Ensemble Modeling of Denoising Autoencoder for Speech Spectrum Restoration

Xugang Lu[1], *Yu Tsao*[2], *Shigeki Matsuda*[1], *Chiori Hori*[1]

1. National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academic Sinica, Taiwan

Abstract

Denoising autoencoder (DAE) is effective in restoring clean speech from noisy observations. In addition, it is easy to be stacked to a deep denoising autoencoder (DDAE) architecture to further improve the performance. In most studies, it is supposed that the DAE or DDAE can learn any complex transform functions to approximate the transform relation between noisy and clean speech. However, for large variations of speech patterns and noisy environments, the learned model is lack of focus on local transformations. In this study, we propose an ensemble modeling of DAE to learn both the global and local transform functions. In the ensemble modeling, local transform functions are learned by several DAEs using data sets obtained from unsupervised data clustering and partition. The final transform function used for speech restoration is a combination of all the learned local transform functions. Speech denoising experiments were carried out to examine the performance of the proposed method. Experimental results showed that the proposed ensemble DAE model provided superior restoration accuracy than traditional DAE models.

Index Terms: Denoising autoencoder, ensemble modeling, speech restoration.

1. Introduction

Estimating clean speech from noisy ones can be regarded as a function approximation problem in which the estimated function is used to describe the mapping relation between noisy and clean speech. Many classical algorithms try to estimate such a function as a linear function for noise reduction, for example, Wiener filtering and signal subspace method [1]. Considering that neural network can be used to learn a universal nonlinear function, it is promising in learning the mapping function for noise reduction. In recent years, with the development of deep learning algorithms in signal processing and pattern recognition [2, 3, 4], the neural network based noise reduction algorithms have been gotten much attention.

Under the line of neural network learning technique, several algorithms have been proposed [5, 6, 7]. Among them, the denoising autoencoder (DAE) based algorithms have been proposed in image denoising and robust feature extraction [5]. We have adopted similar idea in noise reduction and speech enhancement [8, 9]. The advantage of using DAE is that it is simple in concept, and can be easily stacked to a deep denoising autoencoder (DDAE) architecture for further improving the performance. In either DAE or DDAE, the transform function between noisy and clean speech is learned based on a large collection of mixtures of noisy speech and clean speech data pairs [9, 7].

In most studies, it is supposed that the DAE or DDAE can learn any type of complex transform functions between noisy and clean speech. However, since the learning is a kind of statistical average of all training data samples, for large variations of speech patterns and noisy environments, the learned model is lack of focus on local transformations between noisy and clean speech. In unmatched testing conditions, large estimation error may occur due to the weak generalization ability of the learned model. In machine learning, ensemble modeling is one of the efficient strategies for reducing model variance and increasing model generalization ability [10]. The basic idea is shown as in Fig. 1. In this figure, rather than using a single model to learn the mapping function $f(\cdot)$ (as shown in (a)), multiple models $f_1(\cdot)$ and $f_2(\cdot)$ (two as shown in (b)) are used to learn the mapping functions, and the final mapping function is a combination of the learned mapping functions as $f(\cdot) = g(f_1(\cdot), f_2(\cdot))$, where $g(\cdot)$ is a combination function. The ensemble modeling strategy has been used in speech processing for speaker and environment modeling and adaptation [11, 12]. Inspired by the ensemble modeling strategy, in this study, we propose an ensemble model of DAE to learn both the global and local transform functions for noise reduction. In the model, local transform functions are captured by several DAEs, and the final transform function is a combination of all the transform functions of the learned DAEs.

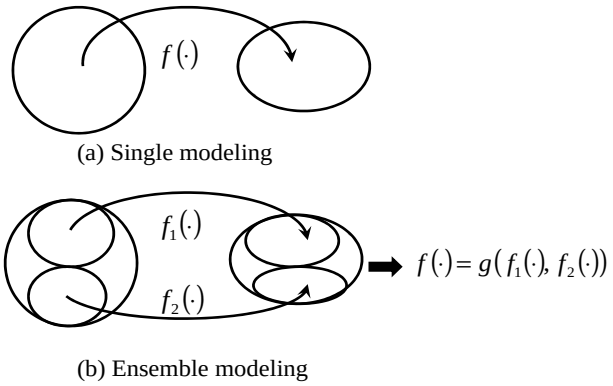

Figure 1: Single modeling (a), and ensemble modeling (b).

