EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A reverberation-time-aware DNN approach leveraging spatial information for microphone array dereverberation

Bo Wu[1], Minglei Yang[1*], Kehuang Li[2], Zhen Huang[2], Sabato Marco Siniscalchi[2,3], Tong Wang[1] and Chin-Hui Lee[2]

Abstract

A reverberation-time-aware deep-neural-network (DNN)-based multi-channel speech dereverberation framework is proposed to handle a wide range of reverberation times (RT60s). There are three key steps in designing a robust system. First, to accomplish simultaneous speech dereverberation and beamforming, we propose a framework, namely DNNSpatial, by selectively concatenating log-power spectral (LPS) input features of reverberant speech from multiple microphones in an array and map them into the expected output LPS features of anechoic reference speech based on a single deep neural network (DNN). Next, the temporal auto-correlation function of received signals at different RT60s is investigated to show that RT60-dependent temporal-spatial contexts in feature selection are needed in the DNNSpatial training stage in order to optimize the system performance in diverse reverberant environments. Finally, the RT60 is estimated to select the proper temporal and spatial contexts before feeding the log-power spectrum features to the trained DNNs for speech dereverberation. The experimental evidence gathered in this study indicates that the proposed framework outperforms the state-of-the-art signal processing dereverberation algorithm weighted prediction error (WPE) and conventional DNNSpatial systems without taking the reverberation time into account, even for extremely weak and severe reverberant conditions. The proposed technique generalizes well to unseen room size, array geometry and loudspeaker position, and is robust to reverberation time estimation error.

Keywords: Deep neural networks (DNNs), Simultaneous speech dereverberation and beamforming, Auto-correlation function, Temporal and spatial contexts, Reverberation-time-aware (RTA)

1 Introduction

In hands-free speech communication systems, the acoustic environment can crucially affect the quality and intelligibility of the speech signal acquired by the microphone(s). In fact, the speech signal propagates through the air and is reflected by the walls, the floor, the ceiling, and any object in the room before being picked up by the microphone(s). This propagation results in a signal attenuation and spectral distortion, called *reverberation*, that seriously degrades speech quality and intelligibility. Many dereverberation techniques have thus been proposed in the past (e.g., [1–5]). One direct way is to estimate an inverse filter of the room impulse response (RIR) [6] to deconvolve the reverberant signal. Wu and Wang [1], Mosayyebpour [2] designed an inverse filter of RIR by maximizing the kurtosis and skewness of the linear prediction (LP) residual, respectively, to reduce early reverberation. However, a minimum phase assumption is often needed, which is almost never satisfied in practice [6]. The RIR can also be varying in time and hard to estimate [7]. Kinoshita et al. [3] estimated the late reverberations using long-term multi-step linear prediction, and then reduced the late reverberation effect by employing spectral subtraction.

Recently, due to their strong regression capabilities, deep neural networks (DNNs) [8, 9] have also been utilized in speech dereverberation. In [10, 11], a DNN-based single-microphone dereverberation system was proposed by adopting a sigmoid activation function at the output layer and min-max normalization of target

*Correspondence: mlyang@xidian.edu.cn
[1]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China
Full list of author information is available at the end of the article

features. An improved DNN dereverberation system we proposed recently [12] adopted a linear output layer and globally normalized the target features into zero mean and unit variance, achieving the state-of-the-art performances.

Microphone array signal processing which utilizes spatial information, is another fundamentally important way for enhancement of speech acquisition in noisy environment [13, 14]. It has recently been shown that the use of the time-varying nature of speech signals could achieve high-quality speech dereverberation based on multi-channel linear prediction (MCLP) [15–17]. Its efficient implementation method which performs in time-frequency-domain, is often referred to as the weighted prediction error (WPE) [15, 16, 18]. The work in [19] designed a feed-forward neural network for mapping microphone array's spatial features into a T-F mask. And [20] utilized a deep neural network (DNN) based multichannel speech enhancement technique, where the speech and noise spectra are estimated using a DNN based regressor and the spatial parameters are derived in an expectation-maximization (EM) like fashion.

In this paper, we aim to provide a robust and efficient DNN-based multi-channel dereverberation framework that takes explicitly and fully advantage of the rich temporal and spatial information provided by the microphone array. A linear uniform array is considered. And possible extension to circular or even ad-hoc array arrangements can be made in the future. First, we propose a single DNN to simultaneously perform beamforming and dereverberation to overcome some of the limitations of the commonly used delay-and-sum beamformers. We refers to the proposed approach as DNNSpatial, because it selectively combines LPS input features of reverberant speech obtained from multiple microphones in an array and map them into the expected output LPS features of anechoic reference speech. The proposed new data utilization strategy based on multi-channel data, leverages upon the complementary information captured in microphone array speech to simultaneously perform beamforming and dereverberation. Its key goal is to discover rich and complex interactions in the signals without any ad-hoc preprocessing but with only data. Different from beamforming, where expert knowledge has to be involved in order to reach the desired result, the DNN will eventually boost the signal in the direction of the desired source, and/or possibly ignore/deemphasize some of the available channels using information available in the data. To better assess the strength of the proposed approach, we also build two standard DNN-based dereverberation systems, namely DSB-DNN and DNNs-DSB, which are direct extensions of the single-microphone case in [12] by combining a beamformer with single or multiple DNN dereverberation systems. The experimental evidence confirms that the proposed DNNSpatial approach outperforms the other two more common DNN-based solutions investigated in this work, in both 2-microphone and 6-microphone settings, according to all of the three objective measures tested. Moreover, the six-microphone array also outperforms the dual-microphone configurations at all reverberation times.

In a single-channel case [12], we found that RT60-dependent frame shift and acoustic context are two key environment-aware parameters in DNN training, which can boost the system's environment robustness. While in multichannel case, we pay more attention to spatial information rather than frame shift, because spatial information captured by microphone array is fundamentally important to speech enhancement of speech acquisition in noisy environment. We next investigate the temporal auto-correlation function of reverberant signals in different reverberant conditions, illustrating a RT60-dependent feature selection is needed to achieve top performances in diverse reverberant cases. Experimental evidence confirms that in stronger reverberant situations, even at a cost of losing spatial contexts, larger temporal contexts with larger aperture size can achieve higher quality dereverberated speech, An environment-aware approach, namely RTA-DNNSpatial, is thus designed to improve the system performance and enhance system robustness, by adopting RT60-dependent temporal-spatial information. Experimental results demonstrate that RTA-DNNSpatial surpasses WPE and DNNSpatial at a wide range of RT60s. It also has good generalization capabilities to unseen room size, array geometry and loudspeaker position, and robustness to RT60 estimation error.

The rest of the paper is organized as follows. We first describe the proposed reverberation-time-aware DNN-based multi-channel dereverberation framework in Section 2. Experimental results are next provided and analyzed in Section 3. The generalization capabilities of the proposed environment-aware model are illustrated in Section 4. Finally we summarize our findings in Section 5.

2 Multi-microphone dereverberation

The well-known delay and sum beamformer (DSB), which is a fundamentally essential method to speech dereverberation [7], is utilized in this paper. It provides us with a foundation from which to explore alternative approaches and techniques. It is of interest also since it has been used as a benchmark for several newly developed dereverberation algorithms [21–23]. Moreover, we adopt an improved DNN dereverberation system we proposed recently [12] by adopting a linear output layer and globally normalizing the target features into zero mean and unit variance. In addition, a linear uniform array is utilized.

2.1 DSB-DNN

Figure 1 shows a potential framework, namely DSB-DNN, consisting of a DSB followed by a single-channel dereverberation DNN model [19, 24]. Here the DNN system is adopted as the mapping function from DSB output to the reference channel's anechoic speech features. This system has a low computational complexity and is easy to accomplish, therefore commonly used in dereverberation applications. The following experimental results show that it is not the optimal choice.

