

The impact of reverberation on spoken language is one of the major problems in automatic speech recognition (ASR), which has been in the focus of many recent studies dealing with robustness in speech processing [1, 2, 3, 4]. Strategies that aim to alleviate the reverberation effect range from speech enhancement and feature extraction over reverberant signal modeling to machine learning approaches, that have lately been shown to strongly improve recognition according to the results from the REVERB Challenge [3]. However, due to the spectral coloration and temporal smearing, reducing the reverberation effect remains a major challenge in front-end audio processing in general, and its application to ASR systems specifically. *Dereverberation* is usually divided into two categories. (a) Reverberation cancellation, which applies a linear filter to the received microphone signals or in front of the loudspeakers based on the estimation of the room impulse responses (RIRs) between the speaker and the microphones [5, 6], and (b) suppression of the late reverberation, which has a major impact on ASR performance [7], by employing a non-linear operation to the received microphone signals in the spectral domain, often ignoring the phases. These dereverberation approaches do not require the complete RIR, but operate on few parameters, e.g. the reverberation time T_{60} [8] and/or the direct-to-reverberation ratio (DRR) [9]. Furthermore, they have been shown to be robust against changes between the speaker and the microphones due to the fact that the late reverberation spectrum is insensitive to these variations [10]. In the single-microphone scenario, a minimum mean square error (MMSE) estimator is commonly employed to suppress the late reverberation spectrum, for which a late reverberation spectral variance (LRSV) estimator is required. If multiple microphones are available, beamforming has been established as a standard since it suppresses noise and reverberation based on the inherent spatial information [11, 1].

This contribution investigates the complementarity of these approaches and explores combinations of late reverberation suppression by beamforming and MMSE-based estimation; ASR word error rates (WERs) are chosen as performance measure, covering (mainly) noise-free conditions in various reverberant environments, i.e. different T_{60} according to different rooms and different DRRs w.r.t. different speaker-microphone distances [12]. More specifically, a minimum variance distortionless response (MVDR) beamformer [13] is used to suppress the influence of late reverberation which allows for a separation from the early reflections as additive noise components [12, 10]. For scenarios in which the speaker-microphone distance is smaller than the critical distance, i.e. if the DRR is larger than 0 dB [12], a generalized LRSV estimator [9] is adopted that considers the direct sound separately to provide a more generalized RIR model compared to Polack's statistical model used in [8]. The combination of the MVDR beamformer and the MMSE estimator is fused by the LRSV estimator, and we investigate two options for this fusion: a direct operation on the multi-channel data and on the output of the beamformer. In a related study, two methods to jointly combine MVDR beamforming and single-channel MMSE estimation have been proposed [14]. However, the focus of [14] was on enhancement rather than the recognition and it was limited to DRRs smaller than 0 dB, whereas this study covers a wide range of DRRs. In [15], a system to jointly suppress the noise and the late reverberation was proposed but the MVDR beamformer and the MMSE estimator were cascaded independently, whereas here, their cooperation is taken into account in the analysis. Furthermore, motivated by the optimal broadband multi-channel MMSE enhancement solution or post-filtering [16], which consists of an MVDR beamformer

This work was partially supported by the projects "Dereverberation and Reverberation of Audio, Music, and Speech" (DREAMS, project no. 316969) and "Knowledge Transfer System" (KNOTS, project no. AAL-2013-6-144) funded by the European Commission (EC) and the BMBF, as well as by the DFG-Cluster of Excellence EXC 1077/1 "Hearing4All".

and a single-channel Wiener filter as a whole, another combination scheme is proposed to explore the late reverberation coherence matrix derived from the LRSV estimate into a generalized post-filter solution [17] instead of the conventional Zelinski approach [18], based on the fact that the late reverberation between channels are inherently and strongly correlated.

The remainder of this paper is organized as follows: Section 2 introduces the system architectures analyzed in this study. Multi-microphone dereverberation including the MVDR beamformer and the post-filtering solution is introduced in Section 3. Section 4 briefly reviews the single-microphone LRSV estimator necessary for different system combinations. The experimental procedure and results of the ASR systems are addressed and discussed in Section 5. Concluding remarks are given in Section 6.