Our work is different from the work of using multi-column neural networks in image classification and robust image denoising [14, 15]. In image classification, the multi-column neural networks were trained for different feature representations, and the final classification is based on an average of the results from multi-column neural networks [14]. In robust image denoising [15], the multi-column neural networks were trained for different noise types and signal to noise ratio (SNR) conditions, and an adaptive weighting of the restored images from

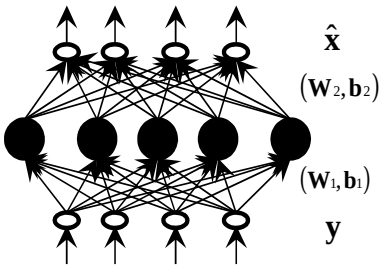

Figure 2: One hidden layer neural associator for speech denoising.

multi-column neural networks was applied to obtain the final estimation. Differently from their work, we borrowed the idea of ensemble learning from machine learning. Each DAE in the ensemble is trained using a data subset obtained from unsupervised data clustering and partition method. Unsupervised data clustering and partition is more suitable for reducing model variation than using the data set from certain noise types and SNR conditions.

The paper is organized as follows. Section 2 introduces the basic architecture of neural denoising autoencoder for speech spectrum restoration. Section 3 describes the proposed ensemble learning of the denoising autoencoder. Section 4 shows experimental results and evaluations. Discussions and conclusion are given in Section 5.

2. Denoising autoencoder

The DAE is widely used in building a deep neural architecture for robust feature extraction and classification [3, 13]. We have used the DAE and its deep version for speech enhancement [9]. The basic processing block of DAE is shown in Fig. 2. This DAE can be regarded as a one hidden layer neural associator with noisy speech as input and clean speech as output. It includes one nonlinear encoding stage and one linear decoding stage for real value speech as:

$$\begin{aligned} h(\mathbf{y}) &= \sigma(\mathbf{W}_1\mathbf{y} + \mathbf{b}_1) \\ \hat{\mathbf{x}} &= \mathbf{W}_2 h(\mathbf{y}) + \mathbf{b}_2, \end{aligned} \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are encoding and decoding matrix as the neural network connection weights, respectively. $\mathbf{y}$ and $\mathbf{x}$ are noisy and clean speech, respectively. $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias vectors of hidden and output layers, respectively. The nonlinear function of hidden neuron is a logistic function defined as $\sigma(\mathbf{x}) = (1 + \exp(-\mathbf{x}))^{-1}$. The model parameters are learned by doing the following optimization as:

$$\begin{aligned} \Theta^* &= \arg\min_{\Theta} \left(L(\Theta) + \alpha \left(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2 \right) \right) \\ L(\Theta) &= \tfrac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \end{aligned} \tag{2}$$

where $\Theta = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ is the parameter set, and $\mathbf{x}_i$ is the i-th training clean sample corresponding to the noisy sample $\mathbf{y}_i$, and N is the total number of training samples. In Eq.2, α is used to control the tradeoff between reconstruction accuracy and regularization on weighting coefficients (it was set $\alpha = 0.0002$ in this study). The optimization of Eq. (2) can be solved by using many unconstrained optimization algorithms. In this study, a Hessian-free algorithm is applied for model parameter learning [16].

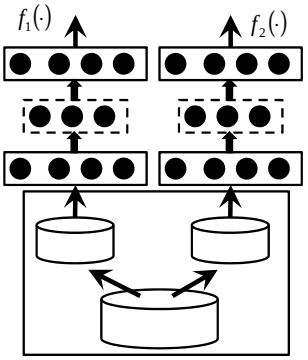

Figure 3: Learning ensemble DAEs based on unsupervised data clustering and partitions.

