Reverberant Speech Recognition Based on Denoising Autoencoder

Takaaki Ishii[1], *Hiroki Komiyama*[1], *Takahiro Shinozaki*[2], *Yasuo Horiuchi*[1], *Shingo Kuroiwa*[1]

[1]Division of Information Sciences, Graduate School of Advanced Integration Science,
Chiba University, Chiba, Japan
[2]Department of Information Processing,
Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Tokyo, Japan

Abstract

Denoising autoencoder is applied to reverberant speech recognition as a noise robust front-end to reconstruct clean speech spectrum from noisy input. In order to capture context effects of speech sounds, a window of multiple short-windowed spectral frames are concatenated to form a single input vector. Additionally, a combination of short and long-term spectra is investigated to properly handle long impulse response of reverberation while keeping necessary time resolution for speech recognition. Experiments are performed using the CENSREC-4 dataset that is designed as an evaluation framework for distant-talking speech recognition. Experimental results show that the proposed denoising autoencoder based front-end using the short-windowed spectra gives better results than conventional methods. By combining the long-term spectra, further improvement is obtained. The recognition accuracy by the proposed method using the short and long-term spectra is 97.0% for the open condition test set of the dataset, whereas it is 87.8% when a multi-condition training based baseline is used. As a supplemental experiment, large vocabulary speech recognition is also performed and the effectiveness of the proposed method has been confirmed.

Index Terms: Denoising autoencoder, reverberant speech recognition, restricted Boltzmann machine, distant-talking speech recognition, CENSREC-4

1. Introduction

One of the advantages of speech based communication is that a speaker and a listener can be separated since the vibration of air transmits for some distance. In fact, human listeners easily recognize utterances uttered in a few meters away. However, the recognition performance of today's automatic recognizers using a distant talking microphone is much lower than the systems using a close talking microphone, even in a quiet condition. This is because the effect of reverberation is significant when a distant talking microphone is used and the direct speech signal is overlapped by reflected signals with various delay timings. The delay is often longer than the length of analysis window used in a feature extractor in a recognition system. As the result, the spectral patterns are blurred that contain acoustic clues necessary for speech recognition. This is especially problematic for applications such as hands-free recognition systems and dialogue robots.

To address the problem, various approaches have been proposed using multiple and single microphones. While using multiple microphones has a potential for powerful performance [1], single microphone approaches have an advantage for usability from an application point of view. As the single microphone approaches, cepstral mean normalization is known to be useful when the distortion has short impulse response [2]. For longer reverberation time, a special scheme has been proposed for obtaining delta and delta-delta features that are less affected by long decay time [3]. However, obtaining good recognition performance using a single microphone is still a difficult problem.

An autoencoder is a multilayer neural network whose output is a reconstruction of the input vector through a small central layer. The purpose of the autoencoder is to convert high dimensional data to low dimensional codes expressed by the central layer [4, 5]. By extending the autoencoder, a denoising autoencoder has been proposed in image processing as an effective denoising method [6]. With the denoising autoencoder, the network is trained so that clean data is reconstructed from a noisy version of it with the hope that the central layer finds codes that are less susceptible to the noise. Recently, denoising autoencoders have been applied to noise robust speech recognition where the network is trained so as to remove additive noise. The evaluations have been performed in closed and open noise conditions, and promising results have been shown [7, 8].

In this paper, we apply a denoising autoencoder to reverberant speech recognition so that clean speech spectrum is reconstructed from reverberant speech. Since a denoising autoencoder is capable of treating a large dimensional input vector, it is expected to be especially useful for dereverberation of speech utterances modeling both sub-phone level rapid changes of speech spectrum and long durational properties of reverberation. In the proposed method, in order to capture long durational effects, a window of multiple short-windowed spectral frames are concatenated to form a composite frame. Additionally, long-windowed spectral frames are incorporated so that long impulse response is properly handled. Digit recognition experiments are performed using the Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments (CENSREC-4) [9] dataset. Moreover, as a supplemental experiment, large vocabulary recognition is performed using the Japanese Newspaper Article Sentences (JNAS) [10] dataset.

