Ueda *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:92
DOI 10.1186/s13634-015-0278-y

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Environment-dependent denoising autoencoder for distant-talking speech recognition

Yuma Ueda[1], Longbiao Wang[2*], Atsuhiko Kai[1] and Bo Ren[2]

Abstract

In this paper, we propose an environment-dependent denoising autoencoder (DAE) and automatic environment identification based on a deep neural network (DNN) with blind reverberation estimation for robust distant-talking speech recognition. Recently, DAEs have been shown to be effective in many noise reduction and reverberation suppression applications because higher-level representations and increased flexibility of the feature mapping function can be learned. However, a DAE is not adequate in mismatched training and test environments. In a conventional DAE, parameters are trained using pairs of reverberant speech and clean speech under various acoustic conditions (that is, an environment-independent DAE). To address the above problem, we propose two environment-dependent DAEs to reduce the influence of mismatches between training and test environments. In the first approach, we train various DAEs using speech from different acoustic environments, and the DAE for the condition that best matches the test condition is automatically selected (that is, a two-step environment-dependent DAE). To improve environment identification performance, we propose a DNN that uses both reverberant speech and estimated reverberation. In the second approach, we add estimated reverberation features to the input of the DAE (that is, a one-step environment-dependent DAE or a reverberation-aware DAE). The proposed method is evaluated using speech in simulated and real reverberant environments. Experimental results show that the environment-dependent DAE outperforms the environment-independent one in both simulated and real reverberant environments. For two-step environment-dependent DAE, the performance of environment identification based on the proposed DNN approach is also better than that of the conventional DNN approach, in which only reverberant speech is used and reverberation is not blindly estimated. And, the one-step environment-dependent DAE significantly outperforms the two-step environment-dependent DAE.

Keywords: Speech recognition, Dereverberation, Denoising autoencoder, Environment identification, Distant-talking speech

1 Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance because of mismatches between the training and test environments. There are two different approaches, namely, front- and back-end-based methods, for dealing with this problem [1]. Many front-end-based approaches [1–8] have been proposed to reduce the effect of reverberation in the observed speech signal. The back-end-based methods, on the other hand, attempt to modify the acoustic model and/or decoder to suit the respective reverberant environment [9, 10]. In this paper, we focus on front-end-based approaches for distant-talking speech recognition.

Many single- and multi-channel dereverberation methods have been proposed to suppress reverberation [2–4, 11–14]. Single-channel dereverberation approaches are much easier and cheaper to implement in real applications than multi-channel ones. In this paper, dereverberation is performed using a single-channel speech signal. Cepstral mean normalization (CMN) can be considered the most general single-channel approach

*Correspondence: wang@vos.nagaokaut.ac.jp
[2]Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka 940-2188, Japan
Full list of author information is available at the end of the article

[15–17]. Having been extensively examined, it has been shown to be a simple and effective way of reducing reverberation by normalizing cepstral features. However, the dereverberation of CMN is not completely effective in environments with late reverberation. Several studies have focused on mitigating this problem [3, 4, 12]. A reverberation compensation method for speaker recognition using spectral subtraction [18], where late reverberation is treated as additive noise, was proposed in [3]. A method based on multi-step linear prediction (MSLP) was proposed for both single and multiple microphones [4, 12]. The method first estimates late reverberations using long-term MSLP, and then suppresses these with subsequent spectral subtraction. Wolfel proposed a joint compensation of noise and reverberation by integrating an estimate of the reverberation energy derived by an auxiliary model based on MSLP, into a framework, which so far, tracks and removes nonstationary additive distortion by particle filters in a low-dimension logarithmic power frequency domain [19].

Neural network (NN)-based approaches have been proposed for feature transformation [20, 21]. Bottleneck features extracted by a multi-layer perceptron (MLP) can be used for nonlinear feature transformation [20]. However, deep networks of MLPs with many hidden layers have a high computational cost, the lower layers in a DNN architecture are hard to train because of vanishing gradients. Deep belief networks, which employ an unsupervised pre-training method using a restricted Boltzmann machine (RBM), have been proposed to train better initial values of deep networks [22]. Deep neural networks (DNNs) with pre-training have been shown to achieve better performance than the conventional MLP without pre-training [22]. There are many DNN, recurrent neural network (RNN) and long short-term memory (LSTM) [23–29] based speech enhancement and feature enhancement approaches that have been proposed for speech enhancement for human listening and robust speech recognition and that have shown good performance for the REVERB challenge task [40]. Recently, the denoising autoencoder (DAE), one type of DNN, has been shown to be effective in many noise reduction applications because higher-level representations and increased flexibility of the feature mapping function can be learned [30–33]. Ishii et al. applied a DAE to spectral-domain dereverberation resulting in improved word accuracy of large-vocabulary continuous speech recognition (LVCSR) [34]. Previously, we found that cepstral domain DAE-based dereverberation is efficient for distant-talking speech recognition [35]. As shown in [35], DAE worked well especially with strong reverberation. However, the results of DAE with small reverberation are not good compared to other methods. Typically, in the training of a DAE [26, 34, 35], data incorporating various environmental conditions are used. Although this training method is suitable for training models in various environments, the nonlinear transformation ability of DAE training using various conditions that do not match those of the test data is lower for certain acoustic conditions of the test set. Thus, the performance of a DAE cannot be sufficiently improved for an unknown test reverberant condition.