2.2 DNNs-DSB

Another possible solution is that the DNN models are performed first on each channel independently and the resulting outputs are fed to the DSB [20], as illustrated in Fig. 2. It has a very high computational cost, especially when the number of microphones M becomes large. $r_i(t)$ $(i = 1, ..., M)$ is the i-ch reverberant signal, with Y_{r_i} and Φ_{r_i} representing its magnitude and phase spectrums, respectively. $Y_{r_i}^{'}$ denotes the DNN-based dereverberated magnitude. The enhanced waveform $s_i(t)$ is then reconstructed from the estimated spectral magnitude $Y_{r_i}^{'}$ and the phase of reverberant speech Φ_{r_i} with overlap-add method [12]. Finally, the system output $s(t)$ is obtained by performing delay and sum beamforming. The DSB relies on the time delay between $s_i(t)$ and $s_j(t)$ $(i \neq j)$ [7]. However, on the reconstructed signals, neither the magnitude nor phase guarantees the delay assumptions across channels. For example, the phase of the reverberant speech Φ_{r_i} could be different from the phase of reconstructed signal Φ_{s_i}, because for a spectrogram-like matrix in the time-frequency domain, it is not guaranteed there exists a time-domain signal whose STFT is equal to that matrix [25]. Thus the following DSB is not reliable, and the dereverberated signal might have relative worse performance on temporal-domain measures, like frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [26].

2.3 DNNSpatial

Both the DSB-DNN and DNNs-DSB rely on the DSB module, but actually it is possible to embed the beamformer in DNN, which encourages us to design a single-DNN configuration without DSB to accomplish simultaneous speech dereverberation and beamforming. We propose a speech dereverberation framework, namely DNNSpatial, by selectively combining input LPS features of reverberant speech from multiple microphones in an array and map them into the expected output LPS features of anechoic reference speech based on DNNs. The block diagram of DNNSpatial system is illustrated in Fig. 3, while the detailed training procedure is in the bottom panel.

A clean signal $x(t)$ is first passed through a M-microphone array. In our experiments, the received signal of each channel $r_i(t)$ $(i = 1, ..., M)$ is then divided into 32 ms time frame with 16 ms frame shift. A 512-point DFT of each overlapping windowed frame is computed. Then 257-dimension LPS feature vectors [27] are used for DNN training. We use $R_i(k, c)$ to denote the log magnitude in the time-frequency (T-F) unit for time frame k and frequency channel c in the ith channel. Therefore, in the LPS domain, each frame can be represented as a vector $\boldsymbol{r}_i(k)$:

$$\boldsymbol{r}_i(k) = [R_i(k, 1), R_i(k, 2), ..., R_i(k, 257)]^T \tag{1}$$

In order to incorporate temporal dynamics, we include the spectral features of neighboring frames into a feature vector. Therefore, $\boldsymbol{r}_i(k)$ is extended to

$$\widetilde{\boldsymbol{r}}_i(k) = [\boldsymbol{r}_i(k - d_i); ...; \boldsymbol{r}_i(k); ...; \boldsymbol{r}_i(k + d_i)] \tag{2}$$

where d_i denotes the number of neighboring frames in each side in the ith channel. Note that it has been shown in [11] that using the frames in both sides is an optimal feature extension strategy. The input vector is a concatenation of selective speech frames spatially from different microphones and temporally from various acoustic contexts in different microphones. Therefore, the concatenated vector for the DNN feature mapping is:

$$\hat{\boldsymbol{r}}(k) = [\widetilde{\boldsymbol{r}}_1(k); ...; \widetilde{\boldsymbol{r}}_i(k); ...; \widetilde{\boldsymbol{r}}_M(k)] \tag{3}$$

The desired output of the neural network is the spectrogram of anechoic reference speech $y(t) = x(t - t')$ (t' is the time delay between loudspeaker and reference microphone) in the current frame k, denoted by a 257-dimensional feature vector, whose elements correspond to the log magnitude in each frequency bin at the kth frame. The channel between the loudspeaker and the first microphone is considered as the reference channel; therefore the target feature is denoted as $\boldsymbol{y}_1(k)$.

A fixed-length context window is used in order to ensure a fair comparison and a controlled environ-

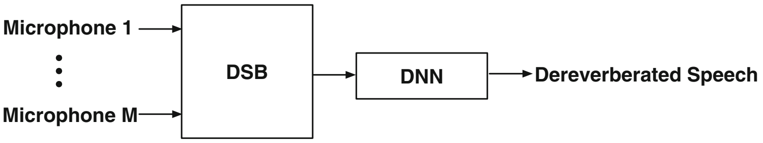

Fig. 1 A block diagram of the DSB-DNN system

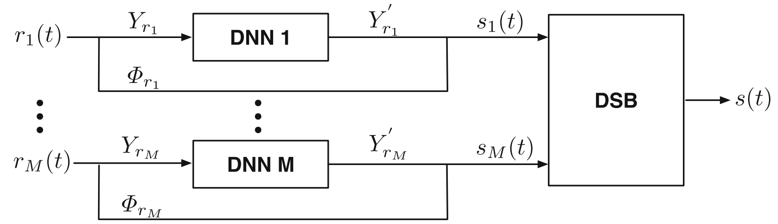

Fig. 2 A block diagram of the DNNs-DSB system

ment (i.e., $\sum_{i=1}^{M}(2d_i + 1)$ is a fixed value). It also avoids a brute-force approach that always selects the same number of frames for each microphone, which may eventually leads to a too big DNN input size as the number of microphone increases and thus causes dramatic performance deterioration [28]. The proposed approach has a low computational complexity, since only a single DNN is used. It is also independent on the delay assumptions because no beamformer is utilized.

