DOI 10.1007/s11265-015-1007-3

Single-channel Dereverberation for Distant-Talking Speech Recognition by Combining Denoising Autoencoder and Temporal Structure Normalization

Yuma Ueda[1] · Longbiao Wang[2] · Atsuhiko Kai[1] · Xiong Xiao[3] · Eng Siong Chng[4] · Haizhou Li[5]

Received: 15 November 2014 / Revised: 9 April 2015 / Accepted: 21 April 2015 / Published online: 15 May 2015

Abstract In this paper, we propose a robust distant-talking speech recognition by combining cepstral domain denoising autoencoder (DAE) and temporal structure normalization (TSN) filter. As DAE has a deep structure and nonlinear processing steps, it is flexible enough to model highly nonlinear mapping between input and output space. In this paper, we train a DAE to map reverberant and noisy speech features to the underlying clean speech features in the cepstral domain. For the proposed method, after applying a DAE in the cepstral domain of speech to suppress reverberation, we apply a post-processing technology based on temporal structure normalization (TSN) filter to reduce the noise and reverberation effects by normalizing the modulation spectra to reference spectra of clean speech. The proposed method was evaluated using speech in simulated and real reverberant environments. By combining a cepstral-domain DAE and TSN, the average Word Error Rate (WER) was reduced from 25.2 % of the baseline system to 21.2 % in simulated environments and from 47.5 % to 41.3 % in real environments, respectively.

Keywords Speech recognition · Dereverberation · Denoising autoencoder · Environment adaptation · 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. According to [1], the approaches in dealing with reverberation problem can be classified as front-end-based and back-end-based approaches. The front-end-based approaches [1–10] attempt to reduce the effect of reverberation from the observed speech signal. The back-end-based methods attempt to modify the acoustic model and/or decoder so that they are suitable for reverberant environment [11, 12]. In this paper, we focus on the front-end-based approached for distant-talking speech recognition.

✉ Longbiao Wang
wang@vos.nagaokaut.ac.jp

Yuma Ueda
ueda@spa.sys.eng.shizuoka.ac.jp

Atsuhiko Kai
kai@sys.eng.shizuoka.ac.jp

Xiong Xiao
xiaoxiong@ntu.edu.sg

Eng Siong Chng
aseschng@ntu.edu.sg

Haizhou Li
hli@i2r.a-star.edu.sg

[1] Graduate School of Engineering, Shizuoka University, Hamamatsu 432-8561, Japan

[2] Nagaoka University of Technology, Nagaoka 940-2188, Japan

[3] Temasek Laboratories @ NTU, Nanyang Technological University, Singapore 138632, Singapore

[4] School of Computer Engineering, Nanyang Technological University, Singapore 138632, Singapore

[5] Human Language Technology, Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore

Many single-channel and multi-channel dereverberation methods have been proposed for robust distant-talking speech/speaker recognition [2–4, 13–17]. Comparing to microphone array, single microphone is much easier and cheaper to be implemented for real applications. Several single-channel dereverberation approaches have been proposed [2–4, 13, 14]. Cepstral mean normalization (CMN) [18–20] may be considered the most general approach. It has been extensively examined and shown as 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 the above problem [3, 4, 14]. A reverberation compensation method for speaker recognition using spectral subtraction [21], in which late reverberation is treated as additive noise, was proposed in [3]. A method based on multi-step linear prediction (MSLP) was proposed by [4, 14] for both single and multiple microphones. The method first estimates late reverberations using long-term multistep linear prediction, 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 multistep linear prediction, into a framework, which, so far tracks and removes nonstationary additive distortion by particle filters in a low-dimension logarithmic power frequency domain [22].

Neural network (NN) based approaches have been proposed for feature transformation [23, 24]. Bottleneck features extracted by a multi-layer perceptron (MLP) can be used a non-linear feature transformation [23]. However, deep networks of MLP with many hidden layers have a high computational cost, and can't learn much further away from the top layer. Deep belief networks (DBNs) which employ an unsupervised pretraining method using restricted Boltzmann machine (RBM) have been proposed to train better initial values of deep networks [29]. DNNs with pretraining have been shown better performance than the conventional MLP without pretraining on automatic speech recognition [29]. Recently, denoising autoencoder (DAE), one of Deep Neural Network (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 [25, 26]. Ishii et al. applied a DAE for spectral-domain dereverberation [27] and found the word accuracy of LVCSR was improved from 61.4 % to 65.2 % for the JNAS database [28]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. Previously, we found that Deep Neural Network (DNN) [29] based cepstral-domain feature mapping is efficient for distant-talking speaker processing [30]. 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.