To improve robustness of speech recognition, the idea of using side information from the environment as additional features, such as speaker-specific side information (e.g., i-vectors) and room information etc. has been proposed previously [36–38]. In this paper, two environment-dependent DAEs are proposed to reduce the influence of mismatches between training and test environments, that is, DAEs are trained and used corresponding to the different environments. In the first approach, various DAEs are trained using speech from different acoustic environments, and the DAE with the condition that best matches the test condition is automatically selected using a DNN (that is, a two-step environment-dependent DAE). The performance of our proposed two-step environment-dependent DAE is dependent on the precision of the automatic environment identification. In this paper, to achieve higher environment identification performance, a DNN using both reverberant speech and reverberation estimated by MSLP is also proposed. In the second approach, reverberation features estimated by MSLP are directly used as an input of the DAE (that is, a one-step environment-dependent DAE or a reverberation-aware DAE). By simultaneously estimating and suppressing the environment-dependent reverberation with a one-step environment-dependent DAE (that is, reverberation-aware DAE), the mismatch between the training data and test data will be reduced. Therefore, better estimation of the clean speech can be expected. In previous work, conventional DAE was trained using speech data under various environments, and the test reverberant speech is transformed using a conventional environment-independent DAE that cannot deal with environmental variation when there is limited training data. In the proposed approach, the test reverberant speech is transformed using an environment-dependent DAE that can estimate the environment-dependent reverberation. Thus, the environment-dependent DAE is more robust to environmental changes than a conventional DAE. The proposed methods are evaluated in both simulated and real reverberant environments.

The remainder of this paper is organized as follows: Section 2 describes the DAE for cepstral-domain dereverberation. The methods for one-step and two-step environment-dependent DAEs are described in Section 3, while the experimental results and a discussion thereof are presented in Section 4. Finally, Section 5 summarizes the paper.

2 Denoising autoencoder for cepstral-domain dereverberation

2.1 Topology of DAE

An autoencoder is a type of artificial neural network (NN), whose output is reconstruction of input, and is often used for dimensionality reduction. DAEs share the same structure as autoencoders, but input data are a noisy version of the teacher signal. In this paper, we use the clean reference speech signal as teacher signal. Autoencoders use feature mapping to convert noisy input data into clean output and have been used for noise removal in the field of image processing [30]. Ishii et al. applied a DAE for spectral-domain dereverberation [34]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. In this paper, we apply a denoising autoencoder for cepstral-domain dereverberation because there are many LVCSR systems that adopt a cepstral-domain feature as the direct input.

Given a pair of speech samples, clean speech and corresponding reverberant speech, DAE learns the non-linear conversion function that converts reverberant speech features into clean speech. In general, reverberation is dependent on both current and several previous observation frames. In addition to the vector of the current frame, vectors of past frames are concatenated to form input.

For cepstral feature X_i of observed reverberant speech of the $i-th$ frame, cepstral features of $N-1$ frames before the current frame are concatenated with the current frame to form a cepstral vector of N frames. Output O_i of the non-linear transformer based on the DAE is given by

$$O_i = f_L(\ldots f_l(\ldots f_2(f_1(X_i, X_{i-1}, \ldots, X_{i-N+1}))) \quad (1)$$

where f_l is the non-linear transformation function in layer l and N is the number of frames to be used as the input features.

Topology of the cepstral-domain DAE for dereverberation is shown in Fig. 1. In this paper, the number of hidden layers is set to three. In Fig. 1, $W_i (i = 1, 2)$ shows the weighting of the different layers, and W_i^T shows the transposition of W_i[1]. That is to say, W_1 and W_2 are the encoder matrices and W_1^T and W_2^T are the decoder matrices, respectively.

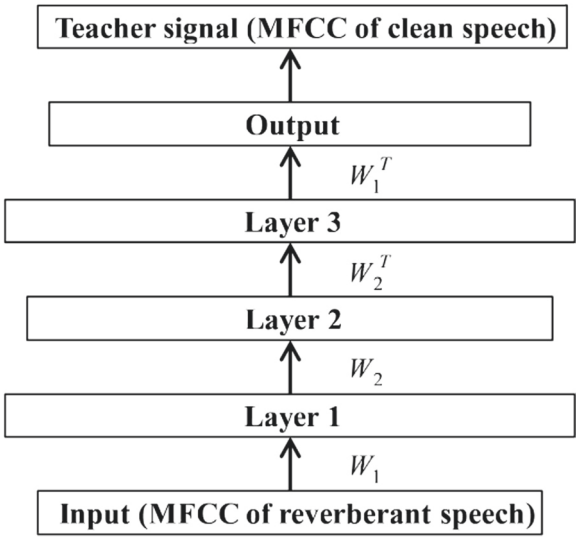

Fig. 1 Topology of a stacked denoising autoencoder for cepstral-domain dereverberation

2.2 Training of DAE

2.2.1 Restricted Boltzmann machine

To train a deep neural network, deep belief networks (DBNs) [22] are used for pre-training because they can obtain accurate initial values of the deep-layer neural networks.

RBM is a bipartite graph shown in Fig. 2. It has a visible and hidden layer in which visible units that represent observations are connected to hidden units that learn to represent features using weighted connection. An RBM is restricted such that there are no visible-visible or hidden-hidden connections. Different types of RBMs are used in the case of binary or real-valued input. Bernoulli-Bernoulli RBMs are used to convert binary stochastic variables to binary stochastic variables. Gaussian-Bernoulli RBMs are used to convert real-valued stochastic variables to binary stochastic variables. Details of RBM are obtained in [22].

To obtain a pre-trained RBM, we trained all the hidden layers by using the Bernoulli-Bernoulli RBM. DBNs are hierarchically configured by connecting these pre-trained RBMs. Here, W_1 and W_2 are learned automatically, and W_1^T and W_2^T are generated from W_1 and W_2 in Fig. 1.