2. SYSTEM COMBINATION STRUCTURES

Based on the mutual influence between the MVDR beamformer and the single-channel MMSE estimator, four different combination strategies are analyzed as depicted in Fig. 1 in order to suppress the reverberation for ASR systems.

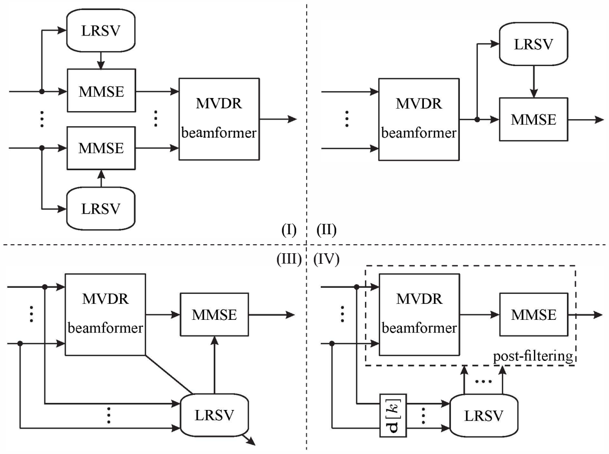

Fig. 1: Four different system combinations (I)-(IV) consisting of MVDR beamformers and MMSE estimators, as well as the respective LRSV estimators.

System (I) shows a straightforward concatenation of the MMSE estimators prior to an independent MVDR beamformer. LRSV estimation operates independently for all the multi-microphone signals and serves as input to the MMSE estimators only. The MMSE estimation does not change the phase information of each microphone signal, i.e. the spatial information for subsequent beamforming is preserved; (II) is another straightforward combination, in which the MMSE estimator follows an independent MVDR beamformer. For analysis, the MVDR-filtered RIR is used by the subsequent MMSE and LRSV estimators. As well, the MMSE and LRSV estimators operate on a single-channel beamformer output, resulting in a low computational complexity; (III) integrates the MVDR beamformer and the successive MMSE estimator by an MVDR refined LRSV version based on a spatially averaged LRSV estimate from all multi-microphone signals. In comparison to (II), this system avoids spatial correlation introduced by the MVDR beamformer output [10] leaking to the LRSV estimator; (IV) illustrates the multi-channel MMSE enhancement scheme according to [16], which can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter. Here the MMSE estimator is actually a post-filter [18, 17].

3. DEREVERBERATION BY MULTI-MICROPHONE BEAMFORMING AND POST-FILTERING

Beamforming and multi-channel post-filtering have been applied to dereverberation in multi-microphone scenarios, cf. e.g. [19, 20, 21, 22]. The MVDR beamformer [13] aims at minimizing the output power of a disturbance while providing a unity gain in the direction of the target source. The filter coefficients of the MVDR beamformer in the short-time Fourier transform (STFT) domain can be derived as

$$\mathbf{W}[\ell,k] = \frac{\mathbf{\Gamma}^{-1}[\ell,k]\mathbf{d}[k]}{\mathbf{d}^{H}[k]\mathbf{\Gamma}^{-1}[\ell,k]\mathbf{d}[k]}\,, \tag{1}$$