The DAE uses its flexible mapping capability to learn a mapping from a window of distorted input features to the clean output features. Due to the limitation on the model complexity, the input window of DAE cannot go too big. In this study, we use a window of 9 frames, which covers roughly 0.1s of speech. However, the effects of reverberation on speech features may be as long as 1 second. Therefore, the DAE is clearly not adequate to deal with the reverberation distortion by itself. In this paper, we apply temporal structure normalization (TSN) [31, 32] to complement DAE for the task of dereverberation. TSN was previously proposed to reduce the effects of transmission channel and additive background noise on speech features for robust speech recognition. It is motivated by the observation that noise and channel modifies the temporal structure of speech features, hence there is a need to restore the clean temporal structure. The furthermore improvement is expected for distant-talking speech recognition by combining cepstral-domain DAE and the TSN filter based feature normalization. The proposed method is evaluated in both simulated and real reverberant environments.

The remainder of this paper is organized as follows: Section 2 describes denoising autoencoder for cepstrl-domain dereverberation. Temporal structure normalization is described in Section 3. The experimental results and discussions are presented in Section 4. Finally, Section 5 summarizes the paper.

2 Denoising Autoencoder for Cepstral-Domain Dereverberation

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 is a noisy version of the output data. Denoising autoencoder use feature mapping to convert noisy input data into clean output, and have been used for noise removal in the field of image processing [25]. Ishii et al. applied a DAE for spectral-domain dereverberation [27]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. Noting that many speech recognition systems adopt a cepstral-domain feature as the direct input, we think that the transformation of cepstral-domain feature may achieve better performance than that of spectral-domain feature. Cepstral-domain denoising autoencoder based dereverberation transforms the cepstrum of reverberant speech to that of clean speech. By using Mel filterbank, the cepstral features take into consideration the fact that human auditory system has higher resolution in low frequency than in high frequency. Hence, error of dereverberant signal and clean teacher signal on cepstral feature automatically emphasizes more on low frequencies than high frequencies. On the other hand, if error on spectrum is used, all frequencies are treated equally important. Moreover, the dimensions of the spectral-domain based features are greater than those of cepstral-domain based features. This introduces greater difficulty in learning for DAE with a deep architecture. Thus, it is expected that the DAE-based cepstral-domain dereverberation should be more efficient than DAE-based spectral-domain dereverberation for speech recognition. 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 featureX_i of observed reverberant speech of $i-th$ 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:

$$O_i = f_L(...f_l(...f_2(f_1(X_i, X_{i-1}, ..., X_{i-N}))) \qquad (1)$$

where f_l is the non-linear transformation function in layer l, 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. Details of parameter turning for DAE is discussed in Section 4.2.1. 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 matrix and W_1^T and W_2^T are the decoder matrix, respectively.

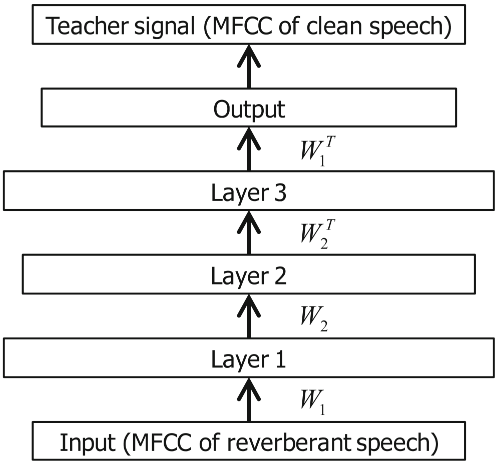
Teacher signal (MFCC of clean speech)
Output
W_1^T
Layer 3
W_2^T
Layer 2
W_2
Layer 1
W_1
Input (MFCC of reverberant speech)

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

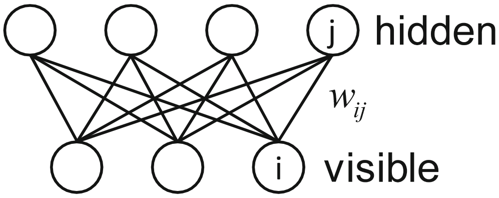

Figure 2 Graphical representation of the RBM.

2.1 Training of DAE

2.1.1 Restricted Boltzmann Machine

To train a deep neural network, Deep Belief Networks (DBNs) [29] 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 visible and hidden layer in which visible units that represent observations are connected to hidden units that learn to represent features using weighted connection. A RBM is restricted that there are novisible-visible or hidden-hidden connections. Different types of RBM is used in the case of binary or real-valued input. Bernoulli-Bernoulli RBMs used to convert binary stochastic variables to binary stochastic variables. Gaussian-Bernoulli RBMs is used to convert real-valued stochastic variables to binary stochastic variables.