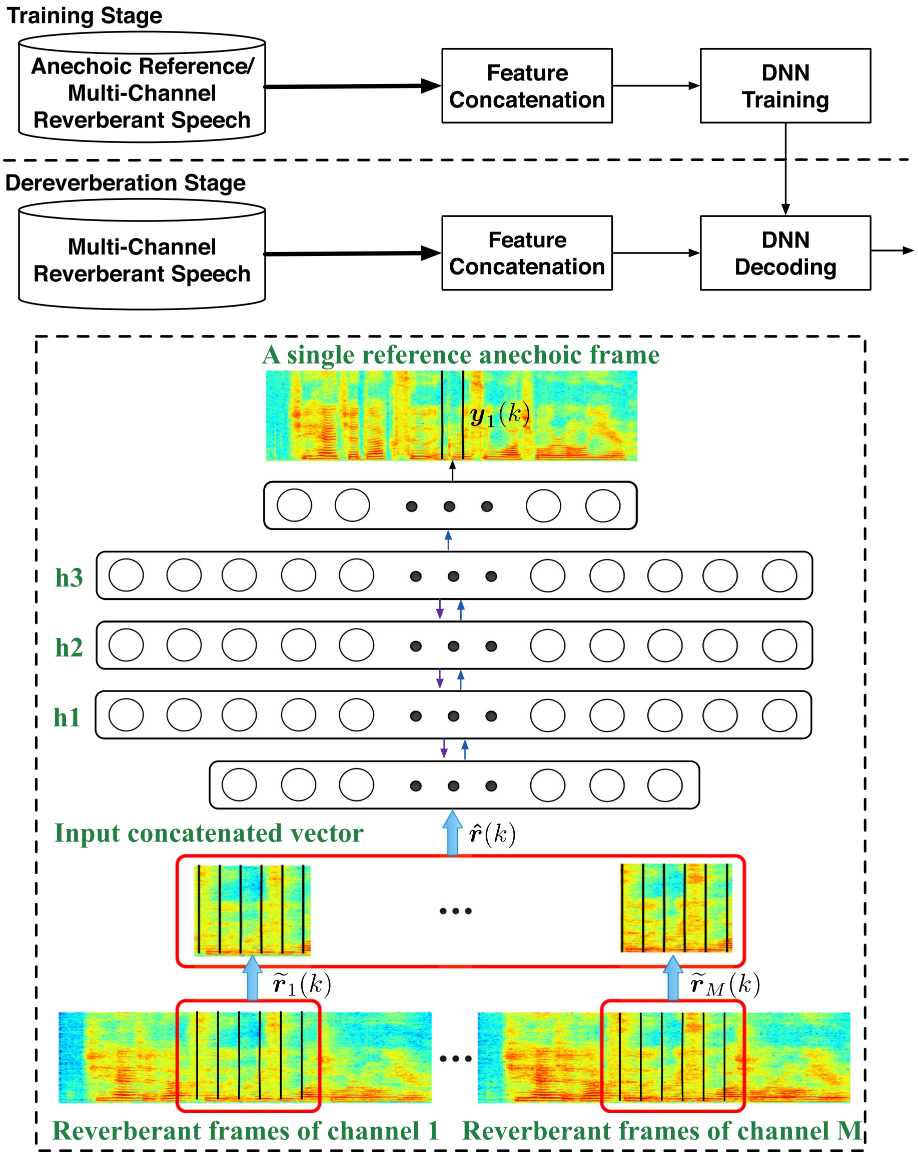
Training Stage
Anechoic Reference/ Multi-Channel Reverberant Speech
Feature Concatenation
DNN Training
Dereverberation Stage
Multi-Channel Reverberant Speech
Feature Concatenation
DNN Decoding
A single reference anechoic frame
$y_1(k)$
h3
h2
h1
Input concatenated vector
$\hat{r}(k)$
$\tilde{r}_1(k)$
$\tilde{r}_M(k)$
Reverberant frames of channel 1
Reverberant frames of channel M

Fig. 3 A block diagram of the proposed DNNSpatial system

2.4 Reverberation-time-aware DNNSpatial

In DNNSpatial, the input vector is a concatenation of speech frames selected temporally (d_i) and spatially (i) from various acoustic contexts in different microphones, and thus there are different possible combinations of temporal and spatial features. Therefore, the effects of temporal-spatial feature combinations on DNNSpatial-based dereverberation performance should be investigated.

2.4.1 Temporal auto-correlation function at different RT60s

In reverberant environments, the source speech, $x(t)$, and the corresponding received signal at the ith microphone, $r_i(t)$, can be related by

$$r_i(t) = x(t) * h_i(t), \tag{4}$$

where $h_i(t)$ is the RIR of the ith channel, which is assumed to be a time-invariant system, and $*$ denotes convolution. Then the room transfer function $H_i\left(e^{j2\pi f}\right)$ is found by taking the Fourier transform of $h_i(t)$, which can be expressed in terms of a direct and a reverberant component, i.e.,

$$H_i\left(e^{j2\pi f}\right) = H_{d,i}\left(e^{j2\pi f}\right) + H_{r,i}\left(e^{j2\pi f}\right). \tag{5}$$

where f denotes the signal frequency.

It is further assumed that the direct and reverberant components are statistically uncorrelated, and thus, the energy spectral density can be written as

$$E\left\{\left|H_i\left(e^{j2\pi f}\right)\right|^2\right\} = \left|H_{d,i}\left(e^{j2\pi f}\right)\right|^2 + E\left\{\left|H_{r,i}\left(e^{j2\pi f}\right)\right|^2\right\}. \tag{6}$$

The direct component is given by [29]

$$H_{d,i}\left(e^{j2\pi f}\right) = \frac{e^{jQD_i}}{4\pi D_i}, \tag{7}$$

where D_i is the distance between the source and the ith microphone. The wave number $Q = 2f/c$ with c being the sound velocity in air.

From the statistical room response model, the reverberant component, $E\left\{\left|H_{r,i}\left(e^{j2\pi f}\right)\right|^2\right\}$, can be written as [29]

$$E\left\{\left|H_{r,i}\left(e^{j2\pi f}\right)\right|^2\right\} = \frac{1-\alpha}{\pi A\alpha}, \tag{8}$$

where α and A are the average wall absorption coefficient and total wall surface area, respectively.

Using Eqs. (7) and (8), we obtain the energy spectral density

$$E\{|H_i\left(e^{j2\pi f}\right)|^2\} = \frac{1}{16\pi^2 D_i^2} + \frac{1-\alpha}{\pi A\alpha}. \tag{9}$$

The temporal auto-correlation function of $r_i(t)$ is given as follows [30],

$$\phi_{r_i r_i}(\tau) = \phi_{xx}(\tau) * R_{h_i}(\tau), \tag{10}$$

where $\phi_{r_i r_i}(\tau)$ and $\phi_{xx}(\tau)$ represent the auto-correlation function of $r_i(t)$ and $x(t)$, respectively. And, $R_{h_i}(\tau) \triangleq h_i(\tau) * h_i(-\tau)$ is called the deterministic auto-correlation function of $h_i(t)$ [31]. The Fourier transform can be written as the following equation:

$$\Phi_{r_i r_i}\left(e^{j2\pi f}\right) = \Phi_{xx}\left(e^{j2\pi f}\right) \times \Phi_{h_i h_i}\left(e^{j2\pi f}\right), \tag{11}$$

where $\Phi_{r_i r_i}\left(e^{j2\pi f}\right)$, $\Phi_{xx}\left(e^{j2\pi f}\right)$ and $\Phi_{h_i h_i}\left(e^{j2\pi f}\right)$ denote the Fourier transforms of $\phi_{r_i r_i}(\tau)$, $\phi_{xx}(\tau)$, and $R_{h_i}(\tau)$, respectively. Because the autocorrelation function and energy spectral density are a pair of Fourier transform for a random energy signal [32],

$$\Phi_{h_i h_i}\left(e^{j2\pi f}\right) = E\left\{\left|H_i\left(e^{j2\pi f}\right)\right|^2\right\}. \tag{12}$$

Moreover, the energy of $\phi_{r_i r_i}(\tau)$ can be calculated by using Parseval's theorem [33]

$$\begin{aligned} E_{\phi_{r_i r_i}} &= \int_{-\infty}^{\infty} |\phi_{r_i r_i}(\tau)|^2 d\tau \\ &= \int_{-\infty}^{\infty} |\Phi_{r_i r_i}\left(e^{j2\pi f}\right)|^2 df \\ &= \int_{-\infty}^{\infty} \left|\Phi_{xx}\left(e^{j2\pi f}\right)\right|^2 \left|\Phi_{h_i h_i}\left(e^{j2\pi f}\right)\right|^2 df. \end{aligned} \tag{13}$$

Substituting Eqs. (9) and (12) into Eq. (13), then we can obtain

$$E_{\phi_{r_i r_i}} = \int_{-\infty}^{\infty} \left|\Phi_{xx}\left(e^{j2\pi f}\right)\right|^2 \left(\frac{1}{16\pi^2 D_i^2} + \frac{1-\alpha}{\pi A\alpha}\right)^2 df. \tag{14}$$

Clearly, $E_{\phi_{r_i r_i}}$ is inversely proportional to α ($0 < \alpha \leq 1$) and RT60 is also inversely proportional to α [34]. As a result, RT60 is proportional to $E_{\phi_{r_i r_i}}$. Therefore, a higher RT60 will result in a larger $E_{\phi_{r_i r_i}}$. Generally, $\phi_{r_i r_i}$ will decrease with the increase of τ, so there will be more energy at high autocorrelation lags in more severe reverberation.

