Multimed Tools Appl (2016) 75:5093–5108
DOI 10.1007/s11042-015-2849-1

Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition

Bo Ren[1] · Longbiao Wang[1] · Liang Lu[2] · Yuma Ueda[3] · Atsuhiko Kai[3]

Received: 2 April 2015 / Revised: 16 July 2015 / Accepted: 31 July 2015 /
Published online: 3 September 2015

Abstract The performance of speech recognition in distant-talking environments is severely degraded by the reverberation that can occur in enclosed spaces (e.g., meeting rooms). To mitigate this degradation, dereverberation techniques such as network structure-based denoising autoencoders and multi-step linear prediction are used to improve the recognition accuracy of reverberant speech. Regardless of the reverberant conditions, a novel discriminative bottleneck feature extraction approach has been demonstrated to be effective for speech recognition under a range of conditions. As bottleneck feature extraction is not primarily designed for dereverberation, we are interested in whether it can compensate for other carefully designed dereverberation approaches. In this paper, we propose three schemes covering both front-end processing (cascaded combination and parallel combination) and back-end processing (system combination). Each of these schemes integrates bottleneck feature extraction with dereverberation. The effectiveness of these schemes is evaluated via a series of experiments using the REVERB challenge dataset.

Keywords Distant-talking speech recognition · Denoising autoencoder · Bottleneck feature · Dereverberation

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

Liang Lu
liang.lu@ed.ac.uk

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

[1] Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka, Niigata 940-2188, Japan

[2] Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

[3] Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8561, Japan

1 Introduction

Hands-free speech input techniques for various applications are increasingly popular. One driving force behind this is the rapid rise in the use of portable devices such as hands-free mobile telephones, tablets, and voice-controlled systems. However, in such distant-talking conditions, the speech recognition performance of today's automatic recognizers is much lower than that of systems that use a close-talking microphone. One significant impact is the presence of reverberation in an enclosed area (i.e., meeting rooms). In such conditions, the direct speech signal becomes overlapped by reflected signals with various delay times, and the delay is often longer than the length of the analysis window used in speech recognition systems. As a result, the special speech clues that are necessary for automatic speech recognition (ASR) become distorted.

A number of approaches have been proposed to overcome these problems. In particular, single microphone approaches have the advantage of usability from an application point of view. Cepstral mean normalization (CMN) [3, 15] is one of the simplest and most efficient approaches, and is known to be useful when the distortion has a short impulse response. However, CMN is not completely effective in environments with late reverberation. To handle such cases, several methods have focused on spectral domain processing, i.e., multi-channel least mean squares (LMS) [16, 26, 31] and multi-step linear prediction (MSLP) [11]. In addition, deep neural networks (DNNs) have been successfully applied to ASR [8, 19, 29]. The emerging deep learning paradigm may therefore enable novel approaches to address this challenge. Some pilot studies using an autoencoder to denoise and dereverberate the speech signals have been reported [10, 25]. The key point of these autoencoders is to train a sufficiently large network structure so as to reconstruct the clean speech from a noisy or reverberant version. However, it is difficult to obtain good recognition accuracy using this kind of single channel dereverberation.

To improve the speech recognition performance, DNNs have been found to be efficient in many speech recognition systems [1, 8]. Making use of the capabilities of DNNs, two popular configurations have been proposed. In the first configuration, a hybrid DNN is used to predict context-dependent hidden Markov model (HMM) states [20, 29]. The second configuration, which is called a tandem, exploits DNNs to perform nonlinear feature transformation. The transformed features are used as the input to either a Gaussian mixture model (GMM) or DNN-based acoustic model. The bottleneck feature (BF), the most famous example of a tandem DNN, has been adopted and enhanced in various systems [5, 7, 13, 18, 21, 27, 30], and improved performance has been observed in many tasks when the BF is used as a discriminative input [6, 14, 17, 28, 32].

Dereverberation methods can significantly suppress the degradation of both early and late reverberation, and tandem DNNs, especially BFs, are often employed in either noisy or reverberant speech recognition systems. In general, BFs and dereverberation are employed separately in ASR systems. Some limited studies have investigated the integration of dereverberation approaches and discriminative BFs. Because BF extraction is not primarily designed for reverberant speech conditions, we are interested in combining BF extraction with some carefully designed dereverberation approaches. In this paper, we describe the integration of BFs with a linear MSLP and a nonlinear denoising autoencoder (DAE). Three schemes, including both front- and back-end processing, are introduced and evaluated, and the effect of each scheme is evaluated using the REVERB challenge database [12], which is focused on reverberant speech recognition.