2.2.2 Backpropagation algorithm

After pre-training, a backpropagation algorithm was applied to adjust the parameters. Backpropagation modifies the weights of the network to reduce the error of the teacher signal and the output value when a pair of signals (input signal and the ideal teacher signal, the cepstral

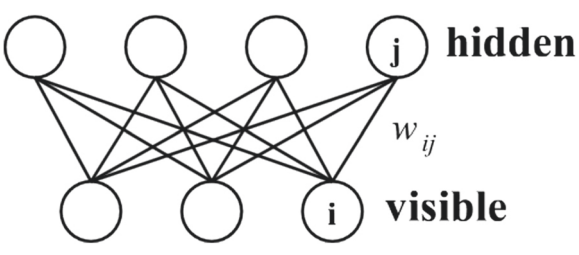

Fig. 2 Graphical representation of the RBM

feature of clean speech) are given. We scaled the cepstral feature value of the input data and teacher signal to between 0 and 1 using a sigmoid function. The minimization in this paper is carried out by minimizing the cross entropy using conjugate gradients [22].

3 Environment-dependent denoising autoencoder

A conventional DAE trained using data under various acoustic conditions is effective for noise reduction and dereverberation. However, it is impossible to deal with mismatched conditions of the training and test data or unseen data with limited training data. We deal with this problem by two approaches. In the first approach, multiple DAEs for each environment are trained and selectively used. However, the reverberation environment is unknown in the test stage. Here, we use a DNN for environment identification because it is more effective than other classifiers such as the Gaussian mixture model (GMM) and support vector machine for audio classification [39]. In the second approach, reverberation features estimated by MSLP are directly used as an input of the DAE (that is, a one-step environment-dependent DAE or a reverberation-aware DAE). In the following, we describe the proposed environment-dependent DAE.

3.1 Environment-independent and environment-dependent DAEs

For a conventional DAE (environment-independent DAE), parameters are trained using pairs of reverberant speech and clean speech under various acoustic conditions. The environment-independent DAE is not robust for mismatches between training and test conditions. To address this problem, we propose two environment-dependent DAEs to mitigate the influence of mismatch in the training and test environments. In the first approach of environment-dependent DAE (that is, two-step environment-dependent DAE), various DAEs are trained using speech from different acoustic environments, and the DAE with the condition that best matches the test condition is automatically selected as shown in Fig. 3. In the second approach (that is, one-step environment-dependent DAE), we add the environmental information (e.g., estimated reverberation features) to input of DAE as shown in Fig. 4. These approaches are expected for reducing the influence of mismatch between training and test environment, and improving LVCSR performance.

3.2 Two-step environment-dependent DAE

3.2.1 Environment identification

First, we divide the training data according to the environment. This is performed manually because the training data environment is known. Next, each DAE, with its respective data and the environment identification model, is trained. The training approach for the environment identification model is the same as that for the DAE with the exception of the input data and giving the reverberation environment as the correct label. In the DAE, the second half of weights of DAE is generated from

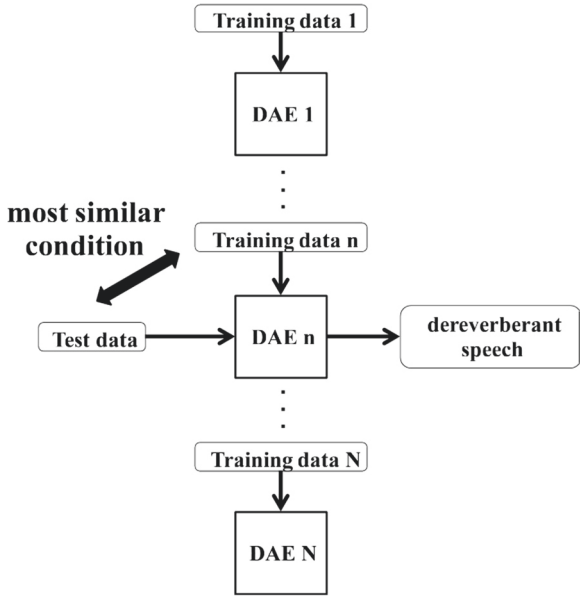

Fig. 3 Two-step environment-dependent DAE

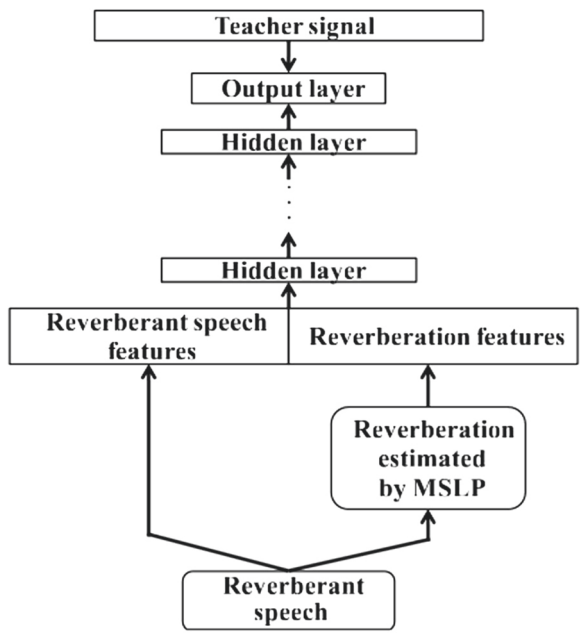

Fig. 4 DNN structure using reverberation estimated by MSLP for environment identification and one-step environment-dependent DAE. The teacher signal is the label of environment for environment identification and MFCC of clean speech for the one-step environment-dependent DAE, respectively

transposition of first half of it. However, in the environment identification DNN, the second half of weights of DNN is not a transposition of the first half of it and is trained by RBM.

