論文 / 著書情報
Article / Book Information

Title	Multi-Task Autoencoder for Noise-Robust Speech Recognition
Authors	Haoyi Zhang, Conggui Liu, Nakamasa Inoue, Koichi Shinoda
Citation	Proc. ICASSP, , , pp. 5599-5603
Pub. date	2018, 4
Copyright	(c) 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	https://doi.org/10.1109/ICASSP.2018.8461446
Note	This file is author (final) version.

MULTI-TASK AUTOENCODER FOR NOISE-ROBUST SPEECH RECOGNITION

Haoyi Zhang, Conggui Liu, Nakamasa Inoue, Koichi Shinoda

Tokyo Institute of Technology, Japan

ABSTRACT

For speech recognition in noisy environments, we propose a multi-task autoencoder which estimates not only clean speech features but also noise features from noisy speech. We introduce the *deSpeeching* autoencoder, which excludes speech signals from noisy speech, and combine it with the conventional denoising autoencoder to form a unified multi-task autoencoder (MTAE). We evaluate it using the Aurora 2 dataset and CHIME 3 dataset. It reduced WER by 15.7% from the conventional denoising autoencoder in the Aurora 2 test set A.

Index Terms— Denoising autoencoder, deSpeeching autoencoder, robust speech recognition

1. INTRODUCTION

Speech recognition is utilized in our everyday life, thanks to its rapid development. However, the existence of noise degrades its performance drastically. Many methods have been proposed to solve this problem. Their examples include vector Taylor series (VTS) [1, 2, 3] and SPLICE [4].

Recently deep learning has been successfully applied to robust speech recognition, and the denoising autoencoder (DAE) has been one of its most effective methods [5]. Using a sufficient amount of data, it can learn non-linear mapping functions from noisy-speech features to clean-speech features. There have been many efforts dedicated to improving it. A deep denoising autoencoder (DDAE) with five hidden layers employs a pre-training method of the Restricted Boltzmann Machine (RBM) [6]. A recurrent layer was inserted to learn the temporal context [7]. An ensemble model of the DAEs enhances the model's generalization ability [8].

The noisy-speech features consist of features from both clean-speech signals and noise signals. What the traditional denoising autoencoder does is to directly model the mapping from noisy-speech features to clean-speech features. It can easily reduce noise signals and get clean-speech features when speech features are dominant in noisy-speech features. In the case of low signal-to-noise ratios (SNRs), however, it performs poorly because noise significantly contaminates the resulting speech signals.

In machine learning, multi-task learning (MTL), which simultaneously trains more than one correlated tasks, is one of the most promising approaches for increasing a model's generation ability [9]. The power of MTL is rooted in its hidden-units-sharing structure, which leads to robust model estimation for all the correlated tasks. MTL has been successfully utilized for various tasks such as low-resource language speech recognition [10] and joint training of triphones and trigraphmes [11].

In this study, we propose a multi-task autoencoder (MTAE) which jointly trains denoising and deSpeeching autoencoders. Here the deSpeeching autoencoder estimates noise features from noisy-speech features. We expect that the deSpeeching autoencoder brings better clean-speech feature prediction power. Our method is particularly effective when SNR is low and noises are non-stationary. To the best of our knowledge, it is the first work to use a neural network to estimate noise audio features and to use it to enhance the estimation of speech audio features. Its effectiveness is validated by speech recognition experiments on the Aurora 2 and the CHIME 3 database. Lee proposed MTL for estimating both clean-speech features and noise features [12]. They employed a traditional DNN structure where the inputs are noisy-speech and estimated noise. On the contrary, our multi-task autoencoder only takes noisy-speech as the input, and also it employs a triangularly shared-units network structure, which was proven to be significantly efficient.

2. RELATED STUDIES

2.1. Robust speech recognition

SPLICE predicts clean-speech from noisy-speech by learning from stereo data [4]. In spectral subtraction, noise spectrum is subtracted from noisy spectrum to obtain clean-speech spectrum [13]. In addition, researchers proposed noise-robust features, such as the ETSI-AFE [14], the PNCC feature [15] and the NMCC feature [16].