3. Ensemble denoising autoencoder

The DAE can be used to learn the mapping function between noisy and clean speech. In most studies, multi-conditional training is applied to train the model parameters, i.e., training data set is composed of mixed noisy conditions with various SNRs. However, the learned model from the multi-conditional training data has large model variation which is not accurate for some local transformations. In order to keep high accuracy for local transformations, we propose ensemble modeling of DAE for speech denoising. The proposed ensemble modeling has two steps, the first step is training ensemble DAEs, and the second step is combination of the ensemble DAEs. Figs. 3 and 4 show the two steps.

3.1. Training of ensemble DAEs

As shown in Fig. 3, the training data set is first clustered and partitioned into several groups (for convenience of explanation, two subgroups are supposed if there is no special mention). The clustering and partitioning is based on an unsupervised clustering algorithm. In this study, a simple K-means clustering algorithm is used. The data subsets are clustered by minimizing the following objective function:

$$J = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \left\| \mathbf{y}_j^{(i)} - \mathbf{C}_i \right\|_2^2, \tag{3}$$

where K is the total cluster number, N_i is the sample number in cluster i, and $\mathbf{C}_i$ is the average of the i cluster (or centroid vector). The clustering is done on training data set of mixtures of noisy speech spectrum. After clustering, similar noisy speech patterns (in Euclidian distance sense) are clustered into several subsets. The advantage of using an unsupervised clustering on data is that data partitioning can be automatically obtained based on their local statistical structure. In addition, based on this unsupervised clustering, speech spectra collected from one sentence in one noisy condition may be clustered into different data clusters. Based on the partitioned data subsets, multiple DAEs are trained.

3.2. Combination of ensemble DAEs

After multiple DAEs are trained, a combination function on the ensemble DAEs is applied (as shown in Fig. 4). The combination function can be any type of linear or nonlinear functions defined as:

$$f(\cdot) \triangleq g(f_1(\cdot), f_2(\cdot)), \tag{4}$$

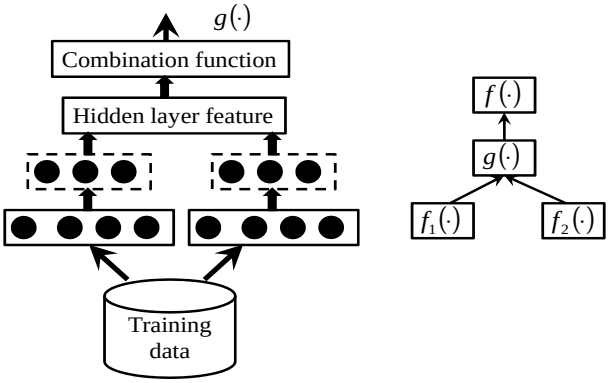

Figure 4: Combination of DAEs in ensemble modeling with a learned combination function $g(\cdot)$.

where the combination function $g(.)$ can be learned from the training data set. For simplicity, a linear weighting function is applied for each training sample in this study, therefore the global mapping function is estimated as:

$$f(\cdot) \triangleq \lambda_1(\cdot) f_1(\cdot) + \lambda_2(\cdot) f_2(\cdot), \quad (5)$$

where $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are the weighting coefficient functions.

Similarly as used in robust image denoising [15], for each training sample $\mathbf{y}$, they are learned by minimizing the restoration error as:

$$\Lambda^* \triangleq \arg \min_{\Lambda=[\lambda_1(\mathbf{y}),\lambda_2(\mathbf{y})]} \|f(\mathbf{y}) - \mathbf{x}\|_2^2, \quad (6)$$

where $f(\mathbf{y}) = \lambda_1(\mathbf{y}) f_1(\mathbf{y}) + \lambda_2(\mathbf{y}) f_2(\mathbf{y})$ is the estimated restoration as defined in Eq. 5. For overcoming overfitting problem in solving the problem in Eq. 6, the constraints $0 \leq \lambda_1(\mathbf{y}), \lambda_2(\mathbf{y}) \leq 1$ and $\lambda_1(\mathbf{y}) + \lambda_2(\mathbf{y}) = 1$ are added.