The organization of the rest of this paper is as follows. In Section 2, autoencoders and denoising autoencoders are briefly reviewed. In Section 3, the proposed dereverberation methods using denoising autoencoders are explained. Experiments using CENSREC-4 dataset are shown in Section 4 and an experiment using JNAS is described in Section 5. Finally, conclusions are given in Section 6.

2. Autoencoder and denoising autoencoder

An autoencoder is a multilayer neural network with a small central layer. Since it is difficult to directly optimize weights in a deep autoencoder having many layers, an initialization step called pre-training is conducted.

In the pre-training, first an "encoder" network is trained layer by layer as a stack of restricted Boltzmann machines (RBMs). RBM is a bipartite graph in which visible units representing observations are connected to hidden units. There is no direct connection between the visible units and between the hidden units. An energy function is defined over the parameters associated with the network and the values of the hidden and visible units. A joint distribution of the hidden and visible units is given as a Boltzmann distribution specified by the energy function like as in a physical system in thermal equilibrium. Because it is not feasible to learn large RBM by exact maximum-likelihood method, the contrastive divergence method is used as an approximate training [11]. A simplest type of RBM is the binary RBM in which both the hidden and visible units are binary and stochastic. For real-valued input, a Gaussian-binary RBM can be used in which the hidden units are binary and the visible units are real valued [12].

After the encoder network is obtained, a "decoder" network is made by turning the encoder network upside down making the input layer as output and the output layer as input. Finally, an initial network of an autoencoder is obtained by gluing the output of the encoder network to the input of the decoder network.

Given the initial autoencoder, backpropagation is applied to adjust the parameters slightly setting the same data at both the input and output layers. The training of denoising autoencoder is mostly the same as that of the autoencoder except that the backpropagation is performed so that clean data is reconstructed from noisy version of it. After the training, the network is used to reconstruct clean data from noisy input.

3. Proposed method

We apply the denoising autoencoder for dereverberation purpose as a front-end of a speech recognizer. The input of the denoising autoencoder is a window of spectral frames of reverberant speech and the output is a window of spectral frames of clean speech. By applying a denoising autoencoder to a time series of overlapping windows of frames of reverberant speech, a time series of overlapping windows of frames of estimated clean speech spectrum is obtained. To obtain a time series of unit spectral frames from the output overlapping windows of the frames, an average is computed at each time frame as shown in Figure 1. Features for back-end processing are computed from the obtained frame of a spectral vector. For example, Mel Frequency Cepstrum Coefficients (MFCC) [13] are obtained by applying mel filter banking and discrete cosine transform to the averaged frame.

To obtain necessary time resolution for speech recognition, short width window ($\approx$ 25 ms) is used for spectrum analysis. However, that window width is not enough to capture long impulse response of reverberation. Therefore, use of long window (ex. 500 ms) is additionally investigated with the same frame shift width as that of the short-term analysis. For the spectral vector from the long window, mel filter banking is applied and the dimensionality is reduced. Then, the long-term spectral frames with the reduced dimension is treated similarly as the short-term spectrum, and a composite frame is formed by concatenating a window of the frames. The window of frames of the short-term spectrum vectors and the one of the long-term spectrum vectors are concatenated to form a single large composite vector. A denoising autoencoder is trained using the composite vectors as the input and output. However, when features for speech recognition like MFCC are made from the output of the denoising autoencoder, the dimensions of the vectors that correspond to the long-term spectrum are not needed. Therefor, these dimensions in the output side of the denoising autoencoder are simply removed after the pre-training process.

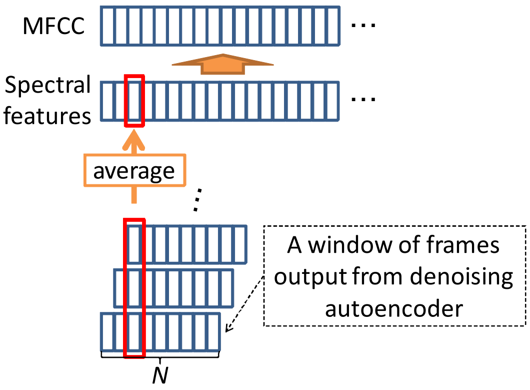

Figure 1: Procedure to obtain a frame from overlapping windows of frames that are output from a denoising autoencoder.