Figure 4 shows an utterance's temporal auto-correlation function, corrupted by reverberation in a simulated room of dimension 6 by 4 by 3 m (length by width by height) at RT60 = 0.2 and 2 s, respectively. Since the energy of the reverberant utterance will significantly affect the auto-correlation, we scale $\phi_{rr}(0)$ to 1 to fairly compare the auto-correlations of the received signals at different reverberant environments. The positions of the loudspeaker and microphone are at (2, 3, 1.5) and (4, 1, 2) meters. It can be observed that $\phi_{rr}(\tau)$ at high autocorrelation lags at RT60 = 2 s is much stronger than that at RT60 = 0.2 s, which is consistent with the above theoretical analysis.

Consequently, in more severe reverberation, the temporal correlation of the consecutive reverberant frames will become stronger. Thus, the temporal context needs

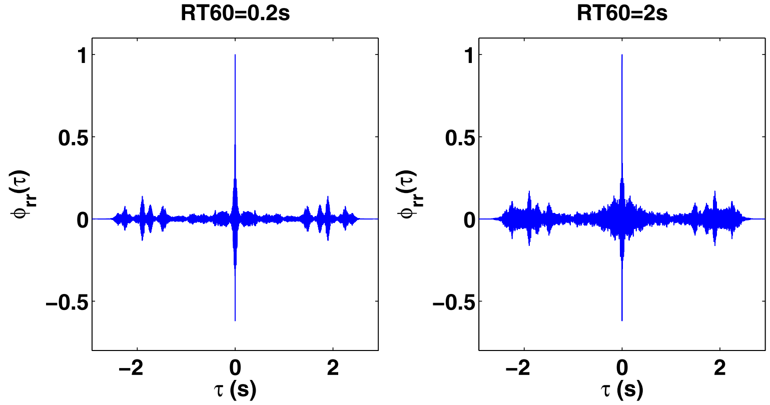

Fig. 4 Temporal auto-correlation function of an utterance in TIMIT dataset [51], corrupted by reverberation at RT60 = 0.2 and 2 s, respectively

to be more emphasised; however it may cause the loss of spatial context because of the fixed-length context window scenario, which has been adopted in this paper for a fair comparison among different approaches, as discussed in Section 3.2. While in weaker reverberant conditions, the spatial context can be guaranteed by decreasing the redundant temporal context. Based on the above findings, an approach exploiting the combinations of temporal and spatial information at different RT60s, should be more appropriate in multi-channel dereverberation systems.

2.4.2 *Channel selection with less spatial contexts*

In stronger reverberant conditions, the temporal context should be emphasised, resulting in less spatial contexts, i.e., only a subset of microphones is available. To achieve better performances, we use the array aperture size as a measure for selecting the channels.

For example, in a fixed uniform linear array, the beamwidth is [13]

$$\theta = 2cos^{-1}\left[\frac{c}{Mdf}\right], \tag{15}$$

where M and d denote the number of microphones and the spacing between neighboring sensors, respectively. $L = Md$ is the array aperture. Therefore, the spatial discrimination capability depends on the array aperture size, i.e., discrimination improves with a larger aperture size [13]. If we assume only 2 elements in the M-microphone array are available in highly reverberant cases, in order to get the largest aperture size, the ith and Mth channels with the largest spacing should be chosen.

In addition, in order to avoid spatial aliasing, the spacing between neighboring sensors has to satisfy the spatial sampling theorem [35, 36], i.e.,

$$d \leq \frac{\lambda}{2} = \frac{c}{2f}, \tag{16}$$

where λ is the wavelength of the speech signal. Since microphone signals are naturally broadband [37], one should sample at half of the wavelength corresponding to the smallest wavelength (or highest temporal frequency) of interest. For example, for a two-element array, the spacing is only about 2.1 cm to prevent aliasing for up to 8 kHz. Clearly, spatial aliasing is somewhat of a misunderstood phenomenon [38], since the human binaural auditory system does not experience problems localizing broadband sounds with an average spacing of 20 cm (corresponding to aliasing above 850 Hz). And it has been revealed in [38] that the spatial Nyquist criterion has little importance for microphone arrays. Therefore, in array design, we could expect a large aperture size by setting a large spacing.

For the purpose of improving the system performance and enhancing system robustness, we propose an environment-aware framework, namely RTA-DNNSpatial, by incorporating the temporal-spatial characteristics at distinct reverberant conditions and the channel selection strategy. A block diagram of the proposed RTA-DNNSpatial system is illustrated in Fig. 5 which is an improved version over our proposed DNNSpatial system illustrated in Fig. 3, by integrating the RT60-dependent temporal and spatial contexts (red parts), into training and dereverberation. In the training and dereverberation stages, the feature selection of temporal and spatial contexts is dependent on the utterance-level RT60, while an RT60 estimator is required in the dereverberation stage. A detailed description of how RT60 affects the combinations of temporal and spatial features will be presented later in Section 3.2.1.

3 Experiments and result analysis

3.1 DNNSpatial

The experiments were conducted in a simulated room of dimension 6 by 4 by 3 meters (length by width by

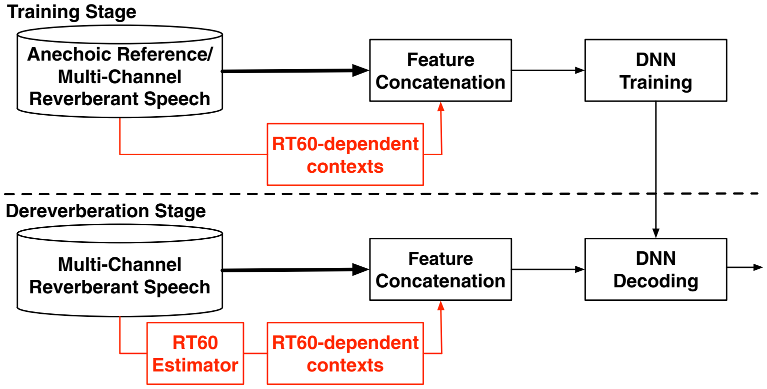

Fig. 5 A block diagram of the proposed RTA-DNNSpatial system

height). The position of the loudspeaker was at (2, 3, 1.5) meters. Both 2-microphone and 6-microphone arrays were considered. For the 2-microphone array, the positions of the microphones were at (4, 1, 2) and (4, 1.2, 2) meters, respectively. For the 6-microphone array, the positions of the microphones were at (4, 1, 2), (4, 1.1, 2), (4, 1.2, 2), (4, 1.3, 2), (4, 1.4, 2), and (4, 1.5, 2) meters. Ten RIRs were simulated using an improved image-source method (ISM) [39] with reverberation time (RT60) ranging from 0.1 to 1.0 s, with an increment of 0.1 s. To learn a high-quality DNN model, all 4620 training utterances from the TIMIT set were convolved with the generated RIRs to build a large multi-condition training set, resulting in about 40 h of reverberant speech at each microphone. To test DNN's generalization capability in mismatch conditions, RIRs with RT60 from 0.1 to 1.0 s with the increment of 0.05 s (rather than 0.1 s) were convolved with 100 randomly selected utterances from the TIMIT test set to construct the test set. This resulted in a collection of 19 × 100 reverberant utterances at each microphone.

Kaldi [40] was used to train DNNs, with 3 hidden layers, 2048 nodes for each layer. The number of pre-training epochs for each RBM [41] layer was 1. The learning rate of pre-training was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. The configuration parameters were chosen according to a previous investigation on speech enhancement [42]. Input and target features of DNN were globally normalized to zero mean and unit variance [43].