In a Bernoulli-Bernoulli RBMs, the weights on the connections and the biases of the individual units define a probabillity distribution over the joint states of the visible and hidden units via an energy function. The energy of a joint configuration is:

$$E(\mathrm{v}, \mathrm{h}|\theta) = -\sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij} v_i h_j - \sum_{i=1}^{\mathcal{V}} a_i v_i - \sum_{j=1}^{\mathcal{H}} b_j h_j \qquad (2)$$

where $\theta = (w, a, b)$ and w_{ij} represents the symmetric interaction term between visible unit i and hidden unit j while a_i and b_j are their bias term. $\mathcal{V}$ and $\mathcal{H}$ are the numbers of visible and hidden units.

The probability that a RBM assigns to a visible vector v is:

$$p(\mathrm{v}|\theta) = \frac{\sum_{\mathrm{h}} exp(-E(\mathrm{v}, \mathrm{h}))}{\sum_{\mathrm{v}}\sum_{\mathrm{h}} exp(-E(\mathrm{v}, \mathrm{h}))} \qquad (3)$$

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

Since there are no hidden-hidden connections, the conditional distribution $p(\mathrm{h}|\mathrm{v}, \theta)$ is factorial and is given by:

$$p(h_j = 1|\mathrm{v}, \theta) = \sigma\left(b_j + \sum_{i=1}^{\mathcal{V}} w_{ij} v_i\right) \tag{4}$$

where $\sigma(x) = (1 + exp(-x))^{-1}$. Similarly, since there are no visible-visible connections, the conditional distribution $p(\mathrm{v}|\mathrm{h}, \theta)$ is factorial and is given by:

$$p(v_j = 1|\mathrm{h}, \theta) = \sigma\left(a_i + \sum_{j=1}^{\mathcal{H}} w_{ij} h_j\right) \tag{5}$$

In a Gaussian-Bernoulli RBMs, the energy of a joint configuration is:

$$E(\mathrm{v}, \mathrm{h}|\theta) = \sum_{i=1}^{\mathcal{V}} \frac{(v_i - a_i)^2}{2} - \sum_{i=1}^{\mathcal{V}} \sum_{j=1}^{\mathcal{H}} w_{ij} v_i h_j - \sum_{j=1}^{\mathcal{H}} b_j h_j \tag{6}$$

The conditional distribution $p(\mathrm{h}|\mathrm{v}, \theta)$ is factorial and is given by:

$$p(v_i = 1|h, \theta) = N\left(v_i; a_i + \sum_{j=1}^{\mathcal{H}} w_{ij} h_j, 1\right) \tag{7}$$

where $N(\mu, \mathcal{V})$ is a Gaussian with mean μ and variance V.

Maximum likelihood estimation of RBM is to maximize the log likelihood $log(p(\mathrm{v}|\theta))$ for the parameters θ. Therefore, the weight update equation is given by:

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{8}$$

where $\langle \cdot \rangle_{data}$ is the expectation that v_i and h_j are on together in the training set and $\langle \cdot \rangle_{model}$ is the same expectation calculated from the model. Because compute $\langle v_i h_j \rangle$ is expensive, using contrastive divergence (CD) approximation for the compute gradient. It is possible to compute $\langle v_i h_j \rangle$ by once the Gibbs sampling.

To obtain a pre-trained RBM, we trained all hidden layers by using the Bernoulli-Bernoulli RBM. DBNs are hierarchically configured by connecting these pre-trained RBMs. In DAE, output is reconstruction of input, so network's first half layer are for encoding, and second half layer are for decoding. 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.

2.1.2 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 feature of clean speech) are given.

3 Temporal Structure Normalization

Noise and reverberation distorts speech features in multiple aspects, e.g. the timbre of speech and also the temporal structure of speech up to 1 second. The DAE uses its flexible mapping capability to learn a mapping from a window of distorted input features to the clean output features. Due to the limitation on the model complexity, the input window of DAE cannot go too big. In this study, we use a window of 9 frames, which covers roughly 0.1s of speech. However, the effects of reverberation on speech features may be as long as 1s. For example, the evaluation data in this study has a T60 time up to 0.7s. Therefore, the DAE is clearly not adequate to deal with the reverberation distortion by itself.

In this section, we describe a method called temporal structure normalization (TSN) to complete DAE for the task of dereverberation. TSN was previously proposed to reduce the effects of transmission channel and additive background noise on speech features for robust speech recognition. It is motivated by the observation that noise and channel modifies the temporal structure of speech features, hence there is a need to restore the clean temporal structure. In TSN, temporal structure of speech features are represented by the modulation spectra of speech signal, i.e. the power spectral density function (PSD) of feature trajectories. The restoration of temporal structure of clean features is implemented by performing a linear filtering on the feature trajectories, where the linear filter weights are designed independently for each feature trajectory and speech utterance to normalize the PSD of the feature trajectory to a reference PSD. The reference PSD of a feature trajectory is estimated as the mean of PSD of clean utterances and represent the clean temporal structure of clean speech.