4. Digit recognition experiments using CENSREC-4 dataset

4.1. Experimental setup

The CENSREC-4 dataset contains a clean training set that consists of 8440 connected digit utterances. Multi-condition training set is defined based on the clean set where it is divided to four parts and one of four types of impulse responses is convoluted respectively. The speech data is sampled at 16 kHz.

Five layer denoising autoencoders were used in the experiments. Gaussian-binary RBM was used as the first layer RBM, and binary-binary RBM was used as the second. For the pre-training, 2110 utterances were randomly selected from the clean training set and 2110 utterances were randomly selected from the multi-condition training set. In total, 4220 utterances were used to train a denoising autoencoder. The mini-batch size was 100 and the number of mini-batches was 2730. The number of iterations was 100 for the first and second layers, respectively. For the fine-tuning, the same utterances as the pre-training were used at the input side of the denoising autoencoder and the corresponding clean utterances were used at the output side. The backpropagation was based on the conjugate gradient optimization [14]. Mean square error was used as the objective function. The mini-batch size was 1000 and the number of mini-batches was 273.

The short-term window width was 25 ms and the long-term window was 500 ms. The frame shift was 10 ms. The mel filter bank size to reduce the long-term spectral vector size was 24. In addition to the power spectrum, log energy was appended that was estimated from speech window without the hamming window weighting. The total dimension of the short-term window based composite vector was $(256 + 1) \times N$ and

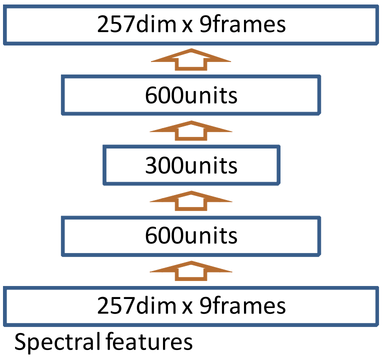

Figure 2: Default network topology of denoising autoencoder.

that for the vector with both short and long term window was $(256+1+24+1) \times N$ where N is the number of windowed frames.

Once the denoising autoencoder was trained, all the utterances in the clean training set was fed into it and Hidden Markov Models (HMMs) were trained from the output. The features used for the HMM was MFCC (12 dim) and log energy, their delta, and delta-delta. The total dimension was 39. For the evaluation, test set A and B were used that were defined in the dataset. The test set A was closed for the reverberation environments in the training set and the test set B was open.

In the following experiments, unless described explicitly, a "default" configuration was used for denoising autoencoders where nine short-term spectral frames were used in concatenation (i.e. N=9). The first (i.e., input) layer size of the autoencoder was 2313, which corresponded to the feature dimension. The second layer size was 600 and the third (i.e., the center) layer size was 300. The pre-training was applied and the number of iterations for the backpropagation was 30. Figure 2 shows the network with the default configuration. This configuration was decided based on preliminary experiments considering both recognition performance and available computational resources. While the configuration was mostly optimal for the tested conditions, we observed a small improvement ($\approx 0.3\%$) when the hidden layer sizes were doubled, i.e., when the second layer size was 1200 and the third layer size was 600.

4.2. Results

Figure 3 shows the relationship between the number of iterations of the backpropagation and mean square errors between clean and dereverberated spectral vectors output from the denoising autoencoders. The errors were evaluated for the training set and for the test set A and B. As expected, test set B had larger errors than test set A since it was open to the training set. The reduction of the errors were mostly saturated at 10 iterations but further slight improvements were observed for the training set and for the test set A.

Table 1 shows total training CPU costs and the effect of pre-training for recognition accuracy of the test set B. While pre-training occupied a significant part of the training cost, it gave nice improvement in recognition accuracy. Running 10 iterations for backpropagation with the pre-training gave better performance than running 30 iterations for backpropagation without the pre-training. The best result was obtained when 30 iterations was performed for the backpropagation after the pre-training.