In addition, frequency weighted segmental SNR (fwSegSNR) [26], short-time objective intelligibility (STOI) [44], and perceptual evaluation of speech quality (PESQ) [45] were used to evaluate the system performance.

3.1.1 Single-microphone DNN-based dereverberation systems

We first show the performances of our recently proposed single-channel DNN dereverberation system "DNN-baseline" in [12] with 11 frames of input feature expansion; another single-channel DNN dereverberation system "DNN-HWW" in [10] without post-processing [25]; the signal processing dereverberation method "WPE" with its single channel mode. The WPE code is available at http://www.kecl.ntt.co.jp/icl/signal/wpe/index.html. "Rev" represents unprocessed reverberant speech.

Figure 6 illustrated that when compared to unprocessed reverberant speech, our proposed DNN-baseline could achieve a significant fwSegSNR improvement of 3.5 dB on the average at all RT60s, including mismatched conditions of RIRs and unseen speakers. This results demonstrated that our proposed single-channel DNN system had both powerful regression and generalization capabilities. Furthermore, the proposed system was superior to WPE and DNN-HWW at each RT60, achieving average fwSegSNR increase of 2.5 and 1.2 dB, respectively.

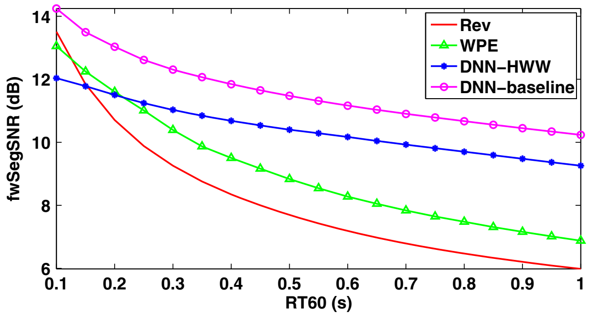

Fig. 6 Average single-microphone fwSegSNR of WPE, DNN-HWW, and DNN-baseline at different RT60s

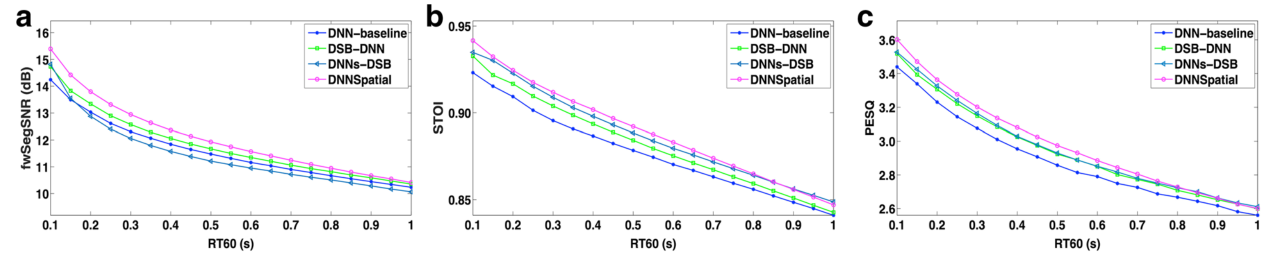

Fig. 7 Average 2-microphone fwSegSNR (**a**), STOI (**b**) and PESQ (**c**) of three DNN-based array dereverberation systems at different RT60s

3.1.2 Multi-microphone DNN-based dereverberation systems

Then we show the performances of three DNN-based array dereverberation systems (i.e., DSB-DNN, DNNs-DSB, and DNNSpatial) in a 2-microphone array. For DNN models in both DSB-DNN and DNNs-DSB, a 11-frame context window was considered during DNN training. While for DNNSpatial, a 10-frame context window was utilized (i.e., $\sum_{i=1}^{2}(2d_i + 1) = 10$), to ensure a fair comparison and a controlled environment. The possible configurations for feature selection in DNNSpatial are (9, 1), (7, 3), and (5, 5). Specifically, (9, 1) implies the number of frames in acoustic context for the first and second microphone is 9 and 1, respectively (i.e., $d_1 = 4$, $d_2 = 0$). We referred to these three configurations as DNNSpatial9-1, DNNSpatial7-3, and DNNSpatial5-5, respectively. We selected the last one as our DNNSpatial configuration without any special purpose. And note that the phase used to do reconstruction in DNN-based systems were the phase of the delay-and-sumed speech signal, so that all systems are comparable regardless of phase issue. As shown in Fig. 7, compared with DNN-baseline, although DSB-DNN could substantially increase the fwSegSNR scores, the improvements were small. DNNs-DSB only improved fwSegSNR below RT60 $\leq$0.15 s, and showed decreases at all other RT60s tested, as explained in Section 2.2. While our proposed DNNSpatial could significantly improve the speech quality and achieved the best fwSegSNR scores at all RT60s. In terms of STOI and PESQ, the proposed DNNSpatial still could obtain better scores than DSB-DNN and DNNs-DSB at almost all RT60s, demonstrating DNNSpatial performed better than the other two DNN-based array dereverberation models. Moreover, the proposed DNNSpatial framework generalizes well to unseen RT60s.

Figure 8 shows the performances of three DNN-based array dereverberation systems in a 6-microphone array. And, DNNSpatial5-1-1-1-1-1 ($d_1 = 2$, $d_2 = d_3 = d_4 = d_5 = d_6 = 0$, $\sum_{i=1}^{6}(2d_i + 1) = 10$) was adopted in the experiment without any special purpose. It still could obtain the best scores at almost all RT60s, in terms of fwSegSNR, STOI, and PESQ, demonstrating its robustness to the array configuration. Moreover, the 6-microphone array also outperformed dual-microphone configurations at all RT60s according to the three objective measures tested.

3.2 Reverberation-time-aware DNNSpatial (RTA-DNNSpatial)

In Section 3.2.1, our proposed DNNSpatial models were trained to estimate the two contexts, i and d_i, needed to achieve top performances for each RT60. In Section 3.2.2, we further explore a RTA-DNNSpatial system that consider the effects of feature selection by adopting the "optimal" temporal and spatial contexts at each RT60, which is assumed to be known in the dereverberation stage.

3.2.1 Temporal and spatial contexts in feature selection

The following experimental settings were the same as in Section 3.1. The larger microphone array that consisted of six elements was considered. Three-thousand seventy-two nodes for each layer were used to train DNNs. The

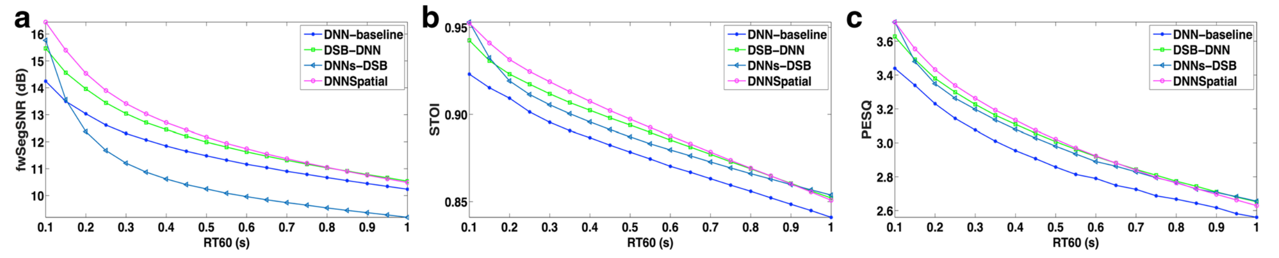

Fig. 8 Average 6-microphone fwSegSNR (**a**), STOI (**b**) and PESQ (**c**) of three DNN-based array dereverberation systems at different RT60s