with k, ℓ and $(\cdot)^H$ representing the frequency bin, frame index and Hermitian transpose, respectively. $\mathbf{d}[k]$ denotes the steering vector of the target and $\mathbf{\Gamma}[\ell,k]$ is the coherence matrix of the interference signal. Note that the direction of arrival (DOA) for $\mathbf{d}[k]$ is beyond the scope of this paper. Various methods exist in the literature to estimate DOAs, cf. e.g. [23]. For this paper, the noise signal is neglected since the focus is removal of reverberation and the received reverberant signal can be modeled as early reflections and late reverberation, denoted in the following as $X[\ell] = X_e[\ell] + X_l[\ell]$, where k is omitted for simplicity. The coherence matrix $\mathbf{\Gamma}$ in (1) for systems (I)-(IV) in Fig. 1 can be replaced either by the identity matrix $\mathbf{I}_M$ which leads to the delay-and-sum (DS) beamformer [11], or by a diffuse noise field $\mathbf{\Gamma}_{\text{diff}}$ resulting in the superdirective (SD) beamformer [13]. When $\mathbf{\Gamma}_{\text{diff}}$ is chosen, a white noise gain constraint WNG_{max} is used in (1) since the SD beamformer is sensitive to uncorrelated noises [13]. Alternatively, the LRSV coherence $\mathbf{\Gamma}_{\mathbf{X}_l\mathbf{X}_l}$ estimated from M received microphone signals can be also adopted as the coherence matrix $\mathbf{\Gamma}$ in (1) especially for (IV).

(IV) is solved by a generalized post-filter approach [17] due to the inherent correlation among late reverberation $\mathbf{X}_l[\ell]$ from all the microphone signals $\mathbf{X}[\ell]$. The coefficients derived in (IV) can be expressed by an MVDR beamformer with a post-filter [16] as $\mathbf{W}_{\text{IV}}[\ell] = \mathbf{P}[\ell]\mathbf{W}[\ell]$, and the post-filter transfer function [18] is

$$\mathbf{P}[\ell] = \max\left(\frac{\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\tilde{\phi}_{x_e x_e}^{(ij)}[\ell]}{\frac{1}{M}\sum_{i=1}^{M}\tilde{\phi}_{xx}^{(ii)}[\ell]}, P_{\min}\right)\,, \tag{2}$$

where $\tilde{\phi}_{xx}^{(ii)}$ is the auto-correlation of the speech signal in microphone channel i. A lower bound constraint $P_{\min}$ is proposed to alleviate speech distortions for ASR systems. The cross-correlaton item $\tilde{\phi}_{x_e x_e}^{(ij)}$ between ith and jth channel is estimated per frame by [17]

$$\tilde{\phi}_{x_e x_e}^{(ij)} = \frac{\Re\{\tilde{\phi}_{xx}^{(ij)}\} - \frac{1}{2}\Re\{\Gamma_{x_l x_l}^{(ij)}\}(\tilde{\phi}_{xx}^{(ii)} + \tilde{\phi}_{xx}^{(jj)})}{1 - \Re\{\Gamma_{x_l x_l}^{(ij)}\}}\,, \tag{3}$$

where $\Re\{\cdot\}$ calculates the real part of a complex signal. A time alignment is required for $\mathbf{X}[\ell]$ [17], which can be achieved by the steering vector $\mathbf{d}[k]$ as depicted in Fig. 1 (IV). In (3) a first-order recursive update of the auto- and cross-correlation calculations is applied and a maximal threshold is introduced to avoid the denominator being non-positive. $\Gamma_{x_l x_l}$ is the element of the LRSV coherence matrix $\mathbf{\Gamma}_{\mathbf{X}_l\mathbf{X}_l}$, which can be estimated as follows.

4. SINGLE-MICROPHONE LRSV ESTIMATOR

The LRSV estimator plays an important role for the combined systems in Fig. 1. Considering the reverberant situations that the speaker-microphone distances are smaller than the critical distance with positive DRRs [12], as well as a boosted DRR value achieved by the beamformer [24] e.g. in (II), a generalized statistic reverberation model [9] is used here which separates the direct path from Polack's RIR model as used in [8], resulting in the spectral variance of the RIR h in the STFT domain as

$$\lambda_h[\ell] = \begin{cases} \beta_d & \text{for } \ell = 0\,, \\ \beta_r \mathrm{e}^{-2\alpha\ell\tau_s} & \text{for } \ell \geq 1\,, \end{cases} \tag{4}$$

where β_d and β_r denote the variances of the direct path and the reverberant part. τ_s is the STFT time shift (hop size in s), and the decay coefficient α is related to T_{60} by $\alpha = 3\ln(10)/T_{60}$. Accordingly, the DRR can be linked as [9]

$$\mathrm{DRR} = 10\log_{10}\left(\frac{1-\mathrm{e}^{-2\alpha\tau_s}}{\mathrm{e}^{-2\alpha\tau_s}}\,\frac{\beta_d}{\beta_r}\right)\,. \tag{5}$$