The rest of this paper is organized as follows. Section 2 reviews two dereverberation approaches. The BF is described in detail in Section 3, and Section 4 introduces the three

proposed schemes. The experimental procedure and results are discussed in Section 5. Finally, our conclusions and future work are described in Section 6.

2 Dereverberation approach

2.1 Denoising autoencoder

An autoencoder is a type of artificial neural network whose output is a reconstruction of the input. They are often used for dimension reduction [9]. DAEs share the same structure as autoencoders, but the input data is a noisy version of the output. 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 [24]. In speech recognition, DAEs have been applied for dereverberation in both the spectral domain [10] and the cepstral domain [22]. Because cepstral domain features such as mel-frequency cepstral coefficients (MFCCs) are conventionally used in ASR systems, we adopt the cepstral domain DAE described by Ueda et al. [22].

Unlike for conventional autoencoders, we use samples that include the corresponding clean and reverberant speech. The DAE learns the nonlinear conversion function that converts the reverberant speech into clean speech. In general, reverberation is dependent on several previous observation frames, as well as the current frame. In addition to the vector of the current frame, vectors from past frames are concatenated to form the input. For cepstral feature X_i of the observed reverberant speech in frame i, cepstral features from the $N-1$ frames before the current frame are concatenated to form a cepstral vector of N frames. Output O_i from the nonlinear DAE-based transformer is given by:

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

where f_l is the nonlinear transformation function in layer l. The topology of cepstral-domain DAEs for dereverberation is shown in Fig. 1. In this paper, we consider three hidden layers. Details of parameter tuning for DAEs were discussed in [22]. In Fig. 1, $W_i (i = 1, 2)$ denotes the weighting of the different layers, and W_i^T is the transpose of W_i.[1] That is to say, W_1 and W_2 form the encoder matrix, and W_1^T and W_2^T constitute the decoder matrix.

2.2 Multi-step linear prediction

The use of long-term MSLP for dereverberation has been described for both single and multiple microphones [11]. Long-term MSLP was originally used to estimate the entire impulse response of speech components [4]. For dereverberation, the long-term MSLP is generally used to identify only the late reverberation [11]. Assuming that $x_1(n)$ is the speech signal recorded by the first distant microphone, and that N and D are the number of filter coefficients and the step-size (i.e., delay), respectively, MSLP can be formulated as

$$x_1(n) = \sum_{p}^{N} w(p)x_1(n - p - D) + e(n) \tag{2}$$

where $w(p)$ represents the prediction coefficients and $e(n)$ is the prediction error. The coefficients of linear prediction models are estimated in the time domain by minimizing the

[1] W_i and W_i^T correspond to f_L in (1).

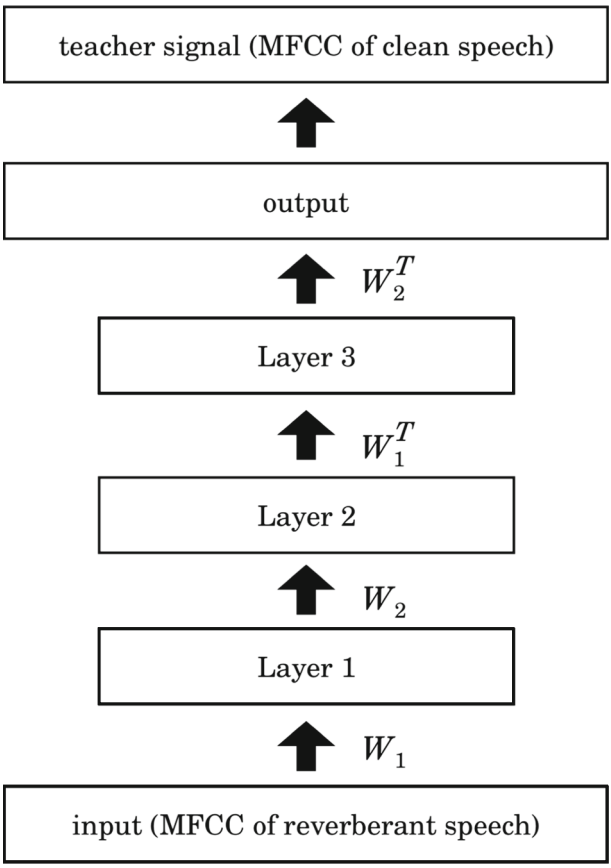


Fig. 1 Topology of a DAE for cepstral-domain dereverberation

mean square energy of $e(n)$. Once the prediction coefficients W have been obtained, they can be applied to the observed signal to estimate the power of the late reverberations as

$$E\{(X_1^T W)^2\} \tag{3}$$

where X_1 and W are the matrix notation of $x_1(n)$ and $w(p)$, respectively. Late reverberations are then converted into the frequency domain and reduced by subsequent spectral subtraction.