2.2. Denoising autoencoder (DAE)

A denoising autoencoder is a variant of an autoencoder (AE). AE consists of two components: an encoder and a decoder. The encoder converts an input $\boldsymbol{x}$ into a representation $h(\boldsymbol{x})$, and the decoder converts $h(\boldsymbol{x})$ into a reconstructed input $g(h(\boldsymbol{x}))$. By minimizing the Mean Squared Error (MSE) between an input $\boldsymbol{x}$ and its reconstructed input $g(h(\boldsymbol{x}))$, AE

learns a good representation of the input. For making representations robust against partial corrupted input data, a denoising autoencoder (DAE) is proposed [17]. Instead of directly using $\boldsymbol{x}$ as an input, denoising autoencoders add noise, such as Guassian noise and random mask noise, to it to make a corrupted input $\tilde{\boldsymbol{x}}$. DAE is trained with MSE between the input $\boldsymbol{x}$ and the reconstructed input $g(h(\tilde{\boldsymbol{x}}))$.

In the field of speech recognition, DAEs has been used for speech enhancement [5]. Let $\boldsymbol{x}$ be an input noisy-speech feature, and $\boldsymbol{y}$ be a prediction of a clean-speech feature. Then the output $\boldsymbol{y}$ is calculated as follows:

$$\boldsymbol{y} = \boldsymbol{W}^1\boldsymbol{h}^0 + \boldsymbol{b}^1, \quad (1)$$

$$\boldsymbol{h}^0 = \sigma(\boldsymbol{W}^0\boldsymbol{x} + \boldsymbol{b}^0), \quad (2)$$

where $\boldsymbol{W}^0$, $\boldsymbol{W}^1$ are learnable weights and $\boldsymbol{b}^0$, $\boldsymbol{b}^1$ are learnable biases, $\boldsymbol{h}^0$ is the output of the hidden layer, and σ is a non-linear activation function. Normally σ is set as a sigmoid function. DAE uses MSE between the network's output $\boldsymbol{y}$ and label $\boldsymbol{y}'$ as a loss function:

$$L(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{y}' - \boldsymbol{y}||^2. \quad (3)$$

A single-layer denoising autoencoder can be extended to a deep denoising autoencoder (DDAE) by using an RBM pre-training method [6]; RBM provides the initialized weights and biases that represent training data's distribution, preventing the over-fitting problem from happening. DDAE generally over-performs DAE.

2.3. Multitask learning (MTL)

The joint training of correlated tasks often improves the performance of all tasks [9]. Multi-task learning (MTL) provides a way to combine several correlated tasks. In MTL of deep nets, correlated tasks share the same input and some parts of hidden units. Each task also possesses its own hidden units for learning its unique knowledge.

The feature sharing structure enables information-sharing between correlated tasks during training, which is the essence of MTL. MTL is effective in the following three reasons. First, it creates more general representations than single task learning (STL). Correlated tasks in MTL focus on different aspects of input data, which makes hidden layers learn all-round features from them. Second, it impedes the overfitting in each task; extra tasks act as an inductive bias that prevents the model from adjusting the original task too well. Third, it improves attribute selection. When training data is limited or contains noise, it is hard for STL to pick out relevant features.

MTL has been widely used in many tasks. For example, in multi-lingual speech recognition, an MTL-based DNN is built where each language has its own output layers while sharing the input layer and several hidden layers [18, 19]. Also, the joint training of triphone and trigraphmes was proposed [11].

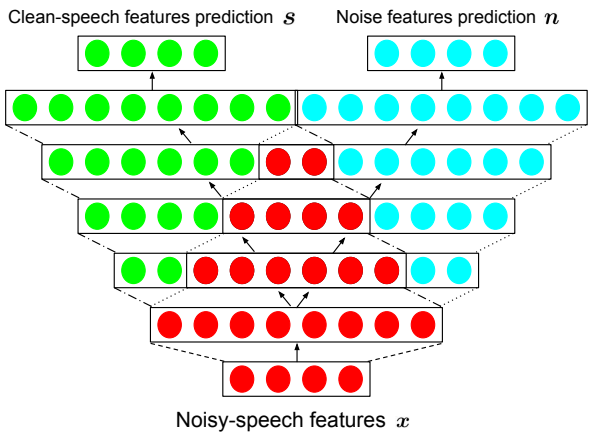

Fig. 1. Multi-task autoencoder: The left side is a denoising autoencoder and the right side is a deSpeeching autoencoder. The green nodes in the upper-left belong to the denoising autoencoder, the blue nodes in the upper-right belong to the deSpeeching autoencoder, and the red nodes in the middle are shared by those two autoencoders.