Indeed, due to the duration of the time shift τ_s, the DRR value is actually related to the clarity index, e.g. if τ_s is set to 50 ms for a longer-term STFT, (5) represents the value of C_{50} [12]. Using (4), the reverberation variance can be computed by [9]

$$\lambda_r[\ell] = (1-\kappa)\mathrm{e}^{-2\alpha\tau_s}\lambda_r[\ell-1] + \kappa\mathrm{e}^{-2\alpha\tau_s}\lambda_x[\ell-1]\,, \tag{6}$$

where $\kappa = \beta_r/\beta_d$ is calculated from the DRR in (5), constraint in the range of $(0,\ 1]$. Then, the LRSV is given by

$$\lambda_l[\ell] = \mathrm{e}^{-2\alpha\tau_s(L_e-1)}\lambda_r[\ell - L_e + 1]\,, \tag{7}$$

where L_e denotes the number of frames which corresponds to the duration of early reflections of the RIR, usually set to 50 ms [8]. An instantaneous estimate of the input reverberant spectral variance λ_x in (6) can be obtained by a smoothed version of $|X[\ell]|^2$ as

$$\lambda_x[\ell] = \eta\lambda_x[\ell-1] + (1-\eta)|X[\ell]|^2\,, \tag{8}$$

where the smoothing constant η is calculated by $\eta = 1/(1+2\alpha\tau_s)$. According to [10], in order to improve the tracking performance of the reverberant speech onset, η shall be set to be lower as η_{att} when $|X[\ell]|^2 > \lambda_x[\ell-1]$. Note that such LRSV estimator requires a priori information of T_{60} and DRR or clarity index at least in full-band mode, which in practice can be estimated by [25] or be jointly estimated by a trained neural network [26].

In (I)-(III), a single-microphone parameterized MMSE spectral magnitude estimator [27] is used to determine the weighting function $G[\ell]$ to obtain the enhanced speech signal. The a priori early reflection to late reverberation energy ratio required for computing $G[\ell]$ is estimated using the decision-directed approach [28]. Subsequently, the estimated desired signal $\widehat{X}_e[\ell]$ is calculated by

$$\widehat{X}_e[\ell] = \max(G[\ell],\, G_{\min})\, X[\ell]\,, \tag{9}$$

where $G_{\min}$ is a lower bound for the weighting function $G[\ell]$, similar to $P_{\min}$ in (2) which alleviates speech distortions. Then, an inverse STFT is conducted to reconstruct the output speech signal in the time domain used for the subsequent ASR experiments.

(I) uses each received microphone signal directly for the LRSV and the MMSE estimator, which can be considered as independent single-microphone dereverberation schemes. (II) needs the MVDR beamformer output as $X[\ell]$ for calculating (8)-(9). Theoretically, if the RIRs are used to generate the reverberant speech from the clean (anechoic) speech, the MVDR output can be modeled as the clean speech convolved with an MVDR-filtered RIR, expressed as

$$h_{\mathrm{II}} = \sum_{i=1}^{M} h^{(i)} * w_{\mathrm{II}}^{(i)}\,, \tag{10}$$

where $*$ denotes the convolution operation and $h^{(i)}$ represents the RIR of the ith channel. $w_{\mathrm{II}}^{(i)}$ are the time domain MVDR beamformer coefficients in (II). In contrast, (III) estimates the LRSV from each received microphone input, but then a spatially averaged version refined by the MVDR coefficients [14] is applied to the posterior MMSE estimator. The LRSV estimate is refined by [14]

$$\lambda_{l,\mathrm{III}}[\ell] = \overline{\lambda}_l[\ell]\,\mathbf{W}_{\mathrm{III}}^H[\ell]\,\boldsymbol{\Gamma}\,\mathbf{W}_{\mathrm{III}}[\ell]\,, \tag{11}$$