Table 1 Average 6-microphone fwSegSNR of different DNNSpatial configurations at different RT60s

RT60 (s)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
Rev	13.51	10.71	9.26	8.35	7.69	7.19	6.79	6.47	6.20	5.98	5.78	5.62	5.47	5.35	5.24	5.14	5.05	4.98	4.90	4.84
WPE	14.78	12.15	11.22	10.72	10.33	10.03	9.74	9.47	9.25	9.03	8.80	8.62	8.40	8.23	8.03	7.82	7.65	7.47	7.32	7.15
DNNbaseline	13.52	12.54	11.91	11.48	11.15	10.90	10.67	10.50	10.32	10.16	10.01	9.88	9.74	9.62	9.51	9.40	9.30	9.21	9.10	9.01
DNNSpatial3-3-1-1-3-3	**15.68**	**14.16**	**13.23**	12.55	12.04	11.61	11.26	10.97	10.70	10.46	10.25	10.05	9.86	9.68	9.52	9.36	9.22	9.08	8.95	8.83
DNNSpatial5-1-1-1-1-5	15.65	14.13	13.21	**12.56**	**12.05**	**11.66**	**11.32**	**11.04**	10.79	10.57	10.38	10.19	10.01	9.84	9.69	9.56	9.43	9.31	9.20	9.08
DNNSpatial7-0-0-0-0-7	15.10	13.74	12.92	12.36	11.93	11.58	11.27	11.03	**10.80**	**10.60**	**10.39**	**10.23**	**10.06**	**9.92**	**9.78**	**9.66**	**9.53**	**9.42**	**9.30**	**9.20**

RT60 was also extended to 2.0 s, in order to explicitly explore the effects of temporal and spatial variations in feature selection on the dereverberation performance.

Table 1 presents the average fwSegSNR scores of a series of sampled DNNSpatial configurations at different RT60s. The numbers in bold denote maximum values at each RT60 among all DNNSpatial models. "Rev" represents unprocessed reverberant speech. The results in Table 1 show that:

(i) A 14-frame context window was used to study how RT60 affected the temporal-spatial features. To ensure a controlled environment and differentiable temporal-spatial contexts, three symmetric configurations of DNNSpatial3-3-1-1-3-3, DNNSpatial5-1-1-1-1-5, and DNNSpatial7-0-0-0-0-7 were investigated. Obviously, DNNSpatial7-0-0-0-0-7 had the longest temporal context but the least spatial context, and the largest aperture array size having only two elements.

Clearly the performances were affected greatly by the combinations of temporal-spatial contexts as these bold numbers show the optimal contexts vary with RT60s. Specially, for conditions of RT60$\geq$ 0.9 s, it was surprising that even at a price of losing some spatial contexts, DNNSpatial7-0-0-0-0-7 still could perform better than DNNSpatial3-3-1-1-3-3 and DNNSpatial5-1-1-1-1-5. This could be explained by the theoretical analysis in Section 2.4 that, in a stronger reverberant environment, (a) larger temporal contexts should be more emphasized because of the stronger temporal correlation of the consecutive reverberant frames; (b) the channels with larger spacing should be selected in order to obtain the largest aperture size, which could compensate the loss of spatial contexts. Similar results were obtained in terms of STOI and PESQ. A "optimal" RT60-dependent feature selection of temporal-spatial contexts was then obtained by extracting the best DNNSpatial configuration at each RT60 in Table 1, illustrated in Table 2.

(ii) The DNNbaseline with 15-frame feature extension was also considered, which was actually a special case of DNNSpatial15-0-0-0-0-0. It was not surprising that at RT60=2.0 s, DNNbaseline with the largest temporal context was superior to DNNSpatial3-3-1-1-3-3. This was also consistent with our findings in Section 2.4.

(iii) Compared with the state-of-the-art signal processing multi-channel dereverberation method WPE with the 6-channel mode, our proposed DNNSpatial frameworks (3-3-1-1-3-3, 5-1-1-1-1-5, 7-0-0-0-0-7) could significantly improve the fwSegSNR by about 1.6 dB on the average at all RT60s tested.

3.2.2 RTA-DNNSpatial (Known RT60)

Inspired by the findings in Table 1, a RTA-DNNSpatial[1] was established by incorporating RT60-dependent temporal-spatial contexts. In the training stage, the utterance-based speech frames spatially from different microphones and temporally from various acoustic contexts in different microphones were concatenated by the optimal temporal and spatial contexts, as shown in Table 2. In the dereverberation stage, as the RT60s were assumed to be known, the test data could be processed in the same manner as the training utterances.

The results in Fig. 9 show that, compared with DNNSpatial systems, RTA-DNNSpatial model could achieved the best fwSegSNR, STOI, and PESQ scores at almost all known RT60s, including extremely weak (RT60 = 0.1 s) and severe (RT6 = 2.0 s) reverberant cases. That indicates that the proposed environment-aware approach is robust enough to handle the slightly and highly reverberant situations. Furthermore, RTA-DNNSpatial worked even better at low RT60s.

4 Discussions on generalization capabilities

Since the DNN is a data-driven mechanism, it is important to evaluate our proposed RTA-DNNSpatial's generalization capabilities to situations not seen in training. Therefore, we directly evaluate RTA-DNNSpatial (obtained in Section 3.2.2) without retraining in a series of mismatched conditions that are most commonly considered in practical applications. fwSegSNR, which is a speech intelligibility indicator [46], was used to evaluate the system performances.

Table 2 RT60-dependent temporal-spatial contexts

RT60 (s)	0.10∼0.30	0.40∼0.80	0.90∼2.00
Strategy	3-3-1-1-3-3	5-1-1-1-1-5	7-0-0-0-0-7

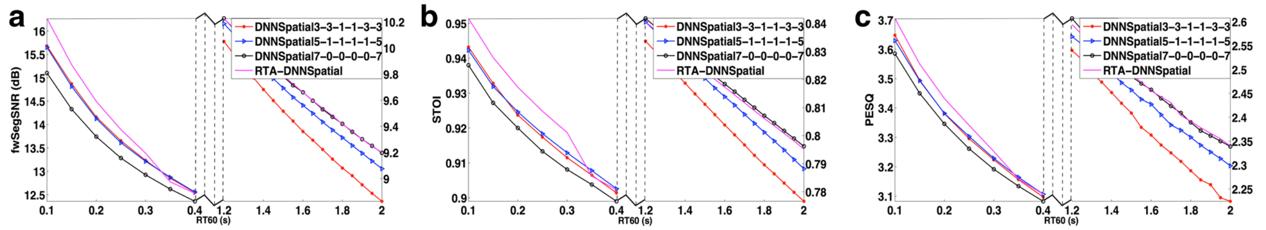

Fig. 9 Average 6-microphone fwSegSNR (**a**), STOI (**b**) and PESQ (**c**) results at different RT60s

4.1 Generalization to room size

The DNN system, which was trained in the room of dimension 6 by 4 by 3 m (length by width by height), was tested in a very different room of dimension 10 by 7 by 3 m, with the positions of loudspeaker and microphones unchanged. Figure 10 shows the generalization results of the room size at different RT60s. Clearly, RTA-DNNSpatial yielded higher fwSegSNR scores than the unprocessed reverberant speech and WPE at each RT60s. The results illustrates that although our proposed DNN model was only trained in a single room, it generalized well to an unseen room size, demonstrating the robustness of our proposed environment-aware approach to new room sizes.

4.2 Generalization to array geometry

Next, we investigated the geometry dependence of the proposed RTA-DNNSpatial. In the training phase, a uniform linear array with an increment of 10 cm was utilized. In the dereverberation stage, a new array pattern was constructed by increasing the increment to 15 cm. As shown in Fig. 11, the RTA-DNNSpatial model still could achieve the highest scores at each RT60s, illustrating that the proposed approach had a powerful generalization capability to the geometrically mismatched case.