In our preliminary study, we found that the PSD function of the clean and reverberant speech features are very different. This is actually expected as reverberation is known to have a blurring effect on the speech spectrum, hence introduce a smoothing effects on the spectrum and hence the features. Therefore, it is natural to apply TSN to deal with the remaining distortions that exist in speech features after DAE.

4 Experiments

4.1 Experimental Setup

4.1.1 Training Dataset

The training dataset provided by gREVERB challengeh (Reverberant Voice Enhancement and Recognition Benchmark) [33] was used. This dataset consists of the clean WSJCAM0 [34] 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 sec. This training dataset was used for both training of acoustic models. Clean speeches are also used by training the reference PSD function of the TSN.

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

4.1.2 Evaluation Test Set

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 RealData in terms of reverberation time and microphone-speaker distance. This setup allows us to perform evaluations in terms of both practicality and robustness in various reverberant conditions. Specifically, the development (Dev.) test set and the final evaluation (Eval.) test set each consists of the following SimData and RealData: SimData is generated from WSJCAM0 corpus [34], and RealData from MC-WSJ-AV corpus [35]. This development dataset was used to determine the optimal parameter for dereverberation and speech recognition. The details of data set of training and test are shown 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. The DAE training was carried out using stochastic mini-batch gradient descent with a mini-batch size of 256 samples. 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.

Multi-step linear prediction (MSLP) algorithm generate inverse filter through the prediction coefficients to estimate inverse system [14]. We estimate the late reverberation components using the inverse filter and apply 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. For the TSN filter, the Yule-Walker method is used to estimate the PSD functions of feature trajectories. The order of the AR model for PSD estimation is set to 6 to obtain proper level of details. A filter length of 33 taps is used for the evaluation if not otherwise stated.

In this study, we used a speech recognition system provided by the gREVERB challengeh task [33], which is based on the hidden Markov model tool kit (HTK) [36]. As an acoustic model, it employs tied-state HMMs with eight Gaussian components per state, trained according to the maximum-likelihood criterion. We use a multi-condition training set for training of acoustic model. This training set 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 roughly from 0.1 to 0.8 sec. Note that the recording rooms used for the SimData, RealData and multi-condition training data are all different. CMLLR [37] is the method for converting the mean and variance of the Gaussian distribution for each state of the hidden Markov models (HMMs) by using the regression matrix to reduce the mismatch between the adaptation data and model. This method is intended to obtain a transformation matrix for modifying the model parameters that maximize the likelihood of the adaptation data. In this paper, we applied CMLLR for unsupervised model adaptation, i.e., environment adaptation.

4.2 Experimental Results

4.2.1 Parameters Tuning for DAE

For DAE-based dereverberation, feature vectors of the current frame and previous eight frames of reverberant speech were used as input. Thirty-nine MFCCs of the current frame of clean speech were used as teacher signals for output, i.e., the dimension of input was 39 × 9 = 351. Optimum value of number of hidden layer and units in each hidden layer were determined from the experimental. Table 3 shows the

Table 1 Quantity of data for Dev. and Eval. set of SimData and RealData and for training dataset.[2]

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

[2] 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	
		Room1	Room2	Room3		Near	Far
SimData	WSJCAM0	0.25s	0.5s	0.7s	20dB	50cm	200cm
RealData	MC-WSJ-AV	0.7s	–	–	–	100cm	250cm

Table 3 Word error rate by DAE-based dereverberation with different number of hidden layers (%).

Number of hidden layer	SimData							RealData		
	Room1		Room2		Room3		Ave.	Room1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
3	16.22	17.85	18.93	28.89	20.28	30.12	22.05	42.30	42.72	42.51
5	15.98	18.24	19.05	28.37	21.09	31.70	22.41	42.67	45.66	44.17
7	16.72	19.64	20.14	32.63	21.93	35.88	24.49	47.16	48.46	47.81
9	19.22	22.00	23.61	36.55	24.51	37.64	27.26	49.03	50.17	49.60

Table 4 Word error rate by DAE-based dereverberation with different number of units in each hidden layer (%).