where $\boldsymbol{\Gamma}$ can be $\mathbf{I}_M$ or $\boldsymbol{\Gamma}_{\mathrm{diff}}$ w.r.t. the corresponding beamformer coefficients $\mathbf{W}_{\mathrm{III}}$, and the spatially averaged LRSV $\overline{\lambda}_l$ is calculated by $\overline{\lambda}_l = \sum_{i=1}^{M}\lambda_l^{(i)}/M$. For (IV), a complex coherence matrix $\boldsymbol{\Gamma}_{\mathbf{X}_l\mathbf{X}_l}$ can be applied to the MVDR beamformer calculation in (1), and such a matrix is also required by the post-filter computation in (2). Its elements $\Gamma_{x_l x_l}^{(ij)}$ are defined as $\Gamma_{x_l x_l}^{(ij)} = \tilde{\phi}_{x_l x_l}^{(ij)}/\sqrt{\tilde{\phi}_{x_l x_l}^{(ii)}\tilde{\phi}_{x_l x_l}^{(jj)}}$ [17], where the auto- and cross-correlation items $\tilde{\phi}_{x_l x_l}$ are obtained from the LRSV estimate $\lambda_{l,\mathrm{IV}}$. It is worthwhile noting that the time alignment by $\mathbf{d}[k]$ in Fig. 1 (IV) has no influence on T_{60} and DRR or clarity index values when estimating $\lambda_{l,\mathrm{IV}}$ by (7).

5. EXPERIMENTAL RESULTS

We used the WSJCAM0 British English corpus [29] as database of clean (anechoic) speech. It contains 7861 utterances for training and another 742 for testing at a sampling rate of 16 kHz. 18 real-world RIRs recorded by a circular microphone array ($M = 8$) with 20 cm diameter from the REVERB Challenge [3] were used for multi-condition training mode and another 6 RIRs [3] for generating various test sets (denoted by T1-T6) with different T_{60} and DRR values, as shown in Fig. 2 (a). The STFT has been computed using a 32 ms Hann window with 1/8 overlap, i.e. $\tau_s = 4$ ms, in order to make the hop size narrow enough to represent the direct path. In other words, the direct path contains the early reflections up to 4 ms for calculating the DRR, which is also equivalent to the clarity index C_4. From pilot experiments, the time instance at which the late reverberation starts was set to 48 ms, i.e. $L_e = 12$ in (7). A white noise gain constraint $\mathrm{WNG}_{\max} = 10$ dB was selected for the MVDR beamformers with $\boldsymbol{\Gamma}_{\mathrm{diff}}$ and measured $\boldsymbol{\Gamma}_{\mathbf{X}_l\mathbf{X}_l}$ in (1). The weighting factor of 0.5 was used in the decision-directed approach. η_{att} was chosen as 0.7η in (8). $G_{\min}$ in (9) was set to 0.1 as a good value regarding the ASR performance, which is also in conformance with [30]. $P_{\min}$ in (2) was chosen as 0.1 and the smoothing factor of the first-order recursive filter was 0.875, as well as a maximal threshold of 0.9 used in (3). Directly from the RIRs, full-band DRR or C_4 was calculated accordingly and T_{60} was determined by using Schroeder's method [31].

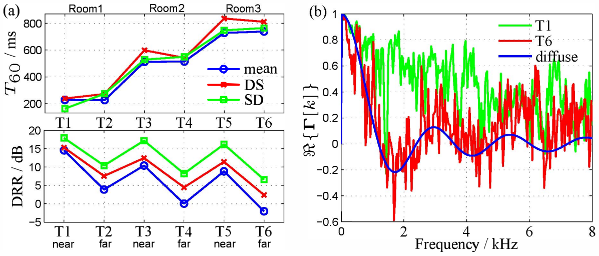

Fig. 2: (a) T_{60} and DRR values of the different test sets T1-T6 (3 rooms and 2 positions), 'mean' denotes the mean values from the 8-microphone RIRs, 'DS' and 'SD' denote the values from DS-filtered and SD-filtered RIRs in (10); (b) shows the real part of the coherence component $\Gamma_{x_l x_l}^{(13)}$ with a specific test utterance from T1 and T6.