For each training sample, we obtain a weighting coefficient set from solving Eq. 6. In real applications, for a testing sample, the weighting coefficient set can be predicted from a regression fitting function of the weighting coefficient sets of trained samples. In estimating the regression function (a linear function was used for simplicity), the input of the regression function is the hidden layer outputs of the DAEs as shown in Fig. 4, and the output is the learned weighting coefficient of the corresponding training samples. With this method, we can adaptively adjust the weighting coefficients for a better restoration than with fixing the weighting coefficients for all testing conditions.

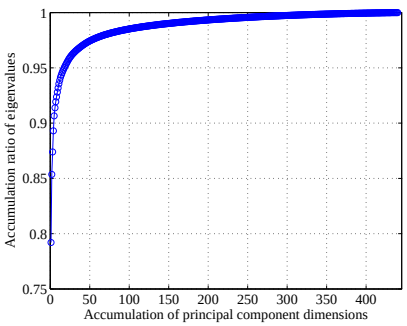

Figure 5: Ratio between accumulation of top eigenvalues and sum of total eigenvalues.

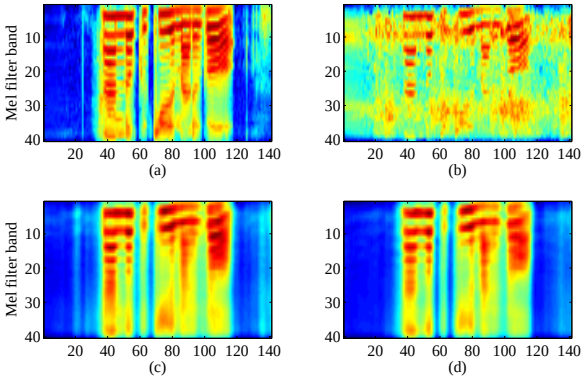

Figure 6: Clean spectrum (a), noisy spectrum (b), denoised spectrum from proposed ensemble modeling (c), denoised spectrum from matched noisy type and SNR condition (d).

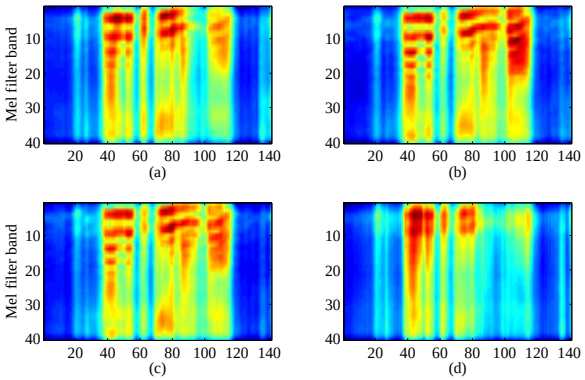

Figure 7: Denoised spectrum from four DAEs in ensemble modeling, respectively.

4. Experiments and evaluations

In this section, we evaluate the performance of the proposed ensemble DAE modeling on speech denoising task. As our first step in this study, we want to confirm whether the ensemble modeling can help in accurate speech spectrum restoration, the performance is measured based on restoration error (RtErr) defined as:

$$\text{RtErr} \triangleq \frac{1}{\#\text{Total}} \sum_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \quad (7)$$

where $\#\text{Total}$ is the total number of testing samples. This criterion measures the restoration error caused by both the speech distortion and noise residual as used in speech enhancement experiments [1].

Table 1: Restoration error (dB) for testing data in subway noise condition

SNR (dB)	DAE_1	DAE_4	DAE_16	Proposed
5	1.200	1.159	1.105	**0.617**
10	1.177	1.165	1.127	**0.534**
15	1.195	1.176	1.126	**0.492**
20	1.204	1.179	1.116	**0.438**

Table 2: Restoration error (dB) for testing data in babble noise condition

SNR (dB)	DAE_1	DAE_4	DAE_16	Proposed
5	1.272	1.173	1.107	**0.657**
10	1.185	1.141	1.080	**0.534**
15	1.226	1.196	1.127	**0.467**
20	1.277	1.237	1.142	**0.416**