4.3 Generalization to loudspeaker position

In real scenarios, the speaker could be anywhere. Therefore, it is of importance to evaluate the generalization capabilities of loudspeaker position. In the training stage, the positions of the loudspeaker was at (2, 3, 1.5) m. In the dereverberation stage, we purposely changed the positions of the loudspeaker to be at (1, 1.5, 2.5) m. This time the room size and microphone array position was kept unchanged. Figure 12 illustrates that compared with WPE, the proposed RTA-DNNSpatial boosted fwSegSNR scores at RT60 = 0.3 s, 1.5 and 2.0 s. It demonstrates that our proposed RTA-DNNSpatial had good generalization capabilities to loudspeaker position.

4.4 Robustness to RT60 estimation error

The above RTA-DNNSpatial models assumed RT60 known in the dereverberation stage. However, it was unavailable in practice. Now, we present experimental results to assess the performance of RTA-DNNSpatial in practical solutions by using an accurate RT60 estimator proposed in [47], marked as "RTA-DNNSpatial-nonoracle". The RT60 is estimated to choose the optimal RT60-dependent temporal and spatial contexts from Table 2, which will be utilized in the dereverberation stage. The results of the two standard DNN-based multi-channel dereverberation configurations were also given. As shown in Fig. 13, compared with unprocessed reverberant speech, WPE, DSB-DNN, and DNNs-DSB, the nonoracle case substantially boosted fwSegSNR scores at all RT60s. Moreover, it is good to know that the two

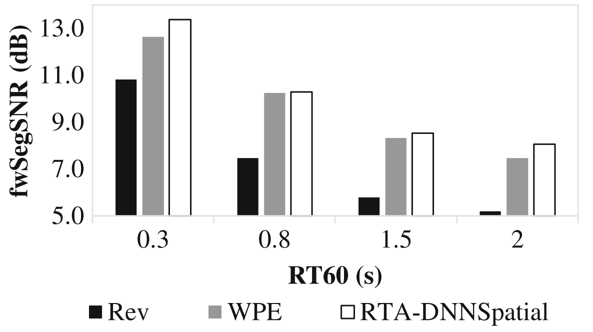

Fig. 10 Average fwSegSNR results at different RT60s, tested in a new room

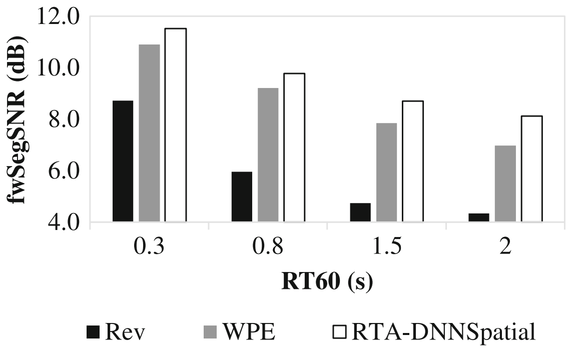

Fig. 11 Average fwSegSNR results at different RT60s, tested in a new array geometry

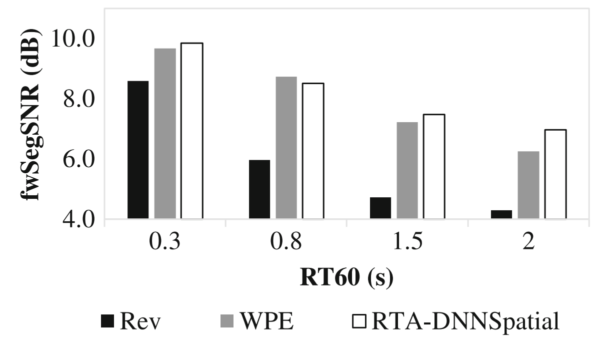

Fig. 12 Average fwSegSNR results at different RT60s, tested at a new loudspeaker position

environment-aware DNNSpatial frameworks were comparable. That is because in our designed RT60-aware DNNSpatial, the estimated utterance-level RT60 was rounded up to the nearest value of the multiples of 0.1 s. That means it allows a ± 0.05 s RT60 estimation error, increasing the robustness of our proposed algorithm to potential estimation errors.

In addition, it is difficult to evaluate the proposed approach on real RIRs. For example, in REVERB Challenge [48], it assumes the scenario of capturing utterances with a 8-ch circular array in reverberant meeting rooms. But the selection of RT60-dependent temporal and spatial contexts are analyzed and determined on a uniform linear array. But we can test our proposed DNNSpatial and RTA-DNNSpatial in its single-channel model on the REVERB Challenge 1-channel real data. As shown in [49], the proposed DNN systems outperform all other methods in all situations - the best performing method listed in the REVERB Challenge is shown for ease of comparison. In addition, although ACE Challenge [50] provides a 8-channel uniform linear array real recorded RIRs, it is tough for our RTA-DNNSpatial to handle all the extreme mismatched conditions together (unseen room size, speaker position, microphone position, array element increment and RT60, etc).

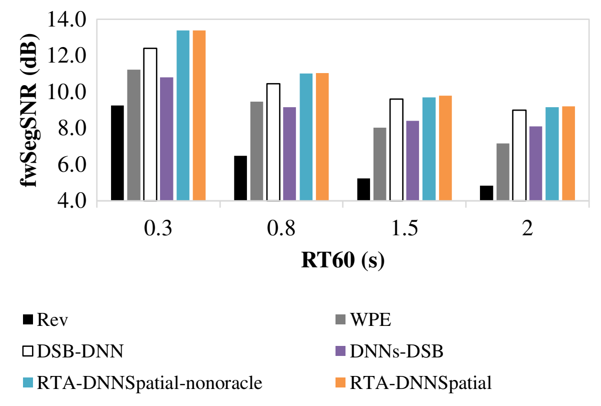

Fig. 13 Average fwSegSNR results at different RT60s, tested with an RT60 estimator

5 Conclusions

In this paper, we first propose a speech dereverberation framework, namely DNNSpatial, by selectively combining input LPS features of reverberant speech from multiple microphones in an array and map them into the expected output LPS features of anechoic reference speech based on DNNs. We compare the proposed single-DNN approach to two standard DNN-based multichannel dereverberation configurations, namely DSB-DNN and DNNs-DSB. Experimental results demonstrate that the proposed single DNNSpatial model without DSB performs better than the other two DNN models with DSB in both 2-microphone and 6-microphone settings according to all the three objective measures tested.

Next, we propose a reverberation-time-aware DNNSpatial framework, namely RTA-DNNSpatial, by adopting RT60-dependent temporal and spatial contexts, to make the system robust enough to handle a wide range of RT60s. Experimental results indicate that it is superior to the state-of-the-art signal processing multi-channel dereverberation algorithm WPE and DNNSpatial models, including slightly and severely reverberant environments. It also generalizes well to mismatched room size, array geometry, and loudspeaker position, and is robust to RT60 estimation error, which will yield significant benefits in many practical applications.

In future studies, we would like to explore the availability of direction of arrival (DOA) information to further improve the system performances.

Endnote

[1] For the input concatenated feature vectors of 3-3-1-1-3-3, 5-1-1-1-1-5, and 7-0-0-0-0-7, the same frequency bin may correspond to different microphones. Based on our preliminary experiments, this inconsistency between frequency and spatial domains will result in degraded dereverberation performances. Therefore during the RTA-DNNSpatial training, the normalized input vectors of different feature combinations were added zeros to remove the inconsistency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61571344). The authors would like to thank Dr. Fengpei Ge at Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences for her valuable suggestions.