Units in each hidden layer	SimData							RealData		
	Room1		Room2		Room3		Ave.	Room1		Ave.
	near	far	Near	Far	Near	Far		Near	Far	
1024-1024-1024	16.22	17.85	18.93	28.89	20.28	30.12	22.05	42.30	42.72	42.51
512-512-512	16.00	18.19	18.56	27.68	20.20	30.37	21.83	42.67	44.22	43.45
1024-512-1024	16.69	19.22	20.56	33.05	22.45	34.77	24.46	47.35	47.51	47.43
512-256-512	16.59	19.42	19.72	31.53	20.90	32.89	23.51	46.29	45.04	45.67

Table 5 Word error rate by spectral-domain DAE and cepstral-domain DAE (%).

DAE-based dereverberation	SimData							RealData		
	Room1		Room2		Room3		Ave.	Room1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
Spectral-domain	14.23	17.70	21.59	38.30	22.97	41.44	26.04	48.72	46.82	47.77
Cepstral-domain	16.22	17.85	18.93	28.89	20.28	30.12	22.05	42.30	42.72	42.51

speech recognition results with different number of hidden layer (1024 hidden units in each hidden layer). According to these results, the number of hidden layer of DAE is set to 3 in the following part of this paper. It is confirmed that the performance of speech recognition is decreased when the number of hidden layer is increased. This can be explained by the complex structure of DNNs: too many layers cause an increase in parameters of DNNs, with the output being overlearned using small training data. In this paper, the size of training data is less than 20 hours, so an appropriate number of layers is sufficient in this task. Table 4 shows the results with different units in each hidden layer. We decided that the number of units in each hidden layer is 1024 according to these results.

4.2.2 Comparison of Spectral-Domain DAE and Cepstral-Domain DAE

Ishii et al. applied a DAE for spectral-domain dereverberation [27] for the JNAS database [28]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. In our study, we applied DAE for cepstral-domain dereverberation for the REVERB-challenge task [33]. In this section, we compare spectral-domain and cepstral-domain DAE-based dereverberation for this task. The results are compared in Table 5. These results indicate that cepstral-domain DAE is better than spectral-domain DAE for reverberant speech recognition on "REVERB Challenge" task.

A possible reason for the better results of cepstral domain DAE is that the measure square error (MSE) of cesptral features are used as the cost function, which is more relevant to the speech recognition task than the MSE of the spectrum. By using Mel filterbank, the cepstral features take into consideration that fact human auditory system has higher resolution in low frequency than in high frequency. Hence, MSE on cesptral feature automatically emphasizes more on low frequencies than high frequencies. On the other hand, if MSE on spectrum is used, all frequencies are treated equally important.

4.2.3 Results by Combining DAE and TSN

Tables 6 and 7 show the speech recognition results with Dev. and Eval. dataset. We compared three kinds of dereverberation methods (MSLP, DAE and TSN) and the combination of these. CMN was applied to all methods. When DAE-based cepstral-domain dereverberation was compared with CMN-based dereverberation, single channel MSLP-based dereverberation and TSN filter-based dereverberation, a remarkable improvement was achieved. DAE worked well especially with strong reverberation, i.e., far-field microphone in gRoom 2hand gRoom 3h of SimData. The performance with CMLLR-based environment adaptation was better than that without CMLLR. However, the results of DAE are worse than that of baseline in gRoom1. This trend is also seen in Table 5. It is considered that the late reverberation in gRoom 1h is relative small and early reverberation can be suppressed by CMN effectively. Hence,

Table 6 Word error rate of each method for Dev. dataset (%).

Dereverberation methods	CMLLR	SimData							RealData		
		Room1		Room2		Room3		Ave.	Room1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
Without dereverberation	None	20.65	20.60	25.59	48.58	28.56	48.89	32.15	79.66	79.02	79.34
	Yes	15.73	18.98	21.32	40.18	24.36	44.21	27.46	56.77	57.42	57.10
CMN	None	16.13	18.85	22.73	43.06	26.73	46.43	28.99	52.21	52.36	52.29
	Yes	13.86	17.01	21.30	36.23	22.77	39.79	25.16	47.79	47.16	47.48
MSLP	None	15.44	18.34	23.10	41.43	26.95	45.38	28.44	53.96	53.25	53.61
	Yes	13.94	16.72	20.56	35.40	22.87	39.02	24.75	48.16	45.39	46.78
DAE	None	17.31	19.10	18.86	30.17	21.74	33.14	23.39	45.54	46.82	46.18
	Yes	16.22	17.85	18.93	28.89	20.28	30.12	22.05	42.30	42.72	42.51
TSN	None	16.57	20.18	23.69	40.25	28.39	44.96	29.01	50.22	51.47	50.85
	Yes	15.36	18.29	21.84	35.99	23.44	40.31	25.87	47.04	46.82	46.93
DAE+TSN	None	17.60	18.46	18.88	29.21	21.61	31.50	22.88	45.23	44.16	44.70
	Yes	15.46	17.43	19.05	27.38	19.26	28.54	21.19	41.92	40.60	41.26

Table 7 Word error rate of each method for Eval. dataset (%).