Table 3: Restoration error (dB) for testing data in car noise condition

SNR (dB)	DAE_1	DAE_4	DAE_16	Proposed
5	1.184	1.146	1.112	**0.622**
10	1.200	1.178	1.122	**0.532**
15	1.208	1.177	1.105	**0.475**
20	1.244	1.221	1.143	**0.437**

The experiments were carried out on the AURORA2J data corpus (continuous Japanese digits speech for noisy environments) [17]. Four types of noises (subway, babble, car, and exhibition) and each with SNR conditions 5, 10, 15 and 20 dB were used. In training, each noisy condition (one combination of noise type and SNR condition) has 422 speech utterances. In test, each noisy condition has 100 speech utterances which are different from training set. The feature used in DAE learning is 40 Mel frequency band spectrum extracted from frame based processing (20 ms frame length with 10 ms frame shift). 11 continuous frames were concatenated to be a vector as input to the DAE. Therefore, the input layer size of DAE is $11 * 40 = 440$ dimensions. Because the dimensions of input vector has high correlation (due to frame based concatenation), the hidden dimension of DAE may use a small number of dimensions for denoising. We did principal component analysis (PCA) of the training data set (mixtures of noisy speech for all noise types and SNR conditions). The ratio between the accumulation of top eigenvalues and sum of total eigenvalues is shown in Fig. 5. From the analysis, we found that that using the top 100 principal components could reconstruct the data set with 98.52% reconstruction accuracy. Based on this investigation, the hidden layer size of DAE is set to 100 in this study.

For comparison, several types of DAE based denoising methods were applied: (1) single DAE model trained with a date set composed of all mixtures of noise types and SNR conditions (multi-conditional training as mostly used), (2) four DAEs model and each DAE is trained with a data set composed of one noise type and mixed SNR conditions, (3) 16 DAEs model, and each DAE is trained with a data set composed of one noise type combined with one SNR condition (totally 4*4=16 combinations), (4) in our proposed ensemble modeling, four DAEs are adopted, and each is trained with one cluster of data set obtained from K-means clustering and partition (four partitions in total). In testing, for method (2) of using four DAEs, the matched noise type trained DAE is chosen for denoising, for method (3) of using 16 DAEs, the matched noise type and SNR condition trained DAE is chosen for denoising, and for our proposed ensemble modeling, the final denoising is based on the weighting combination of the four DAEs. The weighting coefficients are estimated from weighting regression function as introduced in section 3.2.

Table 4: Restoration error (dB) for testing data in exhibition noise condition

SNR (dB)	DAE_1	DAE_4	DAE_16	Proposed
5	1.204	1.171	1.120	**0.598**
10	1.211	1.194	1.139	**0.532**
15	1.197	1.185	1.124	**0.482**
20	1.242	1.223	1.149	**0.431**

Before quantitative evaluation, we visually examine how the restored spectrum looks like. An utterance in factory noise with SNR 5dB condition is tested. The clean spectrum and noisy spectrum are shown in panels (a) and (b) of Fig. 6, respectively. Fig. 7 shows the restored spectrum from the four DAEs in ensemble modeling. Panel (c) of Fig. 6 shows the weighted combination of the restoration from four DAEs in ensemble modeling. Panel (d) of Fig. 6 shows the restoration by using the DAE trained by data set of matched noise type and SNR condition. In these figures, x-axis is the time frame index, y-axis is Mel frequency filter band index. From these figures, we can see that the proposed ensemble modeling got a better restoration than the DAE trained using data set with even matched noise type and SNR condition.

For quantitative evaluation, the results measured with restoration error for each testing condition is shown in tables (1), (2), (3), (4). In the tables, "DAE_1", "DAE_4", "DAE_16" denote the methods (1), (2) and (3) as mentioned above, respectively, and "Proposed" represents our proposed ensemble modeling. The value calculated using Eq. 7 is in dB since the Mel frequency band spectrum is calculated in dB scale. From these four tables, we can see that the restoration error gradually becomes larger and larger from DAE_16 to DAE_4 and DAE_1. This is reasonable since the model focuses more on global transform from DAE_16 to DAE_4 and DAE_1. The proposed ensemble modeling, although only four DAEs were used, got the best performance in all testing conditions. We confirm that final restoration performance is benefitted from the local transform function based restorations which are learned in ensemble modeling.