Although reverberant speech features are used as input in the training of a DAE, this is not sufficient in the training of the identification model. As shown in Fig. 4, reverberation is estimated from reverberant speech, with reverberation features also used as input. By doing this, we can expect the performance of environment identification to be improved. In this paper, estimation of reverberation is based on MSLP [12]. The original MSLP algorithm estimates the reverberation from reverberant speech and suppresses reverberation in the spectral domain. In this study, MSLP is used for reverberation estimation only, and reverberation suppression is not performed.

$$y(n) = \sum_{p=0}^{N-1} w(p)y(n-p-D) + e(n) \quad (2)$$

where $y(n)$ is the observed signal, D is the step size, N is the filter length, $w(p)$ are prediction coefficients, and $e(n)$ is the prediction error. Prediction filter $w(n)$ is estimated by minimizing the mean square energy of the prediction error. Late reverberation is estimated using reverberant speech and the estimated prediction coefficients.

3.2.2 Combination of environment-dependent DAE and automatic environment identification

Figure 5 shows the flowchart for environment-dependent dereverberation using the environment identification technique and multiple DAEs. First, we identify the reverberation environment of the input speech by applying an identification model to the speech. Here, the reverberation features estimated by MSLP are used as input for the model as well as for the training thereof. Next, the DAE corresponding to the identified environment is selected, and dereverberation is applied by it. Since dereverberation is applied by a DAE suited to the environment of the speech, we expect an improvement in performance.

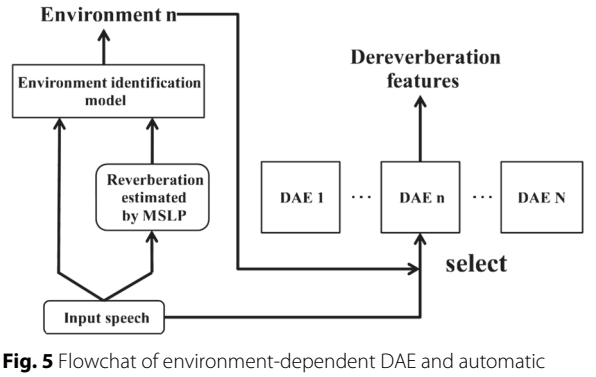

Fig. 5 Flowchat of environment-dependent DAE and automatic environment identification

3.3 One-step environment-dependent DAE

The second approach is almost the same as that for the environment identification in Section 3.2.1. In this approach, estimated reverberation and reverberant speech are directly used as inputs of the reverberation-aware (DAE) as shown in Fig. 4. The one-step environment-dependent DAE can estimate and suppress the environment-dependent reverberation automatically. So, it is expected to reduce the influence of mismatches between training and test environments. On the other hand, the conventional DAE does not use environment-dependent reverberation, so its performance will not be robust for mismatches between training and test conditions.

4 Experiments

4.1 Experimental setup

4.1.1 Training dataset

We used the training dataset provided by the "REVERB challenge" (reverberant voice enhancement and recognition benchmark) [40]. This dataset consists of the clean WSJCAM0 [41] training set and a multi-condition (MC) training set. Reverberant speech is generated from the clean WSJCAM0 training data by convolving the clean utterances with measured room impulse responses and adding recorded background noise. The reverberation times of the measured impulse responses range from approximately 0.1 to 0.8 s. The training data of the "REVERB Challenge" were used to train the environment identification DNN. The environment labels depended on room type and the distance between the microphone and speaker. The training data include three types of rooms and two types of distances between the microphone and speaker, so in total, six types of environments with distinct rooms and distances were used. This training dataset were also used to train the DAEs and acoustic models.

It should be noted that the recording rooms used for the multi-condition training data and test data were different.

4.1.2 Development and evaluation test sets

It is important to note that the proposed dataset consists of real recordings (RealData) and simulated data (SimData), part of which has similar characteristics to the RealData in terms of reverberation time and microphone-speaker distance. This setup allowed us to perform evaluations in terms of both practicality and robustness of various reverberant conditions. Specifically, the development (Dev.) and final evaluation (Eval.) test sets each contained the following SimData and RealData; SimData was generated from the WSJCAM0 corpus [41], and RealData

from the MC-WSJ-AV corpus [42]. This development dataset was used to determine the optimal parameters for dereverberation and speech recognition. Details of the training and test datasets are given in Tables 1 and 2.

4.1.3 Experimental conditions for LVCSR and dereverberation

In this study, Mel-frequency cepstral coefficients (MFCCs) were used as features for LVCSR. The dimension of the MFCCs was 39 including 12 MFCCs plus power and their Delta and Delta-Delta coefficients. MFCC features were normalized using the mean of the entire multi-condition training set. DAE training was carried out using mini-batch conjugate gradients with a mini-batch size of 128 samples. In this paper, the number of hidden layers is set to three. The number of units in each layer is 512, and each unit uses a sigmoid function as an activation function. Because reverberation affects multiple frames, we supply multiple frames at the same time as the input and teacher signals of the DAE. The dimensions of the input data and output are 39 (the dimensions of MFCC per frame) * 9 (the number of segments to supply at the same time) = 351. These parameters were empirically determined. Please refer to [35] for a more detailed description. Fifty epochs with a learning rate of 0.002 were used for all layers during pre-training, and 100 epochs with a learning rate of 0.1 were used for all layers during fine-tuning. Training of the environment identification model architecture was almost the same as for the DAEs. The number of hidden layers is 5, the number of units in each layer is 1024, and 20 epochs were used for all layers during fine-tuning. The number of classes (the number of units at the final softmax layer) is 6, which is determined by the number of environments in the training data (i.e., Room 1, Room 2, and Room 3, each with near and far conditions).