	T1	T2	T3	T4	T5	T6	Avg.
clean-cond.	15.44	26.07	28.49	61.76	36.46	78.58	41.13
ideal matched	**12.29**	15.11	16.84	24.66	19.28	31.70	19.98
multi-cond.	15.00	16.91	18.31	27.17	21.02	35.78	22.36
MMSE $i = 1$	14.14	17.36	18.18	25.84	20.57	31.35	21.24
DS $\mathbf{I}_M$	13.42	15.72	15.68	22.50	17.93	30.27	19.25
SD $\mathbf{\Gamma}_{\text{diff}}$	12.86	14.54	14.02	19.90	15.17	25.76	17.04
(I) $\mathbf{I}_M$	13.62	15.71	15.73	21.03	18.06	27.07	18.53
(I) $\mathbf{\Gamma}_{\text{diff}}$	12.90	14.48	14.28	20.30	16.23	25.24	17.23
(II) $\mathbf{I}_M$	13.59	15.12	15.57	21.09	17.74	27.34	18.40
(II) $\mathbf{\Gamma}_{\text{diff}}$	12.61	14.09	14.01	18.76	**14.89**	23.20	16.26
(III) $\mathbf{I}_M$	13.55	15.54	15.51	20.83	17.55	26.86	18.30
(III) $\mathbf{\Gamma}_{\text{diff}}$	12.68	**13.90**	**13.95**	**18.46**	15.01	**23.12**	**16.18**
(IV) $\mathbf{I}_M$	13.63	16.33	16.08	21.46	17.91	26.77	18.69
(IV) $\mathbf{\Gamma}_{\text{diff}}$	12.66	14.32	14.14	19.69	15.59	23.73	16.68
(IV) $\mathbf{\Gamma}_{\mathbf{x}_l\mathbf{x}_l}$	15.03	18.01	18.44	24.27	21.31	29.62	21.11

Table 1: WERs (%) of each test set with different systems. WER with clean-cond. HMMs and clean (anechoic) test data is 11.06%.

The framework for the ASR experiments was implemented based on the Hidden Markov Model Toolkit (HTK) [32]. Overlapping speech segments of 25 ms duration and 10 ms shift were used for feature extraction. Mel-frequency cepstral coefficients with delta and double-delta coefficients as well as cepstral mean and variance normalization were employed. Context-dependent triphone hidden Markov models (HMMs) with 3 states per model were applied together with 12 Gaussian mixture models per state and a language scaling factor of 14.0 for the 5k-word-bigram language model.

As seen in Table 1, lines 2-5 show the WERs with different training modes all in single-channel scenarios with the reference microphone signal $i = 1$, where the single-microphone MMSE-based dereverberation is applied at line 5. For comparison, lines 6-7 give the results with the beamformer-only systems, and the rest reveals the performance of the combined systems (I)-(IV). Even though the ideal matched training models are applied, ASR systems do suffer from the reverberation effect. Generally, compared to the baseline which employs the multi-condition training mode, dereverberation approaches in both single- and multi-microphone scenarios improve the ASR performance, where beamforming techniques reduce absolute average WERs by 2-4% more than the single-microphone MMSE estimator due to the additional spatial advantage brought by multiple microphones (cf. lines 5-7). Moreover, the SD beamformer surpasses the simple DS beamformer by about 2% absolute WER reduction in all the proposed systems, which indicates that diffuse noise field assumption holds for late reverberation, as also illustrated in Fig. 2 (b) for T6 with large T_{60} and low DRR value.