3 Bottleneck feature extraction

BFs are generated from DNN that include a hidden layer with fewer units than other hidden layers, just as showed in Fig 2. This hidden layer creates a constriction in the network that forces the information pertinent to classification into a low-dimensional representation. BFs are most commonly used in autoencoders where the neural network is trained to predict

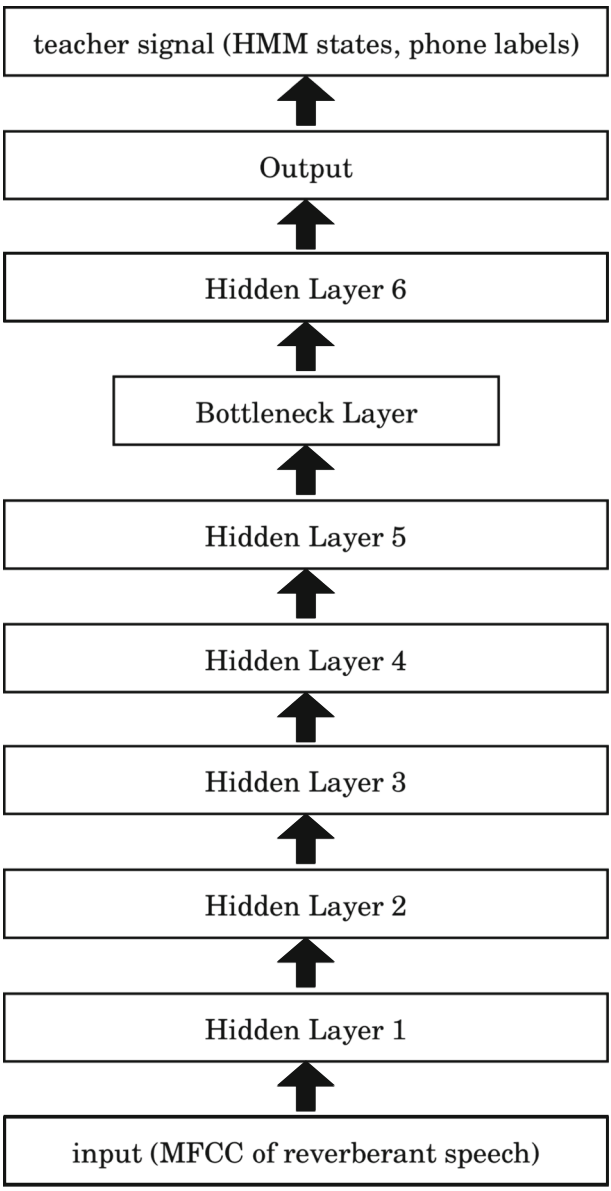


Fig. 2 Topology of bottleneck DNN

the input features themselves [23]. BFs for speech recognition are extracted from DNNs that have been trained to predict phonemes or phoneme states, which are normally generated from an HMM-based recognizer. The inputs to the hidden units of the bottleneck layer are then used as features for further processing, and it is these that are known as Bottleneck Features. Because of the capabilities of DNNs, BFs always represent a nonlinear transformation and discriminative classification pre-processing for speech recognition.

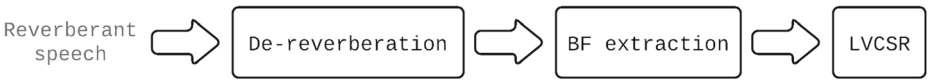


Fig. 3 Cascaded combination diagram

4 Proposed method

It is not always beneficial to combine different approaches. However, some schemes can efficiently integrate a number of techniques. In this paper, to determine whether dereverberation and discriminative BF transforms can be effectively combined, we explore three schemes that consider the characteristics of BF and dereverberations.

4.1 Cascaded combination

As BFs are a kind of feature transform, they are more easily applied to the front-end processing. Thus, a natural idea is to connect them one-by-one in series, as shown in Fig. 3. The dereverberation approach is followed by a BF extractor. This is trained by the dereverberant training features, and is relatively simple and easy to implement. Employed in clean conditions, BF extractors exhibit greatly improved speech recognition performance. It is expected that the BF extractor in the proposed approach can be as effective as in clean conditions, enabling the performance of distant-talking systems to catch up with that of close-talking systems. This is because dereverberation approaches followed by a BF extractor are designed to provide essentially clean features.