The MSLP algorithm generates an inverse filter through the prediction coefficients to estimate the inverse system [12]. We estimated the late reverberation components using the inverse filter and applied dereverberation by power spectral subtraction. For MSLP-based dereverberation, the step size and the order of linear prediction were set to 500 and 750, respectively. We used MSLP to estimate the late reverberation of both the training and test data with the same parameters as for MSLP-based dereverberation.

We used a subspace GMM with maximum mutual information-based discriminative training (MMI-SGMM) [43] and a cross-entropy training DNN for the acoustic model. The KALDI toolkit [44] was used as a decoder for LVCSR. In this study, the numbers of hidden layers and units were set to 3 and 1024 for the DNN acoustic model. The final results were obtained from a confusion network combination of MMI-SGMM with 7000 states and DNN-HMM with 2500 states. Details can be found in [44]. Standard Wall Street Journal 5000-word trigram language model was used for decoding. We used word error rate (WER) to evaluate the speech recognition performance for each method.

4.2 Experimental results

4.2.1 Results of environment-dependent DAE

In this section, we compare the following four dereverberation methods:

- CMVN: Cepstral mean and variance normalization
- MSLP: Multi-step linear prediction
- DAE: Environment-independent denoising autoencoder (conventional DAE)
- Two-step environment-dependent DAE: environment-dependent DAE selected by an estimated environment
- One-step environment-dependent DAE: reverberation-aware DAE using reverberant speech features and late reverberation features estimated by MSLP

Using three types of acoustic models:

- SGMM: MMI-SGMM
- DNN: cross-entropy training DNN-HMM
- SGMM+DNN: system combination of MMI-SGMM and DNN-HMM.

Tables 3 and 4 show the speech recognition results for each method on the Dev. and Eval. datasets, respectively. For the Dev. dataset, DAE-based cepstral-domain dereverberation shows a remarkable improvement when compared with CMVN- and MSLP-based dereverberation.

Table 1 Amount of data for the Dev. and Eval. sets of SimData and RealData and for the training dataset[1]

	SimData		RealData		Training Data
	Dev.	Eval.	Dev.	Eval.	
# of sentences	1484	2176	179	372	7861
	(∼ 3 h)	(∼ 4.8 h)	(∼ 0.3 h)	(∼ 0.6 h)	(∼ 17.5 h)
# of speakers	10	28	5	10	92

[1]The clean and multi-condition training datasets are the same size

Table 2 Details of data set of SimData and RealData

Speech	Corpus	Reverberant time			Signal-to-noise ratio	Distance between the microphones	
		Room 1	Room 2	Room 3		Near	Far
SimData	WSJCAM0	0.25 s	0.5 s	0.7 s	20 dB	50 cm	200 cm
RealData	MC-WSJ-AV	0.7 s	-	-	-	100 cm	250 cm

The DAE works especially well with strong reverberation, i.e., far-field microphone in "Room 2" and "Room 3" of SimData. The cepstral domain environment-independent DAE outperformed CMN and MSLP under almost all conditions. Although the performance of the two-step environment-dependent DAE is better than the conventional environment-independent DAE in some environments, it is worse in some environments. The reason may be that the performance of environment identification and environment special DAE depended on the training data, environment label, and also the environment of test data. The proposed one-step environment-dependent DAE (that is, reverberation-aware DAE) outperformed the conventional environment-independent DAE and two-step environment-dependent DAE on both SimData and RealData in the Dev. and Eval. datasets using SGMM, DNN, and SGMM+DNN. The improvement of the one-step environment-dependent DAE under large reverberation conditions is greater than that under small reverberation conditions. The reason is that the conventional DAE is not effective enough when there are large environmental mismatches between the training and test conditions, and the one-step environment-dependent DAE can reduce the influence of mismatch by estimating the late reverberation and adding it to the input DAE. For SimData in the Dev. dataset, using the one-step environment-dependent DAE with reverberation features estimated by MSLP, the average WER was reduced from 6.36 % with the conventional DAE to 5.77 % using SGMM+DNN, i.e., a relative error reduction rate of 9.28 %. For RealData in the Dev. dataset, the average WER was reduced from 27.46 % with the conventional DAE to 26.66 % using SGMM+DNN,

Table 3 Word error rate of different dereverberation methods for Dev. dataset (%)