3. PROPOSED METHOD

3.1. DeSpeeching autoencoder

First, we build a deSpeeching autoencoder, a neural network that predicts noise features given noisy-speech features. Its input is the noisy-speech feature and the output is the noise feature. The training procedure of the deSpeeching autoencoder is similar to that of DAE: Use MSE loss between predicted noise features and noise features to learn the mapping function from noisy-speech features to noise features. The deSpeeching autoencoder can be extended to the deep deSpeeching autoencoder by using the RBM-based pre-training.

3.2. Multi-task autoencoder (MTAE)

Next, we combine the two autoencoders, the denoising autoencoder and the deSpeeching autoencoder, to form a unified multi-task autoencoder (MTAE). Here, MTAE has a triangularly shaped shared-units as shown in Fig. 1. In terms of the number of units in each hidden layer, we first determine the first hidden layer to be n and the last hidden layer to be $2n$. Intuitively, we increase hidden layer's units linearly. When the number of hidden layer is L, the l^{th} hidden layer has $\left\lceil \frac{n(L+l-2)}{L-1} \right\rceil$ hidden units, consisting of $\left\lceil \frac{n(L-l)}{L-1} \right\rceil$ shared hidden units, $\left\lceil \frac{n(l-1)}{L-1} \right\rceil$ hidden units exclusive to the denoising autoencoder and $\left\lceil \frac{n(l-1)}{L-1} \right\rceil$ hidden units exclusive to the deSpeeching autoencoder. The green nodes connect only to the green and red nodes of the $(l+1)^{th}$ layer, the blue nodes connect only to the green and red nodes of the $(l+1)^{th}$ layer

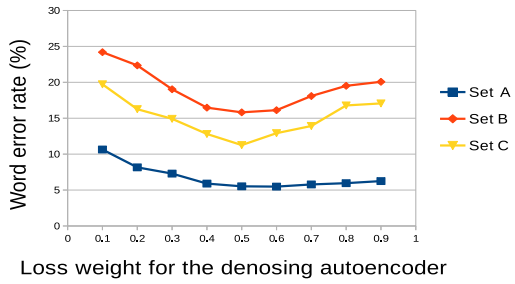

Fig. 2. The performance of MTAE varying the loss weights in Aurora 2 dataset under clean training.

while the red nodes connect to all the nodes of the $(l+1)^{th}$ layer.

The training procedure considers the loss from both of the tasks. The model parameters are learned via the following loss function:

$$L(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{n}) = c||\boldsymbol{s}' - \boldsymbol{s}||^2 + (1-c)||\boldsymbol{n}' - \boldsymbol{n}||^2, \quad (4)$$

where $\boldsymbol{s}$, $\boldsymbol{s}'$ are the predicted clean-speech feature and the original clean-speech feature respectively, $\boldsymbol{n}$, $\boldsymbol{n}'$ are the predicted noise feature and the original noise feature respectively, and c is a weight between the two tasks.

The denoising task and the deSpeeching task collaborate together and provide mutual information. For the conventional DAE, directly modeling the mapping function from noisy-speech features to clean-speech features does not work well when SNR is low, since large noise signals conceal clean-speech features. However, in the case of multi-task autoencoder, the prediction of clean-speech features takes noise features' estimation as prior information by triangularly shaped shared-units. If the denoising task gets accumulated information from shared units, it may somehow figure out noise features, which makes it easy to subtract noise features and predict clean-speech features.

4. EXPERIMENTS

4.1. Experimental settings

We evaluated our proposed method, MTAE, on the Aurora 2 database and the CHIME 3 database. The Aurora 2 database is a continuous digits recognition corpus under noisy environments [20]. The training data set provides two modes: clean condition training and multi-condition training. Each mode contains 8,440 utterances. The test data consists of 3 subsets: Set A, Set B, and Set C. Set A uses the same noise as the multi-condition training set, Set B uses four different types of noise (restaurant, street, airport and train station), Set C uses two types of noise (suburban train and street) with different channel conditions. The CHIME 3 database is a speech recognition corpus under noisy environments based on WSJ corpus [21]. It consists of 7,138 artificial noisy utterances that achieved by mixing clean WSJ0 utterances with noise background (cafe, bus, street, pedestrian area). For evaluation, we use the real noisy data ("et05_real"). We carried out the Aurora 2 and the CHIME 3 speech recognition system by using the KALDI toolkit [22].