Dereverberation methods	CMLLR	SimData							RealData		
		Room1		Room2		Room3		Ave.	Room1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
Without dereverberation	None	22.30	23.21	25.33	42.97	29.16	47.23	31.70	80.45	74.48	77.47
	Yes	17.57	18.96	22.53	34.96	25.61	41.97	26.93	57.07	56.82	56.95
CMN	None	21.18	20.69	22.74	38.54	28.29	45.21	29.44	57.75	54.32	56.04
	Yes	16.62	19.25	21.51	32.58	25.06	39.87	25.82	49.89	47.23	48.56
MSLP	None	19.55	20.52	22.16	37.26	27.93	43.99	28.57	58.29	54.83	56.56
	Yes	16.08	18.37	20.90	31.58	25.01	38.58	25.09	49.70	46.76	48.23
DAE	None	21.30	21.92	20.61	28.01	23.80	32.43	24.68	49.50	49.43	49.47
	Yes	17.93	19.38	20.02	27.46	22.23	30.47	22.92	44.14	46.19	45.17
TSN	None	22.13	23.59	23.78	36.05	28.44	44.16	29.69	55.38	52.73	54.06
	Yes	16.88	19.45	22.05	31.74	25.51	40.06	25.95	49.89	47.03	48.46
DAE+TSN	None	21.08	21.72	19.95	27.59	24.07	31.68	24.35	47.97	48.14	48.06
	Yes	17.84	19.11	19.47	27.16	21.75	29.93	22.54	44.04	44.02	44.03

the merit of suppression of late reverberation under light reverberant condition is not very large. DAE+TSN cause some distortion by doing dereverberation due to mismatch between gRoom 1h condition and training condition. So the performance of DAE+TSN is worse than CMN when RT60 is very small. The results indicate that the proposed methods work better in heavy reverberation than in light reverberation.

For Dev. dataset, the average word error rates (WERs) in SimData were improved from 25.16 % of CMN to 22.05 % of cepstral-domain DAE with CMLLR-based environment adaptation. In RealData, WER were improved from 47.48 % to 42.51 % with CMLLR-based environment adaptation. In SimData, by combining cepstral-domain DAE and TSN filter with environment adaptation, the WER was reduced from 25.16 % in the baseline state to 21.19 %, i.e., the relative error reduction rate was 15.8 %. In RealData, the WER was reduced from 47.48 % to 41.26 %, i.e., the relative error reduction rate was 13.1 %. TSN filter didn't work well alone due to reverberation . However, when combined with cepstral-domain DAE, improvement of TSN filter was increased. It was considered that noise reduction capability of TSN filter was improved by dereverberation of cepstral-domain DAE.

For Eval. dataset, the similar trend in Dev. Dataset was obtained. The WERs in SimData were improved from 25.82 % of CMN to 22.92 % of cepstral-domain DAE with CMLLR-based environment adaptation. In RealData, WER were improved from 48.56 % to 45.17 % with CMLLR-based environment adaptation. The proposed combination of DAE and TSN achieved best speech recognition performance. That is, combination of DAE and TSN outperformed the other dereverberation methods for both Dev. dataset and Eval. dataset.

We found that the combinations with MSLP does not produce good results. MSLP work well for dereverberation, but does not work well in combination with other dereverberation methods.

5 Conclusions

In this paper, we proposed a robust distant-talking speech recognition method by combining the cepstral-domain DAE and the TSN filter. The proposed method was evaluated in simulated and real distant-talking environments. DAE-based cepstral-domain dereverberation achieved a remarkable improvement compared with CMN-based dereverberation, MSLP-based dereverberation and TSN filter-based feature normalization in both environments. Furthermore, speech recognition performance was improved by combining the cepstral-domain DAE and the TSN filter. In SimData of Dev. dataset, by combining cepstral-domain DAE and TSN filter with environment adaptation, the WER was reduced from 25.16 % in the baseline state to 21.19 %, i.e., the relative error reduction rate was 15.8 %. In RealData of Dev. dataset, the WER was reduced from 47.48 % to 41.26 %, i.e., the relative error reduction rate was 13.1 %. For Eval. dataset, the similar trend was obtained. In SimData of Eval. dataset, the WER was reduced from 25.82 % to 22.54 %, i.e., the relative error reduction rate was 12.7 %. In RealData of Eval. dataset, the WER was reduced from 48.56 % to 44.03 %, i.e., the relative error reduction rate was 9.33 %.