Dereverberation method	Acoustic model	SimData							RealData		
		Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
CMVN	SGMM	3.83	5.51	6.33	12.60	7.72	14.71	8.45	42.92	44.77	43.85
	DNN	5.43	6.88	6.73	13.73	9.05	16.37	9.70	41.30	43.20	42.25
	SGMM +DNN	3.86	5.21	5.62	11.63	7.20	13.30	7.80	41.42	42.58	42.00
MSLP	SGMM	5.17	5.39	6.77	10.95	7.93	15.90	8.69	36.22	38.08	37.15
	DNN	5.74	6.30	6.91	12.02	8.19	16.56	9.29	37.43	36.97	37.20
	SGMM +DNN	4.01	5.04	5.23	10.25	5.93	12.17	7.11	35.50	36.91	36.21
DAE	SGMM	4.30	5.41	5.45	9.71	5.24	10.63	6.79	26.51	30.08	28.30
	DNN	5.41	6.69	6.24	10.67	6.50	11.55	7.84	28.70	29.19	28.95
	SGMM +DNN	4.20	4.87	5.23	9.24	4.87	9.74	6.36	26.08	28.84	27.46
Two-step environment-dependent DAE	SGMM	3.79	5.90	4.91	9.22	6.08	8.85	6.46	28.45	31.24	29.84
	DNN	5.01	6.91	5.45	10.38	6.87	11.35	7.66	28.95	31.99	30.47
	SGMM +DNN	3.76	4.89	4.63	8.48	6.26	9.20	6.20	27.45	29.12	28.28
One-step environment-dependent DAE	SGMM	4.08	4.67	4.61	8.80	4.70	8.75	5.94	28.95	27.61	28.28
	DNN	4.97	6.47	5.72	10.13	5.98	9.40	7.11	28.95	27.20	28.08
	SGMM +DNN	3.88	4.79	5.10	7.94	4.60	8.28	5.77	25.83	27.48	26.66

Table 4 Word error rate of different dereverberation methods for Eval. dataset (%)

Dereverberation method	Acoustic model	SimData							RealData		
		Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
CMVN	SGMM	5.47	5.88	6.59	12.68	8.29	16.77	9.28	44.84	44.53	44.69
	DNN	6.05	6.71	7.89	13.29	9.13	17.74	10.14	43.82	43.55	43.69
	SGMM +DNN	4.90	5.39	6.33	11.77	7.68	15.57	8.61	43.40	42.98	43.19
MSLP	SGMM	7.05	7.95	8.42	14.34	10.46	19.70	11.32	42.45	43.65	43.05
	DNN	7.95	9.10	9.48	15.98	11.72	21.39	12.60	43.50	44.80	44.15
	SGMM +DNN	4.71	5.18	5.95	9.95	7.32	14.45	7.93	35.52	36.09	35.81
DAE	SGMM	5.07	5.51	6.11	9.53	7.64	12.11	7.66	32.26	32.58	32.42
	DNN	5.86	6.45	7.06	10.85	7.92	12.78	8.49	31.62	32.88	32.25
	SGMM +DNN	4.79	5.40	5.64	9.00	7.06	10.85	7.12	30.02	31.09	30.56
Two-step environment-dependent DAE	SGMM	4.61	6.73	5.47	10.01	7.83	12.32	7.83	30.57	33.05	29.84
	DNN	5.57	8.37	6.16	10.90	7.85	13.14	8.66	31.49	33.59	30.47
	SGMM +DNN	4.25	6.32	5.19	8.95	7.08	11.50	7.22	29.16	31.03	30.10
One-step environment-dependent DAE	SGMM	4.93	5.30	5.82	8.47	7.25	10.47	7.04	28.65	28.66	28.66
	DNN	5.29	6.05	6.43	8.81	7.20	10.97	7.46	27.95	28.26	28.11
	SGMM +DNN	4.54	5.05	5.37	7.62	6.50	9.40	6.41	26.38	27.28	26.83

i.e., a relative error reduction rate of 2.91 %. For the Eval. dataset, a similar trend was observed. The proposed one-step environment-dependent DAE (that is, reverberation-aware DAE) with reverberation features estimated by MSLP outperformed all the other methods. For SGMM+DNN acoustic model, compared with the conventional DAE, relative error reduction rates of 9.97 % for SimData and 12.21% for RealData were achieved. The results show that the proposed one-step environment-dependent DAE is also robust to variations of speaker and speech context.

4.2.2 Comparison of different environment identification models

In this section, we investigate the effect of the environment identification method for the two-step environment-dependent DAE. We compare the performance of the environment identification model, using and not using reverberation estimated by MSLP, for training.

Table 5 shows the speech recognition results on the Dev. dataset for these two methods. The results are based on a system combination of MMI-SGMM and DNN. Bigram training was used for the language model. These results indicate that the performance of environment identificat7ion is improved by using estimated reverberation in training the DNN. By blindly using reverberation estimated by MSLP, the DNN can identify an unknown test environment precisely. Without using the estimated reverberation, the two-step environment-dependent DAE performs worse than the conventional environment-independent one owing to poor environment identification performance. That is to say, the two-step environment-dependent DAE is sensitive to the environment identification performance.

Table 5 Effect of environment identification method for two-step environment-dependent DAE. WER of Dev. dataset (%)

Use reverberation estimated by MSLP	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
No	5.16	7.23	7.32	16.79	7.05	20.35	10.65	42.30	41.90	42.10
Yes	5.01	8.24	6.78	11.78	7.54	13.72	8.85	32.13	33.42	32.78

Table 6 The average WERs (%) of Eval. dataset for the single channel dataset comparing with other teams. Multi-condition training dataset and the trigram language model were used for all teams

Team	Acoustic model	Feature	Dereverberation method	SimData	RealData	Ave.
REVERB-challenge baseline	GMM	MFCC	CMVN	25.27	47.48	36.38
J. Alam et al. [45]	DNN	MFCC	maximum likelihood inverse filtering-based dereverberation	11.1	32.4	21.8
Y. Tachioka et al. [46]	MMI-SGMM	MFCC and PLP	Single-channel dereverberation with estimation of reverberation time	10.05	28.06	19.01
This paper	MMI-SGMM	MFCC	One-step environment-dependent DAE	7.04	28.66	17.85
This paper	DNN	MFCC	One-step environment-dependent DAE	7.46	28.11	17.79
This paper	MMI-SGMM +DNN	MFCC	One-step environment-dependent DAE	6.41	26.83	16.62

4.2.3 Comparison with results of the other participants in the REVERB-challenge

We compared our results with those of the other participants under the same conditions for the training data and language model. A single-channel dataset provided by the REVERB-challenge was used.