We use the MFCC feature of 13 dimensions. To include the context information, we use continuous 11 frames as the input. The input dimension is 143. We use 5 for L and 1024 for n in Subsection 3.2. Therefore, (a, b, c) from the 1st hidden layer to the 5th hidden layer are $(0, 1024, 0)$, $(256, 768, 256)$, $(512, 512, 512)$, $(768, 256, 768)$ and $(1024, 0, 1024)$, respectively.

We train our model using multi-condition training data. In Aurora 2 experiments, 8,440 utterances are divided into 7,806 utterances for training and 634 utterances for validation. In CHIME 3 experiments, we use 6,600 utterances for training and 528 utterances for validation.

Table 1. WER (%) results of MTAE on Aurora 2 under clean training .

SNR	Set A	Set B	Set C	Avg.
Clean	0.68	0.68	0.71	0.69
20dB	1.01	1.25	1.14	1.13
15dB	1.34	2.45	2.16	1.98
10dB	2.29	7.34	5.37	5.00
5dB	5.04	20.10	10.86	12.00
0dB	17.90	47.96	36.78	34.21
-5dB	53.57	73.08	68.32	54.99
Avg.	5.52	15.82	11.26	10.87

Table 2. WER (%) of MTAE and the other methods on Aurora 2 under clean training (averaged over 0-20 dB).

Method	Set A	Set B	Set C	Avg.
MFCC [20]	38.66	44.26	33.86	38.93
ETSI-AFE [14]	12.19	12.91	14.23	13.11
NMCC [16]	16.77	14.91	17.50	16.39
DDAE [6]	6.39	20.44	17.20	14.68
MTAE	5.52	15.82	11.26	10.87

4.2. Experimental results

In Fig. 2, we show the results using different loss weights c in Eq.4, for the denoising and deSpeeching autoencoder. The lowest WERs for all test sets appear when $c = 0.5$. We use 0.5 for c below.

Table 1 shows the Aurora2's detailed results of the proposed MTAE when the speech recognition system is trained by using clean utterances. The WERs of Set A are much lower than that of Set B and Set C due to a lager mismatch between

Table 3. WER (%) of MTAE and DDAE on Aurora 2 under clean and multi-condition training (averaged over 0-20 dB).

Method	A_multi	B_multi	A_clean	B_clean
MFCC [20]	12.19	13.73	38.66	44.26
DDAE [6]	8.92	22.26	6.39	20.44
MTAE	7.71	18.20	5.52	15.82

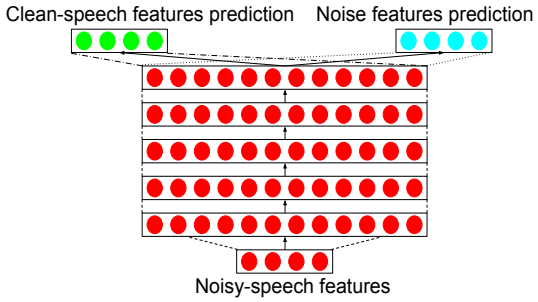

Fig. 3. Structure 1: The traditional MTL structure.

the training data and test data (Set B and Set C). Table 2 compares our method with the other methods. Our method outperformed the other methods in Set A and Set C. In the open noise condition test (Set B), it has higher WER than ETSI-AFE [14] and NMCC [16]. The reason may be that MTAE provides biased information of noise in the open noise condition test.

Table 3 represents Aurora2's results of the conventional DDAE and the proposed MTAE under both clean training and multi condition training. For clean training, the acoustic model is trained with clean data and DDAE/MTAE is trained with multi condition data. For multi-condition training, we train both the acoustic model and the DDAE/MTAE with multi-condition data. It is interesting that the clean training gives lower WERs than the multi condition training for both Set A and Set B. Probably this is because the predicted clean features are more similar to the clean training features than the multi condition training features.