Acknowledgments This work was partially supported by a research grant from the Research Foundation for the Electrotechnology of Chubu (REFEC).

References

1. Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., & Kellermann, W. (2012). Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 114–126.
2. Wu, M., & Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on ASLP*, *14*(3), 774–784.
3. Jin, Q., Schultz, T., & Waibel, A. (2007). Far-field speaker recognition. *IEEE Transactions on ASLP*, *15*(7), 2023–2032.
4. Delcroix, M., & Hikichi, T. (2007). M.Miyoshi, Precise dereverberation using multi-channel linear prediction. *IEEE Transactions on ASLP*, *15*(2), 430–440.
5. Wang, L., Zhang, Z., & Kai, A. (2013). Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach. *Proceedings of ICASSP*, *2013*, 7224–7228.
6. Habets, E.A. (2005). Multi-channel speech dereverberation based on a statistical model of late reverberation. *Proceedings of IEEE ICASSP*, 173–176.
7. Wang, L., Kitaoka, N., & Nakagawa, S. (2006). Robust Distant Speech Recognition by Combining Multiple Microphone-array Processing with Position-dependent CMN. *Eurasip Journal on Applied Signal Processing*, *2006*(95491), 1–11.
8. Wang, L., Kitaoka, N., & Nakagawa, S. (2011). Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. *IEICE Transactions on Information Systems*, *E94-D*(3), 659–667.
9. Wang, L., Odani, K., & Kai, A. (2012). Dereverberation and denoising based on generalized spectral subtraction by nutil-channel LMS algorithm using a small-scale microphone array. *Eurasip Journal on Advances in Signal Processing*, *2012*(12), 1–11.
10. Li, W., Wang, L., Zhou, F., & Liao, Q. (2013). Joint sparse representation based cepstral-domain dereverberation for distant-talking speech recognition. *Proceedings of IEEE ICASSP*, 7117–7120.
11. Hirsch, H., & Finster, H. (2008). A new approach for the adaptation of HMMs to reverberation and background noise. *Speech Communication*, *50*(3), 244–263.
12. Sehr, A., Maas, R., & Kellermann, W. (2010). Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. *IEEE Transactions on ASLP*, *18*(7), 1676–1691.
13. Sadjadi, S.O., & Hasnen, J.H.L. (2011). Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *Proceedings of IEEE ICASSP* (pp. 5448–5451).
14. Kinoshita, K., Delcroix, M., Nakatani, T., & Miyoshi, M. (2006). Spectral subtraction steered by multistep forward linear prediction for single channel speech dereverberation. In *Proceedings of IEEE ICASSP* (Vol. 2006, pp. 817–820).
15. Wang, L., Odani, K., & Kai, A. (2012). Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array. *EURASIP Journal on Advances in Signal Processing, 2012,* 12.
16. Wang, L., Kitaoka, N., & Nakagawa, S. (2011). Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. *IEICE Transactions on Information and Systems*, *E94-D*(3), 659–667.
17. Wang, L., Zhang, Z., & Kai, A. (2013). Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach. *Proceedings of IEEE ICASSP*, *2013*, 7224–7228.
18. Furui, S. (1981). Cepstral Analysis Technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 254–272.
19. Liu, F., Stern, R., Huang, X., & Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of ARPA Speech Natural Language Workshop* (pp. 69–74).
20. Wang, L., Kitaoka, N., & Nakagawa, S. (2007). Robust distant speech recognition by combining position-dependent CMN with conventional CMN. *Proceedings of ICASSP*, 817–820.
21. Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120.
22. Wolfel, M. (2009). Enhanced speech features by single-channel joint compensation of noise and reverberation. *IEEE Transactions on Audio Speech Language Processing*, *17* (2), 312–323.
23. Konig, Y., Heck, L., Weintraub, M., & Sonmez, K. (1998). Nonlinear discriminant feature extraction for robust text-independent speaker recognition. In *Proceedings of RLA2C: ESCA workshop on speaker recognition and its commercial and forensic applications* (pp. 72–75).
24. Zhu, Q., Stolcke, A., Chen, B.Y., & Morgan, N. (2005). Using MLP features in SRI's conversational speech recognition system. *INTERSPEECH*, *2005*, 2141–2144.
25. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*, 3371–3408.
26. Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013). Speech enhancement based on deep denoising autoencoder, In *Proceedings of Interspeech* (pp. 436–440).
27. Ishii, T., Komiyama, H., Shinozaki, T., Horiuchi, Y., & Kuroiwa, S. (2013). Reverberant speech recognition based on denoising autoencoder. In *Proceedings of Interspeech* (pp. 3512–3516).
28. Itou, K., Yamamoto, M., Takeda, K., Kakezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., & Itahashi, S. (1999). JNAS: Janpanese speech corpus for large vocabulary continuous speech recognition research. *J. Acoust. Soc. Jpn. (E)*, *20*(3), 199–206.
29. Hinton, G., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.
30. Yamada, T., Wang, L., & Kai, A. (2013). Improvement of distant-talking speaker identification using bottleneck features of DNN. In *Proceedings of Interspeech* (pp. 3661–3664).
31. Xiao, X., Chng, E.S., & Li, H. (2008). Normalization of the speech modulation spectra for robust speech recognition. *IEEE Transactions on Audio Speech, and Language Processing*, *16*(8), 1662–1674.
32. Xiao, X., Chng, E.S., & Li, H. (2007). Temporal structure normalization of speech feature for robust speech recognition. *IEEE Signal Processing Letters*, *14*(7), 500–503.
33. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W.,