4.2 Parallel combination

Considering that BF extractors are a kind of DNN and are capable of dealing with high-dimensional features, additional information should be useful. In this scheme, both dereverberant and reverberant speech are imported into the BF extractor, and the input information is extended with dereverberant information. This paradigm is illustrated in Fig. 4. A number of dereverberation methods have been designed with the aim of mitigating the degradation caused by reverberation. The processed speech should contain more pure information about the relevant utterance, and thus provide correct information to the BF extractor. Unlike the cascaded combination, a parallel combination does not abandon the reverberant speech containing the complete utterance information. Dereverberation approaches cannot accurately eliminate the impact of reverberation, and may distort or damage the key information required for speech recognition. From this point of view, reverberant and dereverberant speech could supply compensatory information to the BF extractor in a parallel combination.

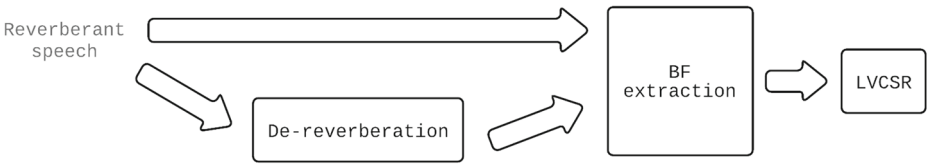


Fig. 4 Parallel combination diagram

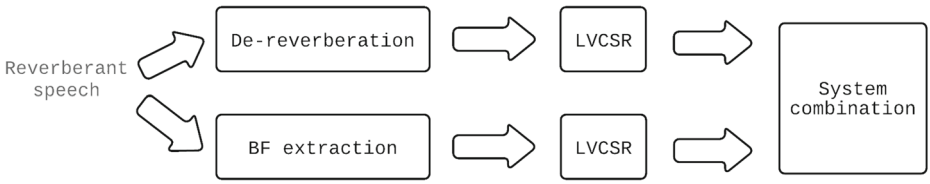


Fig. 5 System combination diagram

4.3 System combination

In fact, dereverberation approaches are mainly intended to obtain clean-like features from a reverberant speech source. However, the BF extractor does not account for any reverberation, but instead attempts to generate a kind of discriminative feature that is easy to distinguish during classification. Because the key points of these two approaches differ, the correct prediction of words and patterns of mistakes are likely to be totally different. From this point of view, it the idea of a confusion network combination (CNC) [2] appears suitable. A CNC should be able to catch the correct predictions from both parts (i.e., the BF extractor and dereverberation approach). In this scheme, a CNC is integrated in the manner shown in Fig. 5. The dereverberation approach and the BF system are first applied separately, and these two separate systems are then combined by the CNC. Thus, correct predictions from each approach can be captured, and system performance will be improved.

5 Experiments

5.1 Experimental setup

5.1.1 Dataset

The data for the experiments were provided by the "REVERB Challenge" [12], and consist of a clean WSJCAM0 training set and a multi-condition (MC) training set. Reverberant speech was generated from the clean WSJCAM0 training data by convoluting the clean utterances with measured room impulse responses and adding recorded background noise. The reverberation times of measured impulse responses ranged from approximately 0.1–0.8 s. The MC training dataset was used to train a DNN-HMM acoustic model, bottleneck DNN, and DAE. To train the DAE, the clean training data were used as target signals. Note that the recording rooms used to obtain the MC training data and the test data were different. It is important to note that the evaluation data consisted of real recordings (RealData) and

Table 1 Details of SimData and RealData datasets

speech	reverberant time			signal-to-noise rate	distance between the microphones	
	room1	room2	room3		near	far
SimData	0.25s	0.5s	0.7s	20dB	50cm	200cm
RealData	0.7s	–	–	–	100cm	250cm

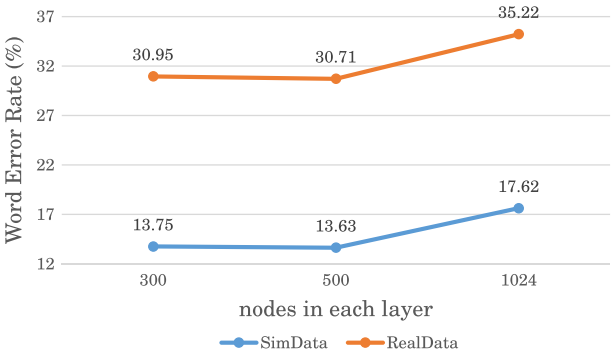


Fig. 6 Performance of BF with respect to the number of nodes in each layer

simulated data (SimData). In particular, the development (Dev.) test set and the final evaluation (Eval.) test set each consist of SimData and RealData. Details of the training and test datasets are presented in Table 1. In the experiments, the Dev. test set was used to optimize the parameters of each component.