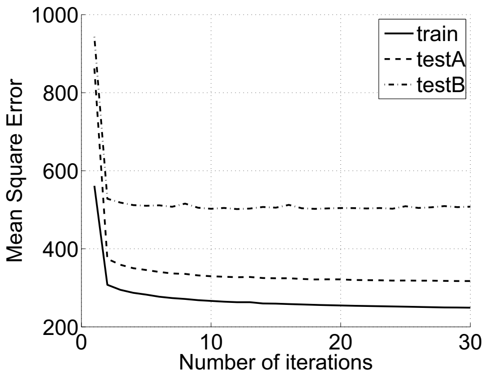

Figure 3: Relationship between the number of iterations of the backpropagation and mean square errors for the training set and for the test set A and B. The errors are obtained between clean spectrum and dereverberated spectrum that are output from the denoising autoencoders.

Table 1: Effect of pre-training for digit recognition accuracy and total training CPU costs. Number of iterations for backpropagation was 0, 10 and 30. Iteration 0 indicates the CPU cost was spent only for pre-training. The test set B was used.

CPU time (Acc)	iter=0	iter=10	iter=30
W/o pre-training	0h (-)	21.3h (85.9)	57.3h (92.5)
W pre-training	17.3h (-)	41.5h (95.8)	72.3h (96.4)

Table 2 shows the relationship between the window size (N) for concatenating spectral frames and recognition accuracy. Larger context is modeled as N increases, but the input layer size also increases. It looks the optimal value for N is around 11.

Table 3 lists recognition accuracies of conventional and proposed methods. In the table, the baseline is the results of normal MFCC (with the energy, delta and delta delta) based system, and CMN is the results with the cepstral mean normalization. The accuracies of these systems are cited from the paper by the group who made the the CENSREC-4 dataset [9]. HDelta is the hybrid delta method proposed by IBM and the accuracies are cited from their paper [3]. For these results, clean indicates clean-condition training and Multi indicates multi-condition training. DAE is the denoising autoencoder results based on the short term spectra and short+long term spectra. The baseline accuracies with the multi-condition training were 92.9% for the closed test set A and 87.8% for the open test set B. Among the conventional methods, the HDelta method gave the highest result where the accuracy for the test set A was 95.7 and that for the test set B was 96.4. The proposed denoising autoencoder based dereverberation with the short term spectral frames gave better results than the HDelta method. The best results were obtained by the proposed denoising autoencoder using the short and long-term spectra, where the recognition accuracies were 98.4% for the test set A and 97.0% for the test set B. The relative error reduction from the HDelta method were 62.8% and 43.2% for the test set A and B, respectively.

Table 2: Window size (N) for spectral frame concatenation and recognition accuracy.

N	testA	testB
5	97.3	95.8
7	97.8	96.3
9	97.9	96.4
11	98.1	96.7
13	97.9	96.1

Table 3: Recognition accuracy of conventional and proposed methods. Baseline is a normal MFCC based recognition, CMN is with cepstral mean normalization, HDelta is the hybrid delta method proposed by IBM, and DAE is the proposed denoising autoencoder based dereverberation. Clean indicates clean-condition training and Multi indicates multi-condition training. For DAE, short means it is based on short-term spectra and long means it is based on short and long term spectra.

Method	testA	testB
baseline(Clean)	83.8	82.8
baseline(Multi)	92.9	87.8
CMN(Clean)	86.5	88.6
CMN(Multi)	91.8	89.7
HDelta(Multi)	95.7	94.7
DAE(short)	97.9	96.4
DAE(short+long)	98.4	97.0

5. Word recognition experiment using JNAS dataset

5.1. Experimental setup

As a supplemental experiment, large vocabulary recognition was performed using the JNAS dataset that contains read speech utterances of news articles. We selected recordings in the JNAS dataset given by four male speakers and four female speakers as a test set, which amounted 110 minutes in total. A training set was defined as 174 recordings that amounted 35 hours in total. There was no overlapping speakers in the training and test sets. For the experiment, reverberant speech was generated by applying convolution using the impulse responses provided in the CENSREC-4 dataset to the clean speech in JNAS recoded by headset microphones. For the training set, four types of impulse responses were randomly convoluted to the clean utterances. For the test set, four impulse responses that were different from the training set were convoluted.