5. Conclusion and discussions

DAE and its deep architecture have been proposed for robust feature learning and classification [3, 13]. And later, they are successfully used for image denoising and classification [5]. We have applied the DAE and its deep version DDAE architecture for noise reduction and speech enhancement [9]. In our previous studies, the DAE or DDAE is trained either as matched noise type and SNR condition or as a multi-conditional training with large mixtures of noise types and SNR conditions. However, the trained model is lack of focus on local transformations between noisy and clean speech. In this study, we introduced an ensemble modeling of DAEs for speech denoising. The advantage of this method is that local transform can be kept well in ensemble modeling. During denoising, the test noisy speech can be adaptively denoised based on several local denoising functions (DAEs) in ensemble modeling. Our experimental results confirmed the effectiveness of the proposed ensemble DAE modeling.

In this study, four DAEs were trained in ensemble modeling. In the future, we need to investigate how many DAEs are optimal for a training data set. In addition, the DAE is a one hidden layer neural network, its deep architecture DDAE has already been proved to improve restoration accuracy. Extending the ensemble modeling on DDAE is our another future work.

6. References

[1] Loizou, P. C., Speech Enhancement: Theory and Practice, CRC Press, 2007.

[2] Hinton, G. E., and Salakhutdinov, R., "Reducing the Dimensionality of Data with Neural Networks," Science, 313: 504-507, 2006.

[3] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H., "Greedy layer-wise training of deep networks," In Advances in Neural Information Processing Systems, 19: 153-160, MIT Press, Cambridge, 2007.

[4] Ranzato, M. A., Huang, F. J., Boureau, Y. L., LeCun, Y., "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," IEEE conference on Computer Vision and Pattern Recognition, 1-8, 2007.

[5] Xie, J., Xu, L., and Chen, E., "Image denoising and inpainting with deep neural networks," Advances in Neural Information Processing Systems 25, 2012.

[6] Burger, S. H., "Image Denoising: Can Plain Neural Networks Compete with BM3D?," CVPR, 2012.

[7] Xu, Y., Du, J., Dai, L., Lee, C., "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," IEEE Signal Processing Letters, 21(1):65-68, 2014.

[8] Lu, X., Matsuda, S., Hori, C., Kashioka, H., "Speech restoration based on deep learning autoencoder with layer-wised learning," INTERSPEECH, Portland, Oregon, Sept., 2012.

[9] Lu, X., Yu, T., Matsuda, S., Hori, C., "Speech enhancement based on deep denoising autoencoder," INTERSPEECH 2013: 436-440.

[10] Dietterich, T. G., "Ensemble Methods in Machine Learning," Int. Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, 1857: 1-15, 2000.

[11] Tsao, Y., Lee, C., "An Ensemble Speaker and Speaking Environment Modeling Approach to Robust Speech Recognition," IEEE Transactions on Audio, Speech and Language Processing, 17(5): 1025-1037, 2009.

[12] Tsao, Y., Lu, X., Dixon, P., Hu, T., Matsuda, S., Hori, C., "Incorporating local information of the acoustic environments to MAP-based feature compensation and acoustic model adaptation," Computer Speech and Language 28(3): 709-726, 2014.

[13] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P., "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of Machine Learning Research, 11(Dec): 3371-3408, 2010.

[14] Ciresan, D. C., Meier, U., Schmidhuber, J., "Multi-column Deep Neural Networks for Image Classification," IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012.

[15] Agostinelli, F., Anderson, M., Lee, H., "Adaptive Multi-Column Deep Neural Networks with Application to Robust Image Denoising," In NIPS 2013.

[16] Martens, J., "Deep Learning via Hessian-free Optimization," In Proceedings of the 27th International Conference on Machine Learning (ICML), 2010.

[17] Nakamura, S., Takeda, K., Yamamoto, K., Yamada, T., Kuroiwa, S., Kitaoka, N., Nishiura, T., Sasou, A., Mizumachi, M., Miyajima, C., Fujimoto, M., and Endo, T., "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. Syst., 88 (D): 535-544, 2005.