Maas, R., Gannot, S., & Raj, B. (2013). The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics (WASPAA-13)*.

34. Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995). Wsjcam0: A british english speech corpus for large vocabulary continuous speech recognition. In *Proceedings of ICASSP* (Vol. 95, pp. 81–84).
35. Lincoln, M., McCowan, I., Vepa, I., & Maganti, H.K. (2005). The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. In *Proceedings of ASRU* (pp. 357–362).
36. Young, S., Kershow, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The HTK book (for HTK version 3.0)*: Cambridge University.
37. Gales, M.J.F., & Woodland, P.C. (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech & Language*, *10*, 249–264.

Yuma Ueda received his B.E. degree from Shizuoka University in 2014. He is currently pursuing the M.E. degree at Shizuoka University. His research interests include robust speech recognition and acoustic signal processing.

Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. and Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2005 and 2008 respectively.

From July 2000 to August 2002, he worked at the China Construction Bank. He was an assistant professor in the faculty of Engineering at Shizuoka University, Japan from April 2008 to September 2012. Since October 2013 he has been an associate professor at Nagaoka University of Technology, Japan. His research interests include robust speech recognition, speaker recognition and sound source localization. He received the “Chinese Government Award for Outstanding Self-financed Students Abroad” in 2008. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Atsuhiko Kai received his D.Eng. degree from Toyohashi University of Technology, Japan, in 1996 and joined the faculty as an assistant professor. He moved to Shizuoka University, Japan, in 1999, and is currently an associate professor.

His scientific interests are spoken language processing and dialogue processing with a focus on speech recognition.

He is a member of IEEE, Institute of Electronics, Information and Communication Engineers (IEICE), Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ) and Japanese Society for Artificial Intelligence (JSAI).

Xiong Xiao received the B. Eng. and Ph.D. degrees in computer engineering from Nanyang Technological University (NTU), Singapore, in 2004 and 2010, respectively. He is now a senior research scientist in the Temasek laboratories, NTU. His research interests include robust speech recognition, keyword search, speech enhancement, and signal processing.

Eng Siong Chng received the B.Eng. (honors) degree in electrical and electronics engineering and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1991 and 1996, respectively. He is currently an Associate Professor in the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. Prior to joining NTU in 2003, he was with the Institute of Physics and Chemical Research, Riken, as a Postdoctoral Researcher working in the area of signal processing and classification (1996), the Institute of System Science (ISS, currently known as I2R) as a member of research staff to transfer the Apple-ISS speech and handwriting technologies to ISS (1996-1999), Lernout and Hauspie (now part of Nuance) as a Senior Researcher in speech recognition (1999-2000), and Knowles Electronics as a Manager for the Intellisonic microphone array research (2001-2002). His research interests are in pattern recognition, signal, speech, and video processing. He has published over 100 papers in international journals and conferences. He is currently leading the speech and language technology program in the Emerging Research Lab at the School of Computer Engineering, NTU.

Haizhou Li received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently the Principal Scientist, Department Head of Human Language Technology in the Institute for Infocomm Research (I^2R), Singapore. He is also an Adjunct Professor at the National University of Singapore.

Prior to joining I^2R, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), a Research Manager at the Apple-ISS Research Centre (1996–1998), a Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and the Vice President in InfoTalk Corp. Ltd. (2001–2003).

Dr Li is currently the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2017). He has served in the Editorial Board of Computer Speech and Language (2012-2014). He is an elected Member of IEEE Speech and Language Processing Technical Committee (2013-2015), the Vice President of the International Speech Communication Association (2013-2014), and the President of Asia Pacific Signal and Information Processing Association (2015-2016). He was the General Chair of ACL 2012 and INTERSPEECH 2014.

Dr Li is a Fellow of the IEEE. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one the two Nokia Visiting Professors in 2009 by the Nokia Foundation.