Table 6 shows the speech recognition results using the trigram language model for each participant. The WER of Alam et al. [45] was 11.1 % on SimData and 32.4 % on RealData. Tachioka et al. [46] achieved a WER of 10.05 % on SimData and 28.06 % on RealData. In our study, for the Eval. dataset, WER was 7.04 % on SimData and 28.66 % on RealData using MMI-SGMM, and 6.41 % on SimData and 26.83 % on RealData using SGMM+DNN.

The results indicate that the performance of our proposed environment-dependent DAE is better than almost all the other participants' methods using the same training data and language model.

5 Conclusions

In this paper, we proposed two environment-dependent DAE for robust distant-talking speech recognition. The proposed method was evaluated using simulated and real distant-talking speech. DAE-based cepstral-domain dereverberation achieved a remarkable improvement compared with CMN- and MSLP-based dereverberation in both environments. Furthermore, speech recognition performance was improved by the environment-dependent DAE compared with the conventional environment-independent DAE. For SimData in the Eval. using the one-step environment-dependent DAE with reverberation features estimated by MSLP, the average WER was reduced from 7.12 % with the conventional DAE to 6.41 % using SGMM+DNN, i.e., a relative error reduction rate of 9.97 %. For RealData in the Eval. dataset, the average WER was reduced from 30.56 % with the conventional DAE to 26.83 % using SGMM+DNN, i.e., a relative error reduction rate of 12.21 %. The results of our proposed dereverberation method are better than almost all of those of the other participants in the REVERB-challenge for single-channel speech and trigram language model conditions.

Endnote

[1] W_i and $W_{i_1^T}$ correspond to f_L in Eq. 1.

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

Acknowledgements
This work was supported by JSPS KAKENHI Grant Number 15K16020, a research grant from the Telecommunications Advancement Foundation (TAF), Japan and a research grant from the Kayamori Foundation of Informational Science Advancement.

Author details
[1]Graduate School of Engineering, Shizuoka University, Johoku Naka-ku, Hamamatsu 432-8561, Japan. [2]Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka 940-2188, Japan.

Received: 18 February 2015 Accepted: 2 November 2015
Published online: 12 November 2015