The network configuration of the denoising autoencoder was the same as the default configuration in the previous section. Due to time constraint for the experiment, only two hours of subset of the training set was used for the training of the denoising autoencoder. After the denoising autoencoder was trained, it was applied to the whole training set and the test set.

The features for HMM were 12 MFCCs with CMN, energy, their delta, and delta delta. Tied state HMMs with 1000 states and 32 mixtures per state were trained using all the training set. Language model was a trigram with 18k vocabulary trained from the JNAS dataset. For recognition, T^3 WFST based decoder [15] was used.

Table 4: Word recognition accuracy evaluated using the JNAS dataset.

System	WACC (%)
baseline(CMN)	61.4
DAE(short)	65.2

5.2. Results

Table 4 shows word accuracies of the experiment using the JNAS dataset. The baseline is the results using the reverberant speech where the HMM was trained using the training set with the reverberations. In this condition, the word accuracy was 61.4%. DAE is the result using the proposed denoising autoencoder based dereverberation and the word accuracy was 65.2%. Although the denoising autoencoder was not tuned for this condition and the amount of data used to train denoising autoencoder was limited, it gave better performance than the baseline.

6. Conclusions

In this study, we have proposed and investigated to apply denoising autoencoders to reconstruct clean speech spectrum from reverberant speech. In order to capture context effects, a window of multiple short-windowed spectral frames are concatenated to form a single composite frame and it was used as the input for the denoising autoencoder. Moreover, to properly handle long impulse response of reverberation, we proposed to combine short and long-term spectra to form an input for the denoising autoencoder. Experiments were performed using the CENSREC-4 dataset that was designed as an evaluation framework for distant-talking speech recognition. Experimental results showed that the proposed denoising autoencoder based dereverberation gave the best results both for the closed test set A and the open test set B. Compared to only using the short-term spectrum, some more improvements were obtained by combining the short and long term spectra. The recognition accuracies by the denoising autoencoder using the short and long term spectra were 98.4% for the test set A and 97.0% for the test set B, where the baseline accuracies with the multi-condition training were 92.9% and 87.8%, respectively. Additionally, a large vocabulary speech recognition experiment was performed and the effectiveness of the proposed method has been confirmed. As the future work, we are planning to increase the amount of training data to train denoising autoencoders. Extending the framework to use multiple microphones and to multi-modal recognition is also interesting.

7. Acknowledgment

This research has been supported by a grant from the Telecommunications Advancement Foundation (TAF). Part of this research has also been supported by KAKENHI (25280058) and KAKENHI(21300060).

8. References

[1] K. Kumatani, J. McDonough, and B. Raj, “Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 127–140, 2012.

[2] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-

term multiple-step linear prediction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 534 –545, may 2009.

[3] O. Ichikawa, T. Fukuda, and M. Nishimura, "Dynamic features in the linear-logarithmic hybrid domain for automatic speech recognition in a reverberant environment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 816–823, 2010.

[4] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep autoencoder," in *Proc. Interspeech*, 2010, pp. 1692–1695.

[6] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[7] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Proceedings of INTERSPEECH*, 2012.

[8] X. Lu and H. K. S. Matsuda, C. Hori, "Speech restoration based on deep learning autoencoder with layer-wised pretraining," in *the International Speech Communication Association*, 2012.

[9] T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura, "Evaluation framework for distant-talking speech recognition under reverberant environments —newest part of the censrec series—," *Proc. LREC'08*, May 2008.

[10] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoust Soc Jpn E*, vol. 20, no. 3, pp. 199–206, 1999.

[11] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, p. 2002, 2000.

[12] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transaction on Acoustic Speech and Singal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[14] M. R. Hestenes and E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, Dec. 1952.

[15] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The titech large vocabulary wfst speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.