System (I) performs very similar to beamformer-only systems where the MMSE estimators provide a higher benefit when combined with the DS beamformer (cf. lines 6-9). This can be partially explained by distortions of the diffuse field caused by the front MMSE estimators, particularly for the near-position test sets T1, T3 and T5, for which the WERs increase compared to the SD beamformer alone. On the contrary, the *posterior* MMSE estimator in (II) helps the beamforming system to improve the average WERs by approx. 1%. Interestingly, (II) shows that beamformers do boost the DRR values while leaving the T_{60} almost unchanged, as visualized in Fig. 2 (a). Such phenomenon also proves the necessity of separating the direct path from the RIR model in (4) in order to achieve accurate LRSV estimates. Compared to (II), (III) gives slightly better ASR performance, indicating that the spatially averaged LRSV version together with the MVDR refinement in (11) achieves a more accurate LRSV estimate, especially for the diffuse field $\mathbf{\Gamma}_{\text{diff}}$. In other words, the spatial correlation introduced by the beamformer blurs the MVDR-filtered RIR in (10) so that it can not exactly extract the true late reverberation for (II). Overall, average WER reductions of 6.2% and 3.8% are obtained by (III) with the SD beamformer compared to the baseline and even to the ideal matched systems, respectively. Such improvements become more obvious for the far-position test sets such as T4 and T6 than the near-position test sets like T3.

A similar trend can be observed for system (IV), for which the SD beamformer still performs best. It can also be observed that the results further degrade when the beamformer in (1) uses the LRSV coherence matrix $\mathbf{\Gamma}_{\mathbf{x}_l\mathbf{x}_l}$. A possible explanation is that the late reverberation behaves non-stationary and its coherence actually does not match the diffuse property, especially for the near-position test sets, as illustrated in Fig. 2 (b) where the coherence component $\Gamma^{(13)}_{x_l x_l}$ of T1 deviates more severe from the diffuse curve than for the distant position case T6. Indeed, (IV) provides improvement over the baseline as well as the beamformers alone, however, does not offer further robustness compared to (II) and (III). It seems that the diffuse signal suppression by the applied post-filter [17] is not optimal as the spatial coherence assumption is not fully fulfilled [33] and only the magnitude of the spatial coherence matrix is used in (2)-(3). The post-filter design derived from the spatial coherence measure [33] gives better speech enhancement quality, and coherent-to-diffuse ratio estimate has shown its robustness [34] to provide more promising ASR performance for system (IV).

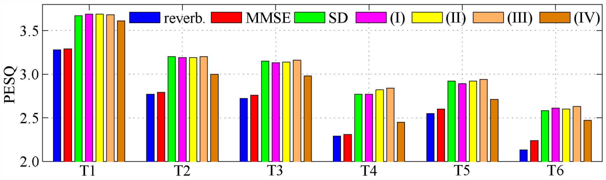

Fig. 3: PESQ scores from the outputs of different system combinations with the clean (anechoic) speech as the reference signal.

In addition, a perceptual evaluation of speech quality (PESQ) [35] has been conducted with the same test utterance used in Fig. 2 (b) as an example. Fig. 3 shows the PESQ scores of the different proposed systems. The scores of the systems with the DS beamformer are not shown since the SD beamformer consistently produces higher PESQ scores, which is in consilience with the WER evaluation in Table 1. In general, multi-microphone dereverberation strategies perform much better than single-microphone approaches, and still, (III) leads to the highest average PESQ score among (I)-(IV) composed by beamforming and MMSE-based filtering.

6. CONCLUSION

This contribution explored possible combination architectures for dereverberation by spectral suppression schemes and (multi-microphone) beamforming with the aim of improving ASR performance in reverberant environments covering a wide range of T_{60} (200 to 800 ms) and DRR (-2 to 15 dB). Results indicate that all the combined systems are able to provide benefits for ASR systems and specifically, the system (III) combining the SD beamformer and the MMSE estimator with the LRSV refinement by the beamformer coefficients achieves 27.7% average relative WER improvement compared to the baseline, as well as 17.3% average relative PESQ boost compared to the reverberant speech signal from the reference microphone. Furthermore, it is also of interest to observe the potential of the spatial late reverberation coherence information to enhance such complex systems that integrate a multitude of (potentially complementary) techniques to deal with reverberation; future work will apply coherence measures to further improve beamforming and the post-filtering.