5.1.2 Baseline system

A state-of-the-art hybrid DNN-HMM acoustic model was used in all of the experiments. The phone targets used to train the DNN-HMM acoustic model were obtained from a pre-trained GMM-HMM with about 2000 triphone HMM states. The DNN was trained with an initial learning rate of 0.015 and a final learning rate of 0.002. Stochastic mini-batch gradient descent (SGD) was used on the MC training examples to minimize the cross-entropy cost function. As the training dataset only contained about 16.5 h of speech, the DNN was designed to have two hidden layers and 500 nodes in each layer.

5.1.3 Optimal conditions for BF extraction and dereverberation

The baseline BFs were extracted from a bottleneck DNN with an input layer consisting of nine adjacent frames of 13-dimensional MFCCs (i.e., a total of $9 \times 13 = 117$). The bottleneck DNN was trained using the same approach as for the DNN-HMM, but with an

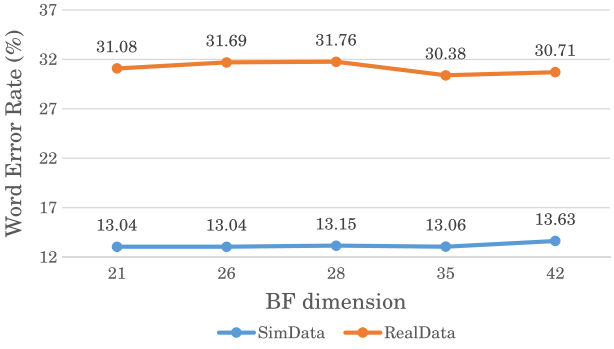


Fig. 7 Performance of BF with respect to the number of BF dimensions

Table 2 WERs (%) of various de-reverberation approaches and BF extraction

	SimData						Ave	RealData		Ave
	Room1		Room2		Room3			Room1		
	near	far	near	far	near	far		near	far	
none	12.96	14.09	14.47	25.43	16.80	31.01	19.13	47.72	45.61	46.67
MSLP	12.64	13.74	13.70	23.48	16.31	28.97	18.14	47.21	46.73	46.97
DAE	12.11	12.84	11.67	22.44	28.78	14.30	17.02	41.94	41.73	41.83
BF	9.40	10.99	11.00	19.32	14.18	25.44	15.05	38.07	38.79	38.43

initial learning rate of 0.005 and a final learning rate of 0.0005. It was constructed to have four hidden layers, and the optimal number of nodes in each hidden layer was determined experimentally. Figure 6 shows the experimental results obtained from the Dev. test set, from which it is clear that 500 nodes should be used for the experimental task. This can be explained by the complex structure of DNNs: a DNN with too many parameters is prone to over-fitting when using a limited set of training data. The dimension of the bottleneck layer was also determined experimentally. The results presented in Fig. 7 show that the word error rates (WERs) for SimData changed little with different BF dimensions. However, the optimal 35-dimensional BFs attained the best performance for RealData and a comparable score for SimData. The other BFs were extracted from a similar bottleneck DNN trained using the dereverberant features corresponding to the dereverberation method (i.e., DAE and MSLP-based dereverberation).

For DAE-based dereverberation, feature vectors from the current frame and the previous eight frames of reverberant speech were used as input. The 39-dimensional MFCCs (which include 12 MFCCs plus their power, Delta, and Delta-Delta coefficients) of the current frame of clean speech were used as teacher signals. An optimized DAE structure with three hidden layers and 1024 units in each layer was used in the experiments. This has been verified by Ueda et al. [22] for the same REVERB challenge. The DAE employed conjugate gradients with a mini-batch size of 256 samples. Pre-training involved 50 epochs with a learning rate of 0.002 for all layers, and the network was then fine-tuned over 100 epochs with a learning rate of 0.1.

The MSLP parameters were taken from the original paper [11], which used a step size of 512 and a filter order of 750 for linear prediction. (We also tested a filter order of 1500, but the performance was worse than for 750.)