Authors' contributions

BW, KHL, SMS, MLY, and CHL conceived and designed the study. BW, KHL, ZH, and SMS performed the experiments. BW and MLY wrote the paper. MLY, TW, and CHL reviewed and edited the manuscript. All authors read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. [2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA. [3]Department of Telecommunications, University of Enna Kore, Enna, Italy.

Received: 3 May 2017 Accepted: 16 November 2017
Published online: 06 December 2017

References

1. M Wu, DL Wang, A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **14**(3), 774–784 (2006)
2. S Mosayyebpour, M Esmaeili, TA Gulliver, Single-microphone early and late reverberation suppression in noisy speech. IEEE Trans. Audio Speech Lang. Process. **21**(2), 322–335 (2013)
3. K Kinoshita, M Delcroix, T Nakatani, M Miyoshi, Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans. Audio Speech Lang. Process. **17**(4), 534–545 (2009)
4. S Gannot, D Burshtein, E Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
5. EAP Habets, J Benesty, A two-stage beamforming approach for noise reduction and dereverberation. IEEE Trans. Audio, Speech Lang. Process. **21**(5), 945–958 (2013)
6. ST Neely, JB Allen, Invertibility of a room impulse response. J. Acoust. Soc. Amer. **66**(1), 165–169 (1979)
7. PA Naylor, ND Gaubitch, eds., *Speech Dereverberation*. (Springer, London, 2010)
8. GE Hinton, S Osindero, YW Teh, A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
9. GE Hinton, RR Salakhutdinov, Reducing the dimensionality of data with neural networks. Science. **313**(5786), 504–507 (2006)
10. K Han, Y Wang, DL Wang, *et al*, Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(6), 982–992 (2015)
11. K Han, Y Wang, DL Wang, in *ICASSP*. Learning spectral mapping for speech dereverberation, (2014), pp. 4628–4632
12. B Wu, K Li, ML Yang, C-H Lee, A reverberation-time-aware approach to speech dereverberation based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 102–111 (2017)
13. J Benesty, J Chen, Y Huang, *Microphone array signal processing, vol. 1*. (Springer, Berlin, 2008)
14. T Lotter, P Vary, Dual-channel speech enhancement by superdirective beamforming. EURASIP J. Adv. Signal Process. **2006**(1), 063297 (2006)
15. T Nakatani, T Yoshioka, K Kinoshita, M Miyoshi, B-H Juang, in *ICASSP*. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation, (2008), pp. 85–88
16. T Nakatani, T Yoshioka, K Kinoshita, M Miyoshi, B-H Juang, Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1717–1731 (2010)
17. A Jukić, T van Waterschoot, T Gerkmann, S Doclo, Multi-channel linear prediction-based speech dereverberation with sparse priors. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(9), 1509–1520 (2015)
18. T Yoshioka, T Nakatani, M Miyoshi, HG Okuno, Blind separation and dereverberation of speech mixtures by joint optimization. IEEE Trans. Audio Speech Lang. Process. **19**(1), 69–84 (2011)
19. J Nikunen, *Distant speech separation using predicted time-frequency masks from spatial features*. (Elsevier Science Publishers B. V., 2015), pp. 97–106
20. S Sivasankaran, AA Nugraha, E Vincent, JA Morales-Cordovilla, in *IEEE Workshop on Automatic Speech Recognition and Understanding*. Robust asr using neural network based speech enhancement and feature simulation, (2015), pp. 482–489
21. M Brandstein, D Ward, *Microphone arrays: signal processing techniques and applications*. (Springer, Berlin, 2013)
22. BW Gillespie, HS Malvar, DA Florêncio, in *ICASSP*. Speech dereverberation via maximum-kurtosis subband adaptive filtering, (2001), pp. 3701–3704
23. SM Griebel, A microphone array system for speech source localization, denoising, and dereverberation. PhD thesis. Citeseer (2002)
24. J Du, Y-H Tu, L Sun, F Ma, H-K Wang, J Pan, C Liu, C-H Lee, The ustc-iflytek system for chime-4 challenge. Technical report, Technical report of CHiME-4 (2016)
25. DW Griffin, JS Lim, Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **32**(2), 236–243 (1984)
26. JM Tribolet, P Noll, BJ McDermott, RE Crochiere, in *ICASSP*. A Study of complexity and quality of speech waveform coders, (1978), pp. 586–590
27. J Du, Q Huo, in *Interspeech*. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions, (2008), pp. 569–572
28. K Li, C-H Lee, in *ICASSP*. A deep neural network approach to speech bandwidth expansion, (2015), pp. 4395–4399
29. BD Radlovic, RC Williamson, RA Kennedy, Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Speech Audio Process. **8**(3), 311–319 (2000)
30. TF Quatieri, *Discrete-time speech signal processing: principles and practice*. (Pearson Education, India, 2002)
31. CW Gardiner, *Handbook of stochastic methods*. (Springer, Berlin, 1985)
32. SM Kay, Statistical signal processing. Estimation Theory. **1** (1993)
33. G Zelniker, FJ Taylor, *Advanced digital signal processing: theory and applications*. (CRC Press, Boca Raton, 1993)
34. WC Sabine, *Collected papers on acoustics*. (Harvard University Press, London, 1922)
35. TD Abhayapala, RA Kennedy, RC Williamson, Spatial aliasing for near-field sensor arrays. Electron. Lett. **35**(10), 764–765 (1999)
36. B Rafaely, B Weiss, E Bachmat, Spatial aliasing in spherical microphone arrays. IEEE Trans. Signal Process. **55**(3), 1003–1010 (2007)
37. K Li, Z Huang, Y Xu, C-H Lee, in *interspeech*. DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech, (2015)
38. J Dmochowski, J Benesty, S Affès, On spatial aliasing in microphone arrays. IEEE Trans. Signal Process. **57**(4), 1383–1395 (2009)
39. EA Lehmann, AM Johansson, Prediction of energy decay in room impulse responses simulated with an image-source model. J. Acoust. Soc. Amer. **124**(1), 269–277 (2008)
40. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, *et al*, in *ASRU*. The Kaldi Speech Recognition Toolkit, (2011), pp. 1–4
41. Y Bengio, Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
42. Y Xu, J Du, L-R Dai, C-H Lee, A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2015)
43. Y Xu, J Du, L-R Dai, C-H Lee, An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. **21**(1), 65–68 (2014)
44. CH Taal, RC Hendriks, R Heusdens, J Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
45. Rec. P. 862 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. Int. Telecommun. Union-Telecommun. Stand. Sector (2001)
46. J Ma, Y Hu, PC Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Amer. **125**(5), 3387–3405 (2009)
47. A Keshavarz, S Mosayyebpour, M Biguesh, TA Gulliver, M Esmaeili, Speech-model based accurate blind reverberation time estimation using an LPC filter. IEEE Trans. Audio Speech Lang. Process. **20**(6), 1884–1893 (2012)
48. K Kinoshita, M Delcroix, S Gannot, EA Habets, R Haeb-Umbach, W Kellermann, V Leutnant, R Maas, T Nakatani, B Raj, *et al*, A summary of the REVERB challenge: state-of-the-art and remaining challenges in

reverberant speech processing research. EURASIP J. Adv. Signal Process. **2016**(1), 1–19 (2016)
49. B Wu, K Li, Z Huang, SM Siniscalchi, M Yang, C-H Lee, in *HSCMA*. A unified deep modeling approach to simultaneous speech dereverberation and recognition for the reverb challenge, (2017)
50. J Eaton, ND Gaubitch, AH Moore, PA Naylor, in *WASPAA*. The ACE Challenge—Corpus Description and Performance Evaluation, (2015)
51. JS Garofolo, Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. Technical report. NIST (1988)