To investigate the effectiveness of the triangular-shaped structure of MTAE, another two types of network structures are examined for comparison: (1) Structure 1: the number of nodes for each hidden layer is the same, as shown in Fig. 3, (2) Structure 2: the same structure with MTAE, but nodes are fully connected between layers, as shown in Fig. 4. Table 4 shows the results. We confirm that our method achieves better performance than these two.

In addition, we present the CHIME 3's results under clean training to show the scalability of the MATE in Table 5. It shows that our method also performs well in real case.

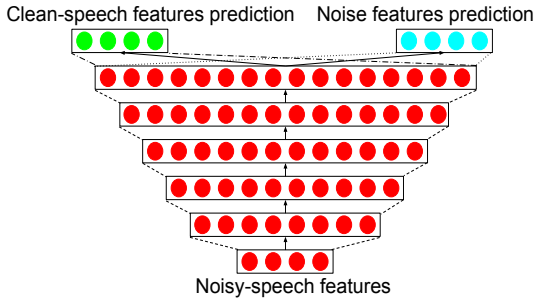

Fig. 4. Structure 2: The same number of nodes with MTAE, but fully connected between layers.

Table 4. The word error rate of Str.1 and Str.2 on Aurora 2 Set A under clean training.

SNR	Str. 1	Str. 2	MTAE
Clean	0.59	0.63	0.68
20dB	1.53	1.67	1.01
15dB	2.19	2.26	1.34
10dB	3.71	3.93	2.29
5dB	7.12	7.80	5.04
0dB	20.44	21.06	17.90
-5dB	56.81	57.52	53.57
Avg.	7.00	7.34	5.52

Table 5. WER (%) of MTAE and the other methods on CHIME 3 under clean training (averaged over 0-20 dB).

Test set	MFCC	DDAE	MTAE
et05_real	50.83	29.71	26.93

5. CONCLUSIONS

We have proposed a multi-task autoencoder (MTAE), which combines the denoising autoencoder and the deSpeeching autoencoder. By simultaneously training these two, the prediction abilities of the denoising autoencoder and the deSpeeching autoencoder are both boosted. Experiment results show that our proposed MTAE reduced WER by 15.7% from the conventional denoising autoencoder in Aurora 2 test set A.

In future, we investigate how many shared hidden units are the optimal. In addition, more different MTL structures should be applied to check how structures affect the effectiveness. It is also interesting to extend our method to multi-task recurrent autoencoder.

6. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI 16H02845 and by JST CREST Grant Number JPMJCR1687, Japan.

7. REFERENCES

[1] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 733–736, 1996.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *INTERSPEECH*, pp. 869–872, 2000.

[3] Y. Zhao and B.H. Juang, "On noise estimation for robust speech recognition using vector taylor series," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 4290–4293, 2010.

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the splice algorithm on the aurora2 database," in *Proceedings of Eurospeech*, pp. 217–220, 2001.

[5] S. Tamura and A. Waibei, "Noise reduction using connectionist models," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 553–556, 1988.

[6] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, "Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU)*, pp. 350–355., 2013

[7] A.L. Maas, Q.V. Le, T.M. O'Neil, O. Vinyals, P. Nguyen, and A.Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *INTERSPEECH*, pp. 22–25, 2012.

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *INTERSPEECH*, pp. 885–889, 2014.

[9] R. Caruana, "Multitask learning," in *Learning to learn*, Springer US, pp. 95–133, 1998.

[10] D. Chen and B. Mark, "Multitask learning of deep neural networks for low-resource speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1172–1183, 2015.

[11] D. Chen, B. Mak, C.C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 5592–5596, 2014.

[12] K.H. Lee, W.H. Kang, T.G. Kang, and N.S. Kim, "Integrated DNN-based model adaptation technique for noise-robust speech recognition," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 5245-5249, 2017.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, pp. 113–120, 1979.

[14] H.G. Hirsch and D. Pearce, "Applying the advanced etsi frontend to the aurora-2 task," in *technical report version 1.1*, 2006.

[15] C. Kim and R.M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, 2012.

[16] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, 2012.

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[18] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 7319–7323, 2013.

[19] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP)*, pp. 7304–7308, 2013.

[20] D. Pearce and H.G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *INTERSPEECH*, pp. 29–32, 2000.

[21] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding (ASRU)*, 2011.