Table 3 WERs (%) of cascaded combination integrating dereverberation with BF

	SimData						Ave	RealData		Ave
	Room1		Room2		Room3			Room1		
	near	far	near	far	near	far		near	far	
MSLP	9.54	10.88	10.34	18.69	14.16	24.70	14.72	35.36	38.62	36.99
DAE	10.38	11.76	11.03	20.64	26.20	12.45	15.41	38.33	38.93	38.63

Table 4 WERs (%) of parallel combination integrating dereverberation with BF

	SimData						Ave	RealData		Ave
	Room1		Room2		Room3			Room1		
	near	far	near	far	near	far		near	far	
MSLP	10.25	11.60	11.27	19.16	14.18	24.58	15.17	37.72	39.06	38.39
DAE	10.23	10.86	11.06	20.59	19.07	15.73	14.59	41.20	41.09	41.15

5.2 Experimental results

The proposed methods were evaluated and analyzed on the Eval. test set. The performance of various dereverberation approaches and BF extraction is presented in Table 2. These results indicate that both techniques are efficient for reverberant speech. The last row of Table 2 represents the usability of BF extraction, and this is used as the baseline for comparison.

5.2.1 Results of each scheme

We now analyze the performance of each combination scheme. The results for the cascaded combination (Table 3) show that integrating BF with MSLP decreases the average WERs in most of the test conditions. However, DAE did not perform well, and degraded the speech recognition in most cases. This may be because both DAE and BF are nonlinear transformations in the feature space, and combining similar transforms serially effectively over-tunes the system.

A similar trend can be observed for the parallel combination (Table 4). MSLP is the best choice for parallel combination, which is more effective in real conditions than for simulated conditions. However, the overall performance of parallel combination is worse than that of cascaded combination.

Table 5 presents the results for the system combination, which integrates BF with different dereverberation approaches. Both DAE and MSLP improve the recognition accuracy when applied in combination with BF. Unlike cascaded combination and parallel combination, DAE is more efficient in this system combination. In short, the linear processing of MSLP is better suited to feature space combination, and the nonlinear processing of DAE is more appropriate for back-end combination.

Table 5 WERs (%) of system combination integrating dereverberation with BF

	SimData						Ave	RealData		Ave
	Room1		Room2		Room3			Room1		
	near	far	near	far	near	far		near	far	
MSLP	9.44	10.76	10.63	19.34	14.14	24.27	14.76	37.85	38.15	38.00
DAE	9.18	10.22	10.21	18.73	15.27	15.39	13.17	36.67	36.73	36.70

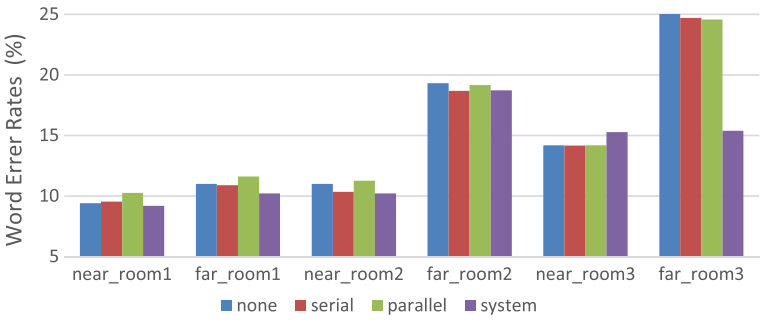


Fig. 8 WERs (%) of each scheme in simulated test conditions

5.2.2 Best scheme for dereverberation and BF

We now determine which scheme is the best choice for combining BF and dereverberation. To obtain a clear view, Fig. 8 shows the WERs of each simulated condition from the best performance of each scheme, and Fig. 9 shows that for each real test condition. Obviously, system combination works better than both the baseline system and the other two combination schemes in most test conditions. However, compared with feature-level combination, the computational cost of the decoding process increases twofold. Comparing the feature-level decoding in the cascaded and parallel combinations, we found that the cascaded combination achieves a small gain in all conditions (both simulated and real conditions). Parallel combination is also comparable to the BF-only system in most conditions. For feature-level combination, the cascaded combination is the best choice.

Inspired by the improved feature-level cascaded combination, we considered whether the system combination could be enhanced using cascaded combination. We applied the system combination between the cascaded combination and BF systems (referred to as "$*$-cascaded + BF", where "$*$" is the dereverberation approach applied in the cascaded combination). The experimental results are presented in Table 6. We found that these new system combinations achieved comparable performance to that of the best system combination of "DAE + BF", and the new-style "DAE-cascaded + BF" achieved even better recognition accuracy with both **SimData** and **RealData**. Note that the best system performance was obtained by applying the system combination between the cascaded combination systems of "DAE-cascaded" and "MSLP-cascaded". This is because cascaded combination-based front-end processing and system combination-based back-end processing achieve complementary improvements in the recognition of distant-talking speech.