References

1. T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)
2. M Wu, D Wang, A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. ASLP. **14**(3), 774–784 (2006)
3. Q Jin, T Schultz, A Waibel, Far-field speaker recognition. IEEE Trans. ASLP. **15**(7), 2023–2032 (2007)
4. M Delcroix, T Hikichi, M Miyoshi, Precise dereverberation using multi-channel linear prediction. IEEE Trans. ASLP. **15**(2), 430–440 (2007)
5. EA Habets, in *Proc.of IEEE ICASSP*. Multi-channel speech dereverberation based on a statistical model of late reverberation (IEEE Pennsylvania Convention Center, Philadelphia, Pennsylvania, USA, 2005), pp. 173–176
6. L Wang, N Kitaoka, S Nakagawa, Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm.IEICE Trans. Inf. Syst. **E94-D**(3), 659–667 (2011)
7. L Wang, K Odani, A Kai, Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array. Eurasip J. Adv. Signal Process. **2012**(12), 1–11 (2012)
8. W Li, L Wang, F Zhou, Q Liao, in *Proc.of IEEE ICASSP*. Joint sparse representation based cepstral-domain dereverberation for distant-talking speech recognition (IEEE Vancouver Convention & Exhibition Center, Vancouver, BC, Canada, 2013), pp. 7117–7120
9. H Hirsch, H Finster. Speech Comm. **50**(3), 244–263 (2008)
10. A Sehr, R Maas, W Kellermann, Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. IEEE Trans. ASLP. **18**(7), 1676–1691 (2010)
11. SO Sadjadi, JHL Hasnen, in *Proceedings of IEEE ICASSP*. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions (IEEE Prague, Czech Republic, 2011), pp. 5448–5451
12. K Kinoshita, M Delcroix, T Nakatani, M Miyoshi, in *Proceedings of IEEE ICASSP 2006*. Spectral subtraction steered by multistep forward linear prediction for single channel speech dereverberation (IEEE Toulouse, France, 2006), pp. 817–820
13. L Wang, N Kitaoka, S Nakagawa, Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. IEICE Trans. Inf. Syst. **E94-D**(3), 659–667 (2011)
14. L Wang, Z Zhang, A Kai, in *Proc.of IEEE ICASSP 2013*. Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach (IEEE Vancouver Convention & Exhibition Center, Vancouver, BC, Canada, 2013), pp. 7224–7228
15. S Furui, Cepstral Analysis Technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**(2), 254–272 (1981)
16. F Liu, R Stern, X Huang, A Acero, in *Proc.ARPA Speech Nat. Lang. Workshop*. Efficient cepstral normalization for robust speech recognition, (1993), pp. 69–74
17. L Wang, N Kitaoka, S Nakagawa, in *Proc.of ICASSP*. Robust distant speech recognition by combining position-dependent CMN with conventional CMN (IEEE Honolulu, Hawaii, USA, 2007), pp. 817–820
18. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. **27**(2), 113–120 (1979)
19. M Wolfel, Enhanced speech features by single-channel joint compensation of noise and reverberation. IEEE Trans. Audio Speech Lang. Process. **17**(2), 312–323 (2009)
20. Y Konig, L Heck, M Weintraub, K Sonmez, in *Proc.of RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications*. Nonlinear discriminant feature extraction for robust text-independent speaker recognition (ESCA, 1998), pp. 72–75
21. Q Zhu, A Stolcke, B-Y Chen, N Morgan, in *Proc.of INTERSPEECH 2005*. Using MLP features in SRI's conversational speech recognition system, (2005), pp. 2141–2144
22. G Hinton, R Salakhutdinov, Reducing the dimensionality of data with neural networks. Science. **313**(5786), 504–507 (2006)
23. Y Xu, J Du, L-R Dai, C-H Lee, An experimental study on speech enhancement based on deep neural net-works. IEEE Signal Proc. Lett. **21**(1), 65–68 (2014)
24. F-J Weninger, S Watanabe, J-L Roux, J Hershey, Y Tachioka, J-T Geiger, G Rigoll, B-W Schuller, *The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement*, (Florence, Italy, 2014)
25. F Weninger, J Geiger, M Wollmer, B Schuller, G Rigoll, Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. Comput. Speech Lang. **28**(4), 888–902 (2014)
26. F Weninger, S Watanabe, Y Tachioka, B Schuller, in *Proc.of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition (IEEE Florence, Italy, 2014), pp. 4623–4627
27. X Xiao, S Zhao, DHH Nguyen, X Zhong, D-L Jones, ES Chng, H Li, in *proceedings of Reverberation Challenge Workshop*. The NTU-ADSC systems for Reverberation Challenge 2014 (Florence, Italy, 2014)
28. M Mimura, S Sakai, T Kawahara, in *Proc.of ICASSP*. Deep autoencoders augmented with phone-class feature for reverberant speech recognition (IEEE Brisbane, Queensland, Australia, 2015), pp. 4356–4369
29. S Araki, T Hayashi, M Delcroix, M Fujimoto, K Takeda, T Nakatani, in *Proceedings of ICASSP*. Exploring multi-channel features for denoising-autoencoder-based speech enhancement (IEEE Brisbane, Queensland, Australia, 2015), pp. 116–120
30. P Vincent, H Larochelle, I Lajoie, Y Bengio, PA Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
31. X Lu, Y Tsao, S Matsuda, C Hori, in *Proc.Interspeech*. Speech enhancement based on deep denoising autoencoder (ISCA Lyon, France, 2013), pp. 436–440
32. X Feng, Y Zhang, JR Glass, in *Proc.of ICASSP 2014*. Speech feature denising and dereverberation via deep autoencoder for noisy reverberant speech recognition (IEEE Florence, Italy, 2014), pp. 1759–1763
33. X Lu, Y Tsao, S Matsuda, C Hori, in *Proc.of INTERSPEECH 2014*. Ensemble modeling of denoising autoencoder for speech spectrum restoration (ISCA Singapore, 2014), pp. 885–889
34. T Ishii, H Komiyama, T Shinozaki, Y Horiuchi, S Kuroiwa, in *Proc.Interspeech*. Reverberant speech recognition based on denoising autoencoder (ISCA Lyon, France, 2013), pp. 3512–3516
35. Y Ueda, L Wang, A Kai, X Xiao, E Chng, H Li, in *Proc.of International Symposium on Chinese Spoken Language Processing 2014*. Single-channel dereverberation for distant-talking speech recognition by combining denoising autoencoder and temporal structure normalization (IEEE Singapore, 2014), pp. 379–383
36. ML Seltzer, D Yu, Y Wang, in *Proceedings of ICASSP*. An investigation of deep neural networks for noise robust speech recognition (IEEE Vancouver Convention & Exhibition Center, Vancouver, BC, Canada, 2013), pp. 7398–7402
37. R Giri, ML Seltzer, J Droppo, D Yu, in *Proceedings of ICASSP*. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning (IEEE Brisbane, Queensland, Australia, 2015), pp. 5014–5018
38. G Saon, H Soltau, D Nahamoo, M Picheny, in *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*. Speaker adaptation of neural network acoustic models using i-vectors (IEEE Olomouc, Czech Republic, 2013), pp. 55–59
39. XL Zhang, J Wu, Deep belief networks based voice activity detection. IEEE Trans. Audio, Speech, Lang. Process. **21**(4), 337–3408 (2013)
40. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, E Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech (IEEE Mohonk Mountain House in New Paltz, New York, USA, 2013)
41. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proc.ICASSP 95*. Wsjcam0: A British English speech corpus for large vocabulary continuous speech recognition (IEEE Detroit, Michigan, USA, 1995), pp. 81–84
42. M Lincoln, I McCowan, I Vepa, HK Maganti, in *Proc.ASRU*. The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments, (2005), pp. 357–362

43. D Povey, L Burget, et al., The subspace Gaussian mixture model—a structured model for speech recognition. Comput. Speech Lang. **25**(2), 404–439 (2011)
44. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, J Silovsky, G Stemmer, K Vesely, in *Proc.of IEEE 2011 workshop on, Automatic Speech Recognition and Understanding*. The Kaldi speech recognition toolkit (IEEE Hawaii, USA, 2011), pp. 1–4
45. J Alam, V Gupta, P Kenny, P Dumouchel, in *Proc.of REVERB Workshop*. Use of multiple front-ends and i-vector-based speaker adaptation for robust speech recognition (Florence, Italy, 2014)
46. Y Tachioka, T Narita, FJ Weninger, S Watanabe, in *Proc.of REVERB Workshop*. Dual system combination approach for various reverberant environments with dereverberation techniques (Florence, Italy, 2014)