7. REFERENCES

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons Ltd, 2009.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[4] F. Xiong, N. Moritz, R. Rehr, J. Anemüller, B.T. Meyer, T. Gerkmann, S. Doclo, and S. Goetze, "Robust ASR in Reverberant Environments using Temporal Cepstrum Smoothing for Speech Enhancement and an Amplitude Modulation Filterbank for Feature Extraction," in *the REVERB challenge*, Florence, Italy, May 2014.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[6] S. Goetze, *On the Combination of Systems for Listening-Room Compensation and Acoustic Echo Cancellation in Hands-Free Telecommunication Systems*, Ph.D. thesis, Dept. of Telecommunications, University of Bremen (FB-1), Bremen, Germany, 2013.

[7] A. Sehr, *Reverberation Modeling for Robust Distant-Talking Speech Recognition*, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, Oct. 2009.

[8] 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, May 2001.

[9] E.A.P. Habets, S. Gannot, and I. Cohen, "Late Reverberant Spectral Variance Estimation based on a Statistical Model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

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

[11] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1st edition, 1993.

[12] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.

[13] J. Bitzer and K.U. Simmer, *Microphone Arrays*, chapter Superdirective Microphone Arrays, pp. 19–38, M. Brandstein and D. Ward (Eds.), Springer, Berlin, Heidelberg, May 2001.

[14] E.A.P. Habets, "Towards Multi-Microphone Speech Dereverberation using Spectral Enhancement and Statistical Reverberation Models," in *42nd Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 2008, pp. 806–810.

[15] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint Dereverberation and Noise Reduction using Beamforming and a Single-Channel Speech Enhancement Scheme," in *the REVERB challenge*, Florence, Italy, May 2014.

[16] K.U. Simmer, J. Bitzer, and C. Marro, *Microphone Arrays*, chapter Post-Filtering Techniques, pp. 39–60, M. Brandstein and D. Ward (Eds.), Springer, Berlin, Heidelberg, May 2001.

[17] I.A. McCowan and H. Bourlard, "Microphone Array Post-Filter based on Noise Field Coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–715, Nov. 2003.

[18] R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, NY, USA, Apr. 1988, vol. 5, pp. 2578–2581.

[19] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[20] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[21] M. Jeub and P. Vary, "Binaural Dereverberation based on A Dual-Channel Wiener Filter with Optimized Noise Field Coherence," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4710–4713.

[22] A. Westermann, J.M. Buchholz, and T. Dau, "Binaural Dereverberation based on Interaural Coherence Histograms," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2767–2777, 2013.

[23] A.L. Swindlehurst and T. Kailath, "A Performance Analysis of Subspace-based Methods in the Presence of Model Errors. I. the MUSIC Algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1758–1774, Jul. 1992.

[24] N.D. Gaubitch and P.A. Naylor, "Analysis of the Dereverberation Performance of Microphone Arrays," in *Proc. of the Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005, pp. 121–125.

[25] J. Eaton, N.D. Gaubitch, and P.A. Naylor, "Noise-Robust Reverberation Time Estimation using Spectral Decay Distributions with Reduced Computational Cost," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[26] F. Xiong, S. Goetze, and B.T. Meyer, "Blind Estimation of Reverberation Time based on Spectro-Temporal Modulation Filtering," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 443–447.

[27] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE Spectral Magnitude Estimation for the Enhancement of Noisy Speech," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[28] 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, Dec. 1984.

[29] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, Michigan, USA, May 1995, pp. 81–84.

[30] 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. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 297–300.

[31] M.R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[32] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, Cambridge, 2009, http://htk.eng.cam.ac.uk/.

[33] J.S. Hu and M.T. Lee, "Multi-Channel Post-Filtering based on Spatial Coherence Measure," *Signal Processing*, vol. 105, pp. 338–349, Dec. 2014.

[34] A. Schwarz and W. Kellermann, "Unbiased Coherent-to-Diffuse Ratio Estimation for Dereverberation," in *Proc. of the Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes, France, Sep. 2014.

[35] ITU-T, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," Feb. 2001.