Fig. 9 WERs (%) of each scheme in real test conditions

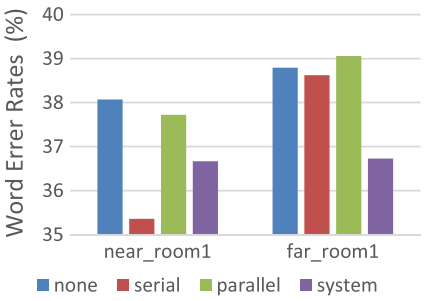


Table 6 WERs (%) of system combination between cascaded combination and BF

	SimData						Ave	RealData		Ave
	Room1		Room2		Room3			Room1		
	near	far	near	far	near	far		near	far	
DAE + BF	9.18	10.22	10.21	18.73	15.27	15.39	13.17	36.67	36.73	36.70
MSLP-cascaded + BF	8.69	10.33	9.92	17.44	13.14	23.20	13.79	34.65	36.12	35.39
DAE-cascaded + BF	8.78	10.16	10.00	18.18	15.34	13.63	12.68	34.17	36.12	35.14
MSLP + DAE	10.57	11.72	11.06	20.82	17.21	15.20	14.43	41.42	40.58	41.00
MSLP-cascaded + DAE-cascaded	8.86	10.25	9.53	17.62	14.93	13.70	12.48	33.92	35.85	34.88

6 Conclusion

In this paper, we have proposed and investigated three schemes that integrate dereverberation approaches and BF extraction, and analyzed the details of each scheme. Based on the experimental results, the most efficient approach appears to be system combination, which is based on a confusion network combination. For the feature-level combination schemes, cascaded combination achieved better performance. In terms of dereverberation approaches, the linear processing of MSLP is better suited to integration with BF in both cascaded and parallel combination. Finally, optimal system performance was obtained by applying the system combination with the cascaded combination systems. We expect this paper to be helpful to other researchers and application engineers.

Acknowledgments This work was supported by JSPS KANKENHI Grant Number 15K16020.

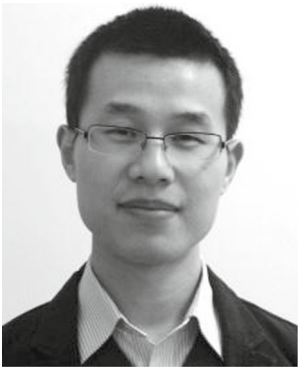
References

1. Abdel-Hamid O, Mohamed A-r, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio, Speech, Lang Process 22(9):1533–1545
2. Evermann G, Woodland PC (2000) Posterior probability decoding, confidence estimation and system combination. In: Proc. Speech Transcr. Work., vol 27. Baltimore
3. Furui S (1981) Cepstral analysis technique for automatic speaker verification. Acoust Speech Signal Process IEEE Trans 29(2):254–272
4. Gesbert D, Duhamel P (1997) Robust blind channel identification and equalization based on multi-step predictors. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol 5, pp 3621–3624
5. Grézl F, Fousek P (2008) Optimizing bottle-neck features for LVCSR. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp 4729–4732
6. Grézl F, Karafiát M, Kontár S, Černocký J (2007) Probabilistic and bottle-neck features for LVCSR of meetings. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc, vol 4. IEEE, pp 757–760
7. Hermansky H, Ellis DPW, Sharma S (2000) Tandem connectionist feature extraction for conventional HMM systems. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. IEEE, pp 1635–1638
8. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag 29(November):82–97

9. Hinton GE, Salakhutdinov RRR (2006) Reducing the dimensionality of data with neural networks. Science 313(July):504–507
10. Ishii T, Komiyama H, Shinozaki Y, Horiuchi T, Kuroiwa S (2013) Reverberant speech recognition based on denoising autoencoder. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp 3512–3516
11. Kinoshita K, Delcroix M, Nakatani T, Miyoshi M (2009) Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans Audio, Speech Lang Process 17:534–545
12. Kinoshita K, Delcroix M, Yoshioka T, Nakatani T, Sehr A, Kellermann W, Maas R (2013) The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Appl. Signal Process. to Audio Acoust. (WASPAA), 2013 IEEE Work. IEEE, pp 1–4
13. Lal P, King S (2013) Cross-lingual automatic speech recognition using tandem features. IEEE Trans Audio, Speech, Lang Process 21:2506–2515
14. Liang L, Renals S (2014) Probabilistic linear discriminant analysis with bottleneck features for speech recognition. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH
15. Liu F-H, Stern RM, Huang X, Acero A (1993) Efficient cepstral normalization for robust speech recognition, Proc. Work. Hum. Lang. Technol. - HLT '93
16. Longbiao W, Kitaoka N, Nakagawa S, Wang L, Kitaoka N, Nakagawa S (2011) Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. IEICE Trans Inf Syst 94(3):659–667
17. Nguyen QB, Gehring J, Muller M, Stuker S, Waibel A (2014) Multilingual shifting deep bottleneck features for low-resource ASR. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp 5607–5611
18. Sainath TN (2012) Auto-encoder bottleneck features using deep belief networks. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp 4153– 4156
19. Sak H, Senior A, Beaufays F (2014) Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv:1402.1128
20. Seide F, Li G, Yu D (2011) Conversational speech transcription using Context-Dependent Deep Neural Networks. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp 437–440
21. Sundermeyer M, Schl R, Ney H (2012) Context-Dependent MLPs for LVCSR : TANDEM, Hybrid or Both? In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH
22. Ueda Y, Wang L, Kai A, Xiao X, Chng E, Li H (2015) Single-channel Dereverberation for Distant-Talking Speech Recognition by Combining Denoising Autoencoder and Temporal Structure Normalization, J. Signal Process. Syst.
23. Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proc. 25th Int. Conf. Mach. Learn. ACM Press, pp 1096–1103
24. 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. J Mach Learn Res 11:3371–3408
25. Wang L, Bo R, Ueda Y, Kai A, Teraoka S, Fukushima T (2014) Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording. In: APSIPA ASC
26. 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 J Adv Signal Process 2012:1–12
27. Xie X, Su R, Liu X, Wang L (2014) Deep neural network bottleneck features for generalized variable parameter HMMs. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. ISCA, pp 2739–2743
28. Yamada T, Wang L, Kai A (2013) Improvement of distant-talking speaker identification using bottleneck features of DNN. In: INTERSPEECH, pp 3661–3664
29. Yu D, Deng L, Dahl GE (2010) Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. In: NIPS Work. Deep Learn. Unsupervised Featur. Learn.
30. Yu D, Seltzer ML (2011) Improved Bottleneck Features Using Pretrained Deep Neural Networks. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp 237–240
31. Zhang Z, Wang L, Kai A (2014) Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation. EURASIP J Audio, Speech, Music Process 2014(1):15
32. Zhang Z, Wang L, Kai A, Yamada T, Li W, Iwahashi M (2015) Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. EURASIP J Audio, Speech, Music Process 2015(1):12

Bo Ren is a Master Student at Nagaoka University of Technology. He received the B.Sc. degrees in information engineering from the Northwest University, Xi'an, China, in 2010. His current research mainly focuses on the front-end processing for de-reverberation for speech recognition.

Longbiao Wang received his Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2008. He was an assistant professor in the faculty of Engineering at Shizuoka University, Japan from April 2008 to September 2012. Since October 2012 he has been an associate professor at Nagaoka University of Technology, Japan. His research interests include robust speech recognition and speaker recognition. He received the "Chinese Government Award for Outstanding Self-financed Students Abroad" in 2008. He is a member of IEEE, the Institute of Electronics, ISCA, APSIPA, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).



Liang Lu is a Research Associate at the University of Edinburgh. He received the B.Sc. and M.Sc. degrees in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2009, respectively. He then obtained his Ph.D. degree from the University of Edinburgh in 2013. His current research mainly focuses on noise robustness, cross-lingual/multilingual acoustic modeling and pronunciation modeling for speech recognition.

Yuma Ueda received his B.E. degree from Shizuoka University in 2014. He entered graduate school of engineering, Shizuoka University. His research interests distant-talking speech recognition.

Atsuhiko Kai received his B.E., M.E. and Dr.Eng. degrees from Toyohashi University of Technology, Toyohashi, Japan, in 1991, 1993 and 1996, respectively. He joined the faculty of Toyohashi University of Technology in 1996 as a Research Associate. In 1999, he joined the Department of Systems Engineering, Shizuoka University, Hamamatsu, Japan, as a Lecturer and is currently an Associate Professor in the Department of Mathematical and Systems Engineering, Shizuoka University. His research interests include speech processing, spoken language processing with a focus on speech recognition and